applied sciences

MDPI

Article

Hybrid Inductive Model of Differentially and Co-Expressed Gene Expression Profile Extraction Based on the Joint Use of Clustering Technique and Convolutional Neural Network

Sergii Babichev [1,2,*,†,‡], Lyudmyla Yasinska-Damri [3,‡], Igor Liakh [4,‡] and Jiří Škvor [1,†,‡]

[1] Department of Informatics, Jan Evangelista Purkyně University in Ústí nad Labem, 400 96 Ústí nad Labem, Czech Republic
[2] Department of Physics, Kherson State University, 73008 Kherson, Ukraine
[3] Department of Computer Science and Information Technologies, Ukrainian Academy of Printing, 79020 Lviv, Ukraine
[4] Department of Information Science and Physics and Mathematics Disciplines, Uzhhorod National University, 88000 Uzhhorod, Ukraine
* Correspondence: sergii.babichev@ujep.cz or sbabichev@ksu.ks.ua; Tel.: +420-777-843-785
† Current address: Pasteurova 3632/15, 400 96 Ústí nad Labem, Czech Republic.
‡ The authors contributed to this work as follows: the first author—50%, the second and the third ones—20%, the fourth author—10%.

Abstract: The development of hybrid models focused on gene expression data processing for the allocation of differentially expressed and mutually correlated genes is one of the current directions in modern bioinformatics. The solution to this problem can allow us to improve the effectiveness of existing systems for complex diseases diagnosis based on gene expression data analysis on the one hand and increase the efficiency of gene regulatory network reconstruction procedures by more careful selection of genes by considering the type of disease on the other hand. In this research, we propose a stepwise procedure to form the subsets of mutually correlated and differentially expressed gene expression profiles (GEP). Firstly, we allocate an informative GEP in terms of statistical and entropy criteria using the Harrington desirability function. Then, we performed cluster analysis using SOTA and spectral clustering algorithms implemented within the framework of objective clustering inductive technology. The result of this step's implementation is a set of clusters containing co- and differentially expressed GEPs. Validation of the model was performed using a one-dimensional two-layer convolutional neural network (CNN). The analysis of the simulation results has shown the high efficiency of the proposed model. The clusters of GEPs formed based on the clustering quality criteria values allowed us to identify the investigated objects with high accuracy. Moreover, the simulation results have also shown that the hybrid inductive model based on the spectral clustering algorithm is more effective in comparison with the use of the SOTA clustering algorithm in terms of both the complexity of the formed optimal cluster structure and the classification accuracy of the objects that contain the allocated gene expression data as attributes. The proposed hybrid inductive model contributes to increasing objectivity during the formation of the subsets of differentially and co-expressed gene expression profiles for further their application in various disease diagnosis systems and for gene regulatory network reconstruction.

Keywords: gene expression profiles; SOTA clustering algorithm; spectral clustering algorithm; convolutional neural network; inductive clustering technique; Harrington desirability function

Citation: Babichev, S.; Yasinska-Damri, L.; Liakh, I.; Škvor, J. Hybrid Inductive Model of Differentially and Co-Expressed Gene Expression Profile Extraction Based on the Joint Use of Clustering Technique and Convolutional Neural Network. *Appl. Sci.* **2022**, *12*, 11795. https://doi.org/10.3390/app122211795

Academic Editors: Jeong Seop Sim and SooJun Park

Received: 6 November 2022
Accepted: 18 November 2022
Published: 20 November 2022

1. Introduction

The application of both data mining and machine learning techniques to develop hybrid models for the purpose of allocation of differentially and co-expressed gene expression profiles (GEP) is one of the current directions in modern bioinformatics. The solution to this problem can allow us to form subsets of informative GEPs which can be used at the next

step for both the creation of various disease diagnostic systems such as cancer, Parkinson's, Alzheimer's, etc., and to more effectively reconstruct gene regulatory networks (GRNs) based on the allocated groups of genes. The following analysis of these GRNs during the implementation of the simulation procedure can contribute to a better understanding of the particularities of the molecular component's interconnection and the influence of these interconnections on the state of target genes by considering the state of the respective disease. In this case, the genes are differentially expressed, the expression values of which significantly differ between patients with various states of disease (for example, healthy and ill). Moreover, the values of gene expressions for all samples should be larger than the appropriate threshold value that can be established empirically based on the joint use of various types of criteria. The allocated gene expression profiles should be also mutually correlated (co-expressed). This means that the gene expression values should be varying consistently with respect to the investigated disease state in the appropriate samples.

It should be noted that the allocation of a set of informative (differentially expressed and mutually correlated) GEPs depends on selection and the use of a proximity metric to evaluate the appropriate GEP proximity level. The traditional metrics such as Euclidean, Manhattan, etc., are not effective for this type of data due to the high dimensionality of the data. In review [1], the authors presented a comparative analysis of various models to form co-expressed gene expression profiles using various types of proximity metrics focused on high-dimensional data. Within the framework of the implementation of various hybrid models, they analyzed the following metrics: correlation proximity, metrics based on the evaluation of mutual information, and the chi-squared test. In [2], the authors compared the effectiveness of various GEP proximity metrics. In this research, they have shown that the correlation proximity metric is less effective in comparison to metrics based on mutual information and the chi-squared test. Moreover, they have also shown that the metrics based on mutual information can disagree with each other when using various methods of Shannon entropy evaluation. As a result of the research, the authors proposed a hybrid metric based on the joint use of various types of mutual information metrics and the chi-squared test. However, it should be noted a significant disadvantage of the proposed technique exists in the high time and computer resources to implement this step of the data processing.

Another challenge that should be solved when processing gene expression data is the formation of subsets of co-expressed GEPs that can be used for the creation of disease diagnosis systems (classification problems) and the reconstruction of gene regulatory networks. There are various clustering algorithms focused on high-dimensional gene expression data processing. So, the self-organizing SOTA clustering algorithm [3] is a logical continuation of Kohonen's neural network [4]. The result of this algorithm operation is a topological binary tree formed in accordance with the cell structure growing algorithm proposed by Fritzke [5]. However, we would like to note that the application of the SOTA algorithm with proximity metrics focused on high-dimensional data (for example, correlation metrics) leads to the division of gene expression data into two subsets that are almost similar in size at one step of the algorithm's implementation. A stepwise procedure implementation of this algorithm is possible; however, in this case there arises the problem of the optimal cluster structure formation. Internal clustering quality criteria can allow us to form an optimal cluster structure, but they do not solve the reproducibility error. Successful results obtained in one dataset do not guarantee obtaining the same results when using other similar datasets.

Another clustering algorithm used for gene expression data clustering is the spectral clustering algorithm [6–9]. This algorithm is one of the modern algorithms which allows us to identify clusters of arbitrary shape based on the application of object similarity matrices. In comparison with traditional algorithms, the spectral clustering algorithm has many fundamental advantages. The results obtained by applying the spectral clustering algorithm are often superior in quality to those obtained using traditional approaches. Moreover, the implementation of the spectral clustering algorithm is straightforward and

can be effectively implemented based on the methods of standard linear algebra. The general difference of the spectral clustering algorithm is that the formation of clusters is not based on the absolute location of objects in the feature space, but instead on the basis of the analysis of the level of these objects' affinity, which is especially useful when forming clusters with complex shapes. The principal problem in the successful implementation of the spectral clustering algorithm is the formation of the optimal cluster structure according to both the relevant clustering quality criteria and the reproducibility error. It should be noted that this problem does not currently have an unequivocal solution.

In this research, we present our point of view to solve this problem solution. We propose a hybrid inductive model of the stepwise procedure to form differentially and co-expressed gene expression profiles based on the joint use of statistical and entropy criteria, spectral and SOTA clustering algorithms implemented within the framework of objective clustering inductive technology, and convolutional neural networks for obtaining the final solution concerning the choice of optimal cluster structure.

The main contributions of this manuscript are the following:

- We propose a technique to form a subset of informative GEPs based on the joint use of statistical criteria and Shannon entropy, where the final solution regarding the level of respective gene informativity is conducted based on analysis of the general Harrington desirability index.
- We implemented the SOTA and spectral clustering algorithms within the framework of objective clustering inductive technology, which allows us to decrease reproducibility error when the gene expression data are clustered.
- We have proposed a hybrid model for the extraction of differentially and co-expressed gene expression profiles based on the joint use of an inductive clustering algorithm (SOTA or spectral clustering) and a convolutional neural network.
- The proposed hybrid inductive model contributes to increasing objectivity during the formation of the subsets of differentially and co-expressed gene expression profiles to further their applications in various disease diagnosis systems and gene regulatory network reconstruction.

2. Literature Review

Gene expression processing based on the joint use of various data mining and machine learning techniques to improve appropriate diagnosis classification is currently very successful. So, in [10], the authors presented research results focused on applying various feature selection and machine learning techniques to allocate the biomarkers with the highest score when performing breast cancer disease detection. In this research, the authors used ROC analysis with the calculation of the AUC (area under curve) to evaluate the appropriate classification model's quality. However, it should be noted that the authors applied principal component analysis as the feature extraction method. In this case, we lose information about individual genes, which is not acceptable in gene regulatory network reconstruction. In [11], the authors introduced a predictive modeling approach to process the gene expression data to infer treatment responses in cancers. In this research, the authors demonstrated the benefits of considering pathway activity estimates in tandem with drug descriptors as features. Using data from The Cancer Genome Atlas, they also showed the applicability of the proposed approach on patient drug response and an independent clinical study describing the treatment journey of three melanoma patients.

The results of the research regarding the development of an approach focused on gene expression subset selection for high-dimensional gene expression data using a multi-objective optimization-based multi-view co-clustering algorithm is presented in [12]. In this research, the authors applied the non-dominated sorting genetic algorithm II evolutionary technique as an optimization strategy. The authors showed that the reduction in dimensions formed by new feature- pace causes decreases in the computational burden and noise level of the original data. However, we would like to note that applying a co-clustering algorithm

can lead to a loss of useful samples during the bicluster allocation. This fact limited the successful application of the proposed method.

In [13], the authors proposed a method for informative gene extraction and following classification of samples based on the extracted genes in diffuse large B-cell lymphoma (DLBCL). In this research, the informative gene expression subset was selected according to the calculation of the cosine angle and distance between the map and the ideal template. However, in this research, the authors do not compare various distance metrics focused on high-dimensional data. The classification accuracy of about 85% obtained during the simulation procedure implementation is very low for this type of data. The research results regarding applying a modified Wrapper feature selection model to address the gene classification challenge by replacing its randomness approach with an extended particle swarm optimization model (EPSO) and PSO are presented in [14]. The simulation results showed that the EPSO method required less processing time to select the optimal features (an average of 62.14 s) than PSO (an average of 95.72 s). Moreover, EPSO's accuracy provided better classification results (from 54% to 100%) than PSO (from 52% to 96%).

In [15–20], the authors developed hybrid models based on the joint use of clustering and classification techniques, where various proximity metrics were used at the stage of gene expression profile filtering. The analysis of the presented sources allows us to conclude that the challenge in objective extraction of informative genes considering resolving power, which allows for the recognition of investigated objects with high accuracy, does not have an unambiguous solution. In most cases, high classification accuracy is achieved when using a small number of the most informative genes. Only in a few cases of the authors' research, the number of extracted genes exceeded 100 when a high classification accuracy was achieved. The significant shortcomings of the presented models should also include the fact that in most cases, the parameters of the corresponding algorithms were determined empirically during the simulation process. In other words, the models are not self-organizing. This fact introduces a large amount of subjectivity into the process of making the final decision regarding gene extraction.

The questions regarding the application of both convolution neural networks and deep learning techniques in various fields of scientific research are presented in [21,22]. So, in [21], the authors considered the application of a 2D convolutional neural network in the torsional capacity evaluation of reinforced concrete beams. A model of diagnosing surface cracks in concrete structures based on CNN application has been considered in [22]. In these studies, the authors have shown CNN's performance in solving complex problems.

The unsolved part of the general problem is the absence of objective techniques to allocate large numbers of differentially and co-expressed gene expression profiles and contribute to the high accuracy of the studied object's classification.

In our research, we present our decision regarding gene expression profile extraction based on the joint use of the Harrington desirability function to form the general criterion to identify the informativity level of gene expression profiles based on the joint use of statistical criteria and Shannon entropy. Then, we implement the hybrid inductive model based on the joint use of a clustering algorithm (SOTA and spectral clustering) and a convolutional neural network (CNN).

3. Material and Methods

3.1. Conceptual Description of the Hybrid Model to Form Subsets of Differentially Expressed and Mutually Correlated GEPs Based on the Joint Application of Clustering Algorithms and CNN

As was noted in the introduction, the main goal of this research is to increase the objectivity of forming subsets of differentially expressed and mutually correlated GEPs, which creates the conditions for increasing the accuracy of the classification of objects whose attributes are the selected gene expression values. The application of the clustering stage allows the division of the set of gene expression profiles into subsets that are mutually close by the appropriate metrics, and the choice of the metrics by which the clusters are formed is very important. In [2], various metrics focused on evaluating the proximity level

of high-dimensional GEPs were investigated, and as a result of the research, a hybrid metric based on the joint application of metrics based on the assessment of mutual information and Pearson's chi-squared test was proposed. This metric is used to form a cluster structure within the framework of our research.

Figure 1 shows the structural block chart of the stepwise procedure of the general concept to form the subsets of differentially expressed and mutually correlated GEPs based on the joint application of clustering and classification methods.

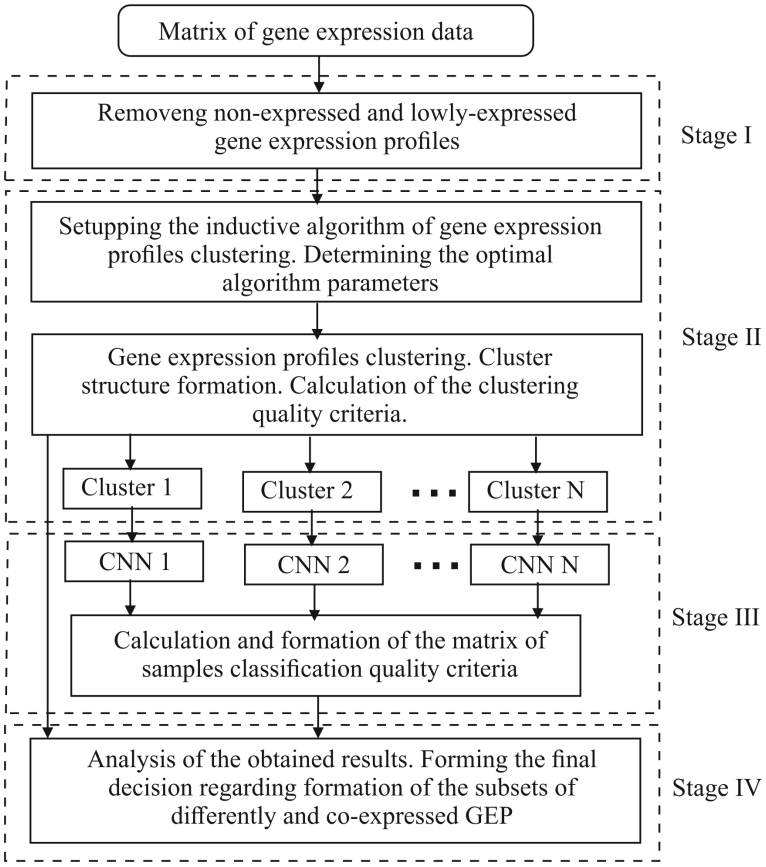

Figure 1. Structural block chart of the stepwise procedure for implementing the concept to form the subsets of differentially expressed and mutually correlated GEPs.

The practical implementation of this concept involves the following stages:

I. Removing non-expressed and low-expressed gene expression profiles. The implementation of this step assumes the removing of genes with zero expression for all samples and non-informative genes by the absolute value of gene expressions, variance, and Shannon entropy.

II. Dividing the subset of the gene expression profiles formed in stage I into two equivalent subsets with subsequent parallel clustering on these data subsets within the range of the clustering algorithm's variation parameter. The final decision regarding the optimal parameters of the corresponding clustering algorithm operation is performed taking into account the analysis of the values of the balance criterion, which contains, as the components, both the internal and external clustering quality criteria. The optimal cluster structures correspond to the maxima values of the balance criterion.

III. Classification of objects containing gene expression data in the selected clusters by applying a one-dimensional two-layer convolutional neural network (CNN). At this stage, a CNN is applied to the gene expression data of each cluster in order to classify the examined objects. The implementation of this stage requires preprocessing of gene expression value vectors in order to set the filters correctly. Vectors of gene expressions are supplemented with genes with zero expression so that the length

of the gene expression vector is divided without residuals by the filter size. The classification stage results in calculation of the classification quality criteria based on the gene expression data allocated in each of the clusters.

IV. Analysis of the obtained results and formation of subsets of differentially expressed and mutually correlated GEPs.
The implementation of this stage involves a comprehensive analysis of both the clustering and classification quality criteria. The subsets of gene expression data that correspond to the maximum values of both the clustering and classification criteria are allocated at this step. It should be noted that this stage's implementation assumes the formation of several subsets of differentially expressed and mutually correlated gene expression profiles according to the maximum values of the relevant criteria. This fact creates the conditions for increasing the objectivity of the studied objects' classification due to the agreed decision-making process regarding the identification of the objects considering the result of the classification of this object based on the gene expression data of all allocated clusters.

3.2. The Technique of Informative Gene Expression Profile Formation Using Statistical Criteria and Shannon Entropy Based on the Use of Harrington Desirability Function

Within the framework of the proposed method, the GEP is considered an informative one if the maximum values of the expression for all samples and variance of this profile are greater and the Shannon's entropy of the GEP is less than the respective boundary values:

$$\{e_{ij}\} = \begin{Bmatrix} \underset{i=\overline{1,n}}{max}(e_{ij} \geq e_{bound}),\ and\ var(e_{ij}) \geq var_{bound}, \\ and\ entr(e_{ij}) \leq entr_{bound} \end{Bmatrix}, j = \overline{1,m} \quad (1)$$

where n is the number of samples or objects to be examined and m is the number of genes.

In [23], the authors presented a solution regarding the division of a set of gene expression profiles into subsets considering the level of gene informativity based on the use of a fuzzy inference system. The simulation results showed that the main disadvantages of the fuzzy model applied for forming subsets of informative GEPs are the high complexity of information processing and high sensitivity to the model parameters, which significantly influence the obtained results. Moreover, the proposed system has allowed the authors to allocate about 13,000 highly informative genes from 25,000 ones. It should be noted that this quantity is very large for the following data processing steps.

As an alternative, we propose a solution to this challenge based on the Harrington desirability function, which is currently successfully used in various fields of scientific research [24,25]. This method is based on the following equation:

$$d = exp(-exp(-Y)) \quad (2)$$

where Y is a dimensionless parameter, the value of which varies within the range from -2 to 5 and d is private desirability that corresponds to one of the criteria used for forming a generalized quality index.

Figure 2 shows Harrington's desirability function, which is the basis for forming the general desirability index, the value of which determines the appropriate decision. It is obvious that the boundary values that distinguish the extreme intervals desirability value variations of 0.2 (unacceptable–bad) and 0.8 (good–excellent) are conditional ones, and they can be adjusted depending on the nature of the model input parameter values variation. The boundaries separating the corresponding intervals within the range $0.37 = 1/e$ (bad–satisfactory) and $0.63 = 1 - 1/e$ (good–excellent) are fixed and correspond to the intersection points of the desirability function. Within the framework of the proposed model, it is assumed that parameter Y and the values of the criteria applied as the model input change with the linear law.

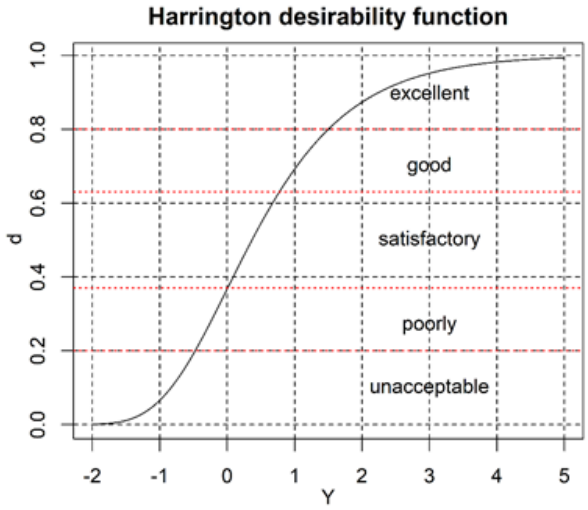

Figure 2. Harrington's desirability function and standard marks on the desirability scale.

The algorithm to calculate the general Harrington desirability index involves the following steps:

1. Determination of the coefficients of the linear equation to transform the values of statistical criteria and Shannon's entropy into the value of the Y parameter, taking into account the boundary values of the respective criteria and the particularities of their change:

$$\begin{gathered} Y_{min} = a_1 + b_1 \cdot max\_expr_{min};\ \ Y_{max} = a_1 + b_1 \cdot max\_expr_{max} \\ Y_{min} = a_2 + b_2 \cdot var_{min};\ \ Y_{max} = a_2 + b_2 \cdot var_{max} \\ Y_{min} = a_3 - b_3 \cdot entr_{max};\ \ Y_{max} = a_3 - b_3 \cdot entr_{min} \end{gathered} \tag{3}$$

where $Y_{min} = -2;\ Y_{max} = 5$.

2. Determination of Y parameter values for each of the criteria used in the model as the input data:

$$\begin{gathered} Y_{max\_expr} = a_1 + b_1 \cdot max\_expr \\ Y_{var} = a_2 + b_2 \cdot var \\ Y_{entr} = a_3 - b_3 \cdot entr \end{gathered} \tag{4}$$

3. Calculation of private desirabilities for each value of the used criteria:

$$\begin{gathered} d_{max\_expr} = exp(-exp(-Y_{max\_expr})) \\ d_{var} = exp(-exp(-Y_{var})) \\ d_{entr} = exp(-exp(-Y_{entr})) \end{gathered} \tag{5}$$

4. Calculation of the general index of the gene expression profiles significance as a geometric mean of the private desirabilities:

$$GI = (d_{max_expr} \cdot d_{var} \cdot d_{entr})^{1/3} \tag{6}$$

A higher value of the general index (6) corresponds to a higher significance level of the appropriate GEP.

Figure 3 shows the structural block chart of the algorithm that forms subsets of informative gene expression profiles based on Harrington's desirability function and evaluating the model's adequacy by calculating the classification accuracy of the investigated samples.

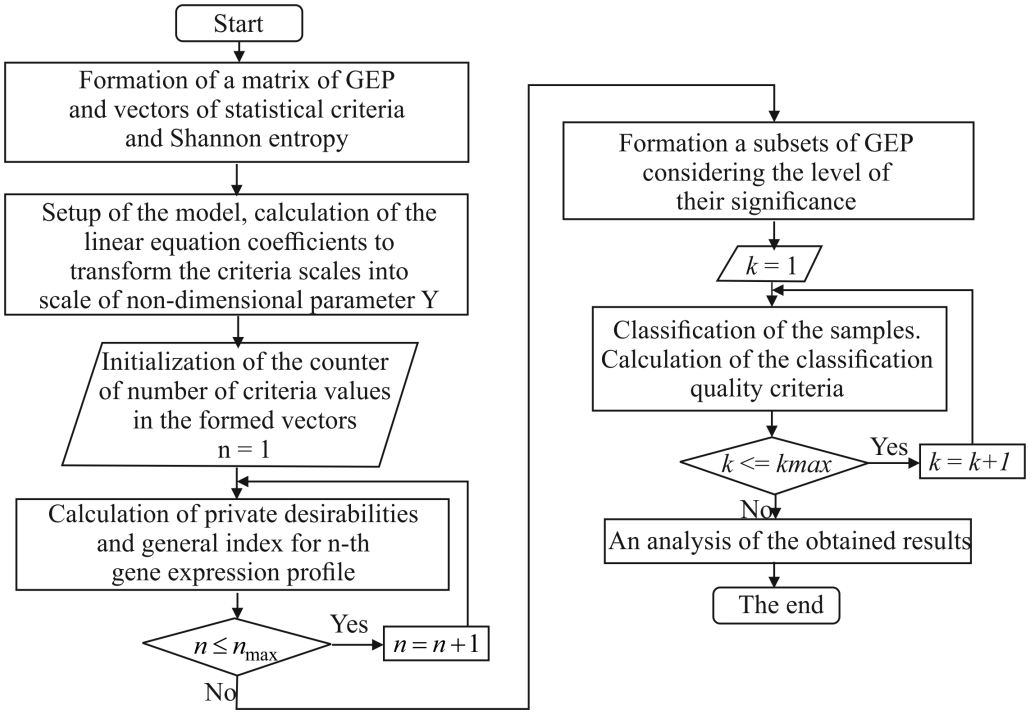

Figure 3. Structural block chart of the algorithm that forms the subsets of informative GEPs based on Harrington's desirability function and assessment of the model adequacy.

Implementation of the model assumes the following stages:

I. Formation of gene expression matrices and vectors of statistical criteria and Shannon entropy values.
   - 1.1. Formation of a gene expression matrix, the values of which are determined by using both DNA microarray experiments or an mRNA molecule sequencing method.
   - 1.2. For each gene expression profile, the maximum expression value, variance, and Shannon entropy are calculated and the vectors of the received values are formed.

II. Formation of subsets of various informativity level gene expression profiles using the Harrington desirability method.
   - 2.1. Calculation of the coefficients of the linear equation to transform the scales of quality criteria of GEPs into the scale of the dimensionless parameter Y, taking into account the boundary values of the corresponding criteria by Equation (3).
   - 2.2. For each value of the statistical criteria and Shannon's entropy, calculate the value of the parameter Y by Equation (4).
   - 2.3. Calculation of private desirabilities for each of the criteria by Equation (5) and the generalized desirability index by Equation (6).
   - 2.4. Formation of subsets of GEPs considering the level of their significance using the results obtained during the implementation of steps two and three of this procedure.

III. Application of the sample classification model using subsets of gene expressions with various significance levels.
   - 3.1. Choice and adjustment of the classifier and formation of data classification quality criteria.
   - 3.2. Classification of objects, the attributes of which are gene expression values with various significance levels.
   - 3.3. Calculation of data classification quality criteria and formation of vectors of these criteria values.

IV. Analysis of the obtained results.

4.1. Analysis of the classification results and the formation of a conclusion regarding the level of adequacy of the proposed model.

3.3. Formation of Criteria for Assessing the Quality of the Cluster Structure

The formation of clustering quality criteria for assessing the cluster structure of GEPs was conducted taking into account the principles of the objective clustering inductive technology (OCIT) [26,27], the application of which involves the assessment of the cluster structure based on the use of both the internal and external clustering quality criteria. In this instance, the final decision regarding the optimal cluster structure formation is performed based on the analysis of the values of the balance criterion, which contains both internal and external criteria as the components. The application of OCIT involves the division of GEPs into two equivalent subsets applying hybrid profile proximity metrics based on the evaluation of both the mutual information using various methods of Shannon's entropy evaluation and Pearson's chi-squared test, which is proposed in [2]. The equivalent subsets contain an equal number of pairwise similar GEPs.

The used hybrid proximity metric combines various methods of Shannon entropy calculation when the mutual information of gene expression profiles is evaluated and the Pearson's chi-squared test. As has been shown in [2], the object classification results differ when various GEP proximity metrics are applied during the mutually correlated gene expression data formation. In order to increase the objectivity of the distance between GEP evaluation, in this research, the authors proposed the stepwise procedure of proximity metric formation. In the first step, the mutual information based on the use of various methods of Shannon entropy and Pearson's chi-squared criterion values are evaluated for appropriate pair of GEPs. Then, the Harrington desirability function is applied to form the general proximity metric. A larger value of the general proximity metric, in this case, corresponds to a greater proximity between this pair of GEPs.

As a rule, the internal clustering quality criteria should take into account both the particularities of the gene expression profiles distribution within clusters relative to the respective cluster median (as the average value of all GEP is an abstraction and does not correspond to the real distribution of expression values in the profile) and the particularities of the clusters distribution (cluster medians) in the feature space. Within the framework of our research, the first component of the internal criterion was calculated as the square root of the sum of the squares of the distances from the GEPs to the median of the cluster where these profiles are allocated. Because within the framework of the research the distance is taken as the joint value of various types of mutual information [2], this component of the internal criterion can be presented as follows:

$$CW = \sqrt{\sum_{k=1}^{K} \sum_{i=1}^{m} MI(e_i M_k)^2} \quad (7)$$

where MI is a hybrid metric based on joint application of the mutual information evaluation and Fisher's chi-squared test (a larger value of this metric corresponds to a smaller distance between appropriate GEPs), e_i is a vector of i-th gene expression values, M_k is the median of the k-th cluster, m is the number of GEPs, and K is the number of clusters.

It is obvious that a larger value of the internal criterion component (7) corresponds to a smaller distance from the gene expression profiles to the median of the respective cluster (or a larger density of the GEPs distribution within the clusters).

The second component of the internal clustering quality criterion is calculated as the square root of the sum of squares of all values of hybrid proximity metric between all pairs of the cluster medians:

$$CB = \sqrt{\sum_{i=1}^{K-1} \sum_{j=i+1}^{K} MI(M_i M_j)^2} \quad (8)$$

In this case, it should be noted that better clustering corresponds to a smaller value of the distance between gene expression profiles in the clusters (larger value of criterion (7)) and a larger distance between the clusters (smaller value of criterion (8)). Taking into account this fact, the formula for calculating the internal clustering quality criterion can be presented as follows:

$$QC_{int} = \frac{CB}{K \cdot CW} \tag{9}$$

where the number of clusters K is, to some extent, a "normalized" coefficient, i.e., formula (9) determines the average density of the gene expression profiles and medians of the corresponding clusters distribution relative to one cluster. In accordance with this criterion, a smaller value of criterion (9) corresponds to better clustering.

The external criterion assumes the presence of two equivalent subsets of gene expression profiles. Whether using this criterion is reasonable is determined by the minimization of the reproducibility error, which is intrinsic to most existing data clustering algorithms. In other words, the clustering results obtained on the same data are not always repeated within the acceptable error when using another equivalent subset of data. It is obvious that if the nature of the gene expression profiles' distribution in equivalent data subsets is similar, the values of the internal criteria (9) obtained on these subsets should not differ significantly from each other. When the reproducibility error increases, the discrepancy between the values of the internal criteria will increase. Taking this fact into account, we propose to calculate the value of the external clustering quality criterion as the normalized difference of the values of the internal criteria obtained on the equivalent data subsets A and B:

$$QC_{ext} = \frac{|QC_{int}^{A} - QC_{int}^{B}|}{QC_{int}^{A} + QC_{int}^{B}} \tag{10}$$

A smaller value of this criterion corresponds to a smaller discrepancy of the clustering results obtained on equivalent data subsets. However, it should be noted that the values of the internal and external criteria can disagree with each other. High internal criterion values (unsatisfactory clustering on equivalent subsets) can be similar to each other. This fact can lead to a low external criterion value. In this case, it is reasonable to calculate the balance criterion, which contains as the components both the internal and external criteria. The balance criterion was calculated using Harrington's desirability function (Figure 2). One of the significant advantages of the Harrington method is that in is not necessary to normalize the input vectors with high objectivity when the output parameter calculation is performed. Normalization of the input data is performed automatically at the stage of the transformation of the input parameters' scales into the scale of the dimensionless output parameter Y, the values of which are varied within the range from -2 to 5. In this instance, the algorithm to calculate the balance criterion involves the following steps:

1. Evaluation of the coefficients a and b in the linear equations considering the boundary values of the appropriate criteria:

$$\begin{aligned} Y_{max} &= a - b \cdot QC_{min} \\ Y_{min} &= a - b \cdot QC_{max} \end{aligned} \tag{11}$$

   where $Y_{min} = -2$; $Y_{max} = 5$, QC_{min}, and QC_{max} are the minimal and maximal values of the internal and external criteria, respectively.
2. Transformation of both the internal and external criteria values into the value of the dimensionless parameter Y:

$$\begin{aligned} Y_{int}^{A} &= a_{int}^{A} - b_{int}^{A} \cdot QC_{int}^{A} \\ Y_{int}^{B} &= a_{int}^{B} - b_{int}^{B} \cdot QC_{int}^{B} \\ Y_{ext} &= a_{ext} - b_{ext} \cdot QC_{ext} \end{aligned} \tag{12}$$

3. Calculation of the private desirability values for each of the criteria:

$$\begin{aligned} d_{int}^{A} &= exp(-exp(-Y_{int}^{A})) \\ d_{int}^{B} &= exp(-exp(-Y_{int}^{B})) \\ d_{ext} &= exp(-exp(-Y_{ext})) \end{aligned} \tag{13}$$

4. Calculation of the balance criterion:

$$QC_{bal} = \sqrt[3]{d_{int}^{A} \cdot d_{int}^{B} \cdot d_{ext}} \tag{14}$$

A higher value of the balance criterion corresponds to better clustering by the group of the used criteria.

3.4. Validation of the Gene Expression Profiles Clustering Model

The main idea of the implementation of this stage is that clusters containing GEPs with a higher level of mutual expression should correspond to higher classification results of objects containing gene expression values of appropriate profiles as attributes. In our research, the assessment of the objects' classification quality was performed using traditional methods based on type I and type II errors with the use of a confusion matrix.

The following classification quality criteria were applied:

- Classification accuracy:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{15}$$

where:

  - TP is the true positive cases;
  - TN is the true negative cases;
  - FP is the false positive cases;
  - FN is the false negative cases.

- F-measure:

$$F = \frac{2 \cdot PR \cdot SE}{PR + SE} \tag{16}$$

where:

  - PR is defined as the ratio of correctly identified positive values to the total number of values identified as positive:

$$PR = \frac{TP}{TP + FP} \tag{17}$$

  - SE (sensitivity) is defined as the ratio of correctly identified positive values to the total number of true positive values:

$$SE = \frac{TP}{TP + FN} \tag{18}$$

Higher values of criteria (15) and (16) correspond to a higher quality of the investigated objects classification.

This research is a continuation of the previous of our research in [28], where we considered the various types of CNNs with various combinations of hyperparameters for the classification of objects, the attributes of which are the gene expression data. In this research, we have also considered the CNN's stability level to the noise component. The results of the research showed a superior performance of a 1D two-layer CNN, where 32 filters and a kernel size of eight were used. Maximal pooling, in this case, was two. This

structure of CNN was used within the current research. Of course, the sizes of filters were adapted considering the length of the input vector of gene expressions. These filter sizes were presented in the experimental part of the paper.

A 1D two-layer convolutional neural network was used as the classifier within the framework of our research. This choice is determined by our previous results, presented in [28]. In this research, we studied various CNN topologies with various combinations of hyperparameters to classify samples, the attributes of which are gene expression values. The results of the research have shown a better performance of 1D two-layer CNN, where 32 filters and kernel size 8 have been used. Maximal pooling, in this case, was 2. This structure of CNN was used within the implementation framework of the classification procedure. The size of the used filters was adapted considering the length of the gene expression vector during the simulation procedure implementation.

3.5. A Hybrid Inductive Model of GEP Clustering Based on the SOTA Clustering Algorithm

As was noted in the introduction, the self-organizing SOTA clustering algorithm [3] forms a binary topological tree based on the joint use of self-organizing Kohonen's maps [4] and the Fritzke algorithm of spatial cellular structure growth [5]. The application of this algorithm leads to an increase in the number of tree nodes in the higher object (gene expression profiles) density areas, while the concentration of objects in the area with lower density does not change at the current step of the algorithm application. An analysis of the simulation results presented by the authors of the SOTA clustering algorithm [3] has shown that the result of the algorithm operation (the structure of the formed topological tree) is determined by the boundary value of the relative change of the variation coefficient value or by the number of iterations on the one hand, and by the value of the parameter for correcting the weights of the sister's cell (parameters for correcting other cells are determined using the conditions: $\alpha_w = 2\alpha_p$; $\alpha_p = 5\alpha_s$).

Within the framework of our research, the boundary value of the relative change of the variation coefficient was taken as 0. Under this condition, the termination of the algorithm's operation was determined either by repeating the nature of the GEP distribution into clusters on two consecutive iterations or by reaching the boundary number of iterations. Determination of the optimal value of the parameter for correcting of the sister's cell weights was performed using the previously described OCIT.

The implementation of the inductive hybrid model of the SOTA algorithm assumes the following stages:

I. Data formation and initialization of the SOTA clustering algorithm.
   - 1.1. Formation of a matrix of gene expression data, where the rows are the investigated objects, and the columns are the genes whose expression values determine the state of the corresponding object.
   - 1.2. Removal of the non-informative GEPs in terms of statistical criteria and Shannon entropy in accordance with the technique presented in Section 3.2.
   - 1.3. Formation of metrics for assessing the proximity level of GEP. Considering the research results presented in [2], we have used a hybrid modified metric based on the integrated application of the hybrid metric of mutual information maximization and Pearson's chi-squared test.
   - 1.4. Formation of two equivalent subsets of GEPs A and B.
   - 1.5. Formation of clustering quality criteria to assess the quality of the cluster structure.
   - 1.6. Initialization of the algorithm. The boundary value of the relative change of the variation coefficient $E_{lim} = 0$ and the range and step variation of the parameter α_s for the correction of the sister's cell weight are set up.

II. GEP clustering and determination of the SOTA algorithm's optimal parameters.
   - 2.1. Initialization of the first value of the parameter for correcting the sister's cell weights: $\alpha_s = \alpha_{smin}$.

2.2. Applying the SOTA clustering algorithm to GEP contained in equivalent subsets A and B. Cluster structure formation.
2.3. If the number of clusters formed on equivalent subsets A and B is the same, then, the calculation of both the internal and external clustering quality criteria and the increase in the α_s value is $\alpha_s = \alpha_s + d\alpha_s$. Otherwise, increase the α_s without calculating the criteria.
2.4. If $\alpha_s \leq \alpha_{smax}$, then, go to step 2.2 of this procedure. Otherwise, use the calculation of the balance criterion values.
2.5. Analysis of the obtained results and fixing the values of the α_s parameter, which correspond to the maximum values of the balance criterion.
2.6. Application of the SOTA clustering algorithm with optimal value of α_s parameters to the full set of GEPs created at step 1.2 of this procedure and formation of a cluster structure.

III. Applying the convolutional neural network (CNN) to gene expression data in the formed clusters.
3.1. Pre-processing of gene expression data in the formed clusters by adding profiles with zero expression to obtain the required number of profiles for the correct application of CNN filters.
3.2. Applying the CNN to gene expression data in the formed clusters. Calculation of the objects classification quality criteria.
3.3. Analysis of the obtained results. The formation of subsets of differentially expressed and mutually correlated GEPs that correspond to the maximum values of both the balance criterion and the classification quality criteria.

3.6. A Hybrid Inductive Model of GEP Clustering Based on the Spectral Clustering Algorithm

Implementation of spectral clustering algorithm within the framework of OCIT to form the clusters of differentially expressed and mutually correlated GEPs assumes the following stages:

I. Data formation and initialization of the model parameters.
1.1. Formation of the gene expression matrix $e_{i,j}$, $i = \overline{1,n}$, $j = \overline{1,m}$, where n is the number of investigated objects and m is the number of genes.
1.2. Formation of a GEP proximity metric.
1.3. Formation of vectors and functions for calculation of clustering quality criteria: internal, external, and balanced.
1.4. Division of the set of GEPs into two equivalent subsets A and B.
1.5. Calculation of similarity matrix for all pairs of GEPs contained in the equivalent subsets.
1.6. Initialization of the range for the number of clusters variation: k_{min}, k_{max}

II. Clustering of gene expression profiles. Calculation of clustering quality criteria.
2.1. Initialization of the initial number of clusters $k = k_{min}$.
2.2. Clustering of GEPs allocated in equivalent data subsets A and B and formation of cluster structures.
2.3. Calculation of internal and external clustering quality criteria.
2.4. If $k < k_{max}$, then increase the number of clusters by one and go to step 2.2 of this procedure. Otherwise, the calculation of the balance criterion using the obtained values of the internal and external criteria.
2.5. Analysis of the received results, fixation of optimal clusterings which correspond to the maximum values of the balance criterion.

III. Classification of gene expression data.
3.1. Formation of subsets of gene expression data allocated in the formed clusters for further application as input data in a CNN.

3.2. Application of CNN to gene expression data contained in each cluster at a current hierarchical level of the cluster structure. Calculation of the classification quality criteria.
3.3. Analysis of the obtained results. Formation of subsets of differentially expressed and mutually correlated GEPs.

4. Experiment, Results and Discussion

4.1. Experimental Data

We used the gene expression dataset GSE19188 [29] from the freely available database GEO (the Geo Expression Omnibus) [30] as the experimental data. This dataset contains the gene expression values for 156 patients examined for lung cancer. Of them, 65 patients were identified as healthy and 91 were identified with lung cancer tumors. Initially, this data contained 54,675 genes.

4.2. Formation of the Subsets of GEPs Considering the Significance Level Using the Harrington Desirability Method

Figure 4 shows the block charts of private desirabilities and the general desirability index, which determine the gene expression profiles' significance level and were calculated for the maximum values of gene expressions for all samples, variance, and Shannon entropy.

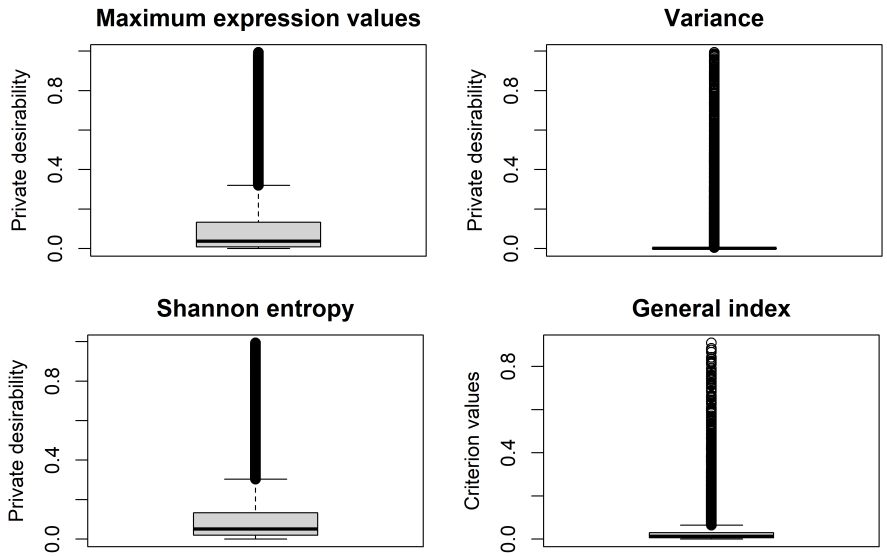

Figure 4. Box plots of private desirabilities and the general desirability index, which determines the gene expression profiles' significance level.

An analysis of the resulting charts allows us to conclude based on the values of private desirability and the value of the general desirability index, most profiles (54,087 from 54,675) are identified as profiles with a very low significance level. In this instance, 15 genes were identified as having a very high significance level, 38 were identified as genes with a high significance level, and 142 and 393 profiles were identified as genes with medium and low significance levels, respectively. However, it should be noted that informative genes are not mutually expressed. This fact is confirmed by the classification results of samples containing, as the attributes, gene expression values allocated in appropriate subsets (Table 1).

Table 1. Simulation results regarding the classification of objects based on gene expression data of various significance level when applying the Harrington desirability method.

Genes Significance	Classification Quality Criteria				
	Accuracy %	Significance	Specificity	F-Score	MCC
Very high	87.1	0.909	0.850	0.890	0.736
High	96.8	1	0.947	0.894	0.935
Medium	91.9	0.862	0.970	0.890	0.841
Low	96.8	1	0.947	0.894	0.935

Indeed, at first glance, the presented results are not logical. However, it should be noted that this fact only shows that the application of the proposed method can allow us only to divide the initial set of gene expression profiles into informative and non-informative ones according to the appropriate quantitative criteria. Moreover, the number of informative genes is determined by the threshold value of the general desirability index, which can vary depending on the nature of the experimental data distribution.

Within the framework of the current research, considering the particularities of the GEPs' general desirability index (GI) values distribution, the boundary value of the GI dividing the gene expression profiles into informative and non-informative ones was chosen at the level of 0.04. In this case, 10,000 gene expression profiles were selected. To assess the model adequacy by applying a classifier, this set of genes was divided into three subsets: $0.04 \leq GI < 0.2$: medium significance (9162 genes); $0.2 \leq GI < 0.37$: high significance (513 genes); $GI \geq 0.37$: very high significance (332 genes). The classifier was applied to the full set of gene expression data (10,000 profiles) and to the gene expression profiles in the formed subsets. The classification results are presented in Table 2.

Table 2. Simulation results regarding the classification of objects based on gene expression data with various significance levels when applying the Harrington desirability method.

Genes Significance	Classification Quality Criteria				
	Accuracy %	Significance	Specificity	F-Score	MCC
Full set of GEP	91.9	1	0.947	0.894	0.935
Very high	91.9	0.862	0.970	0.890	0.841
High	96.8	1	0.947	0.894	0.935
Medium	91.9	1	0.878	0.958	0.842

As it can be seen, in all cases, the value of the classification quality criteria is quite high but not the maximum. Moreover, the classification results of objects containing a large number of gene expression values as attributes (10,000 and 9162) are the same. This fact can be explained by the presence of a large number of differentially expressed gene expression profiles that can be identified as noise. The highest quality of the samples classification is achieved when using 513 genes of high significance level, which can be explained by a significantly smaller number of genes on the one hand and the presence of a larger number of genes designated as informative by the applied criteria on the other hand.

The presented research demonstrates the advisability of using the proposed technique for preprocessing gene expression data to reasonably remove uninformative genes based on measures of GEPs significance. The application of the proposed method allows us to form a subset of the most informative gene expression profiles according to the set of statistical and entropy criteria. In this instance, the number of genes is determined by the threshold value of the general desirability index, which is determined empirically during

the simulation process taking into account the target value of the number of genes, which needs to be allocated for further research.

4.3. Practical Implementation of the Hybrid Inductive Model of GEP Clustering Based on the SOTA Clustering Algorithm

The simulation results regarding the evaluation of the optimal value of the α_s parameter for correcting the sister's cell weights are presented in Figure 5.

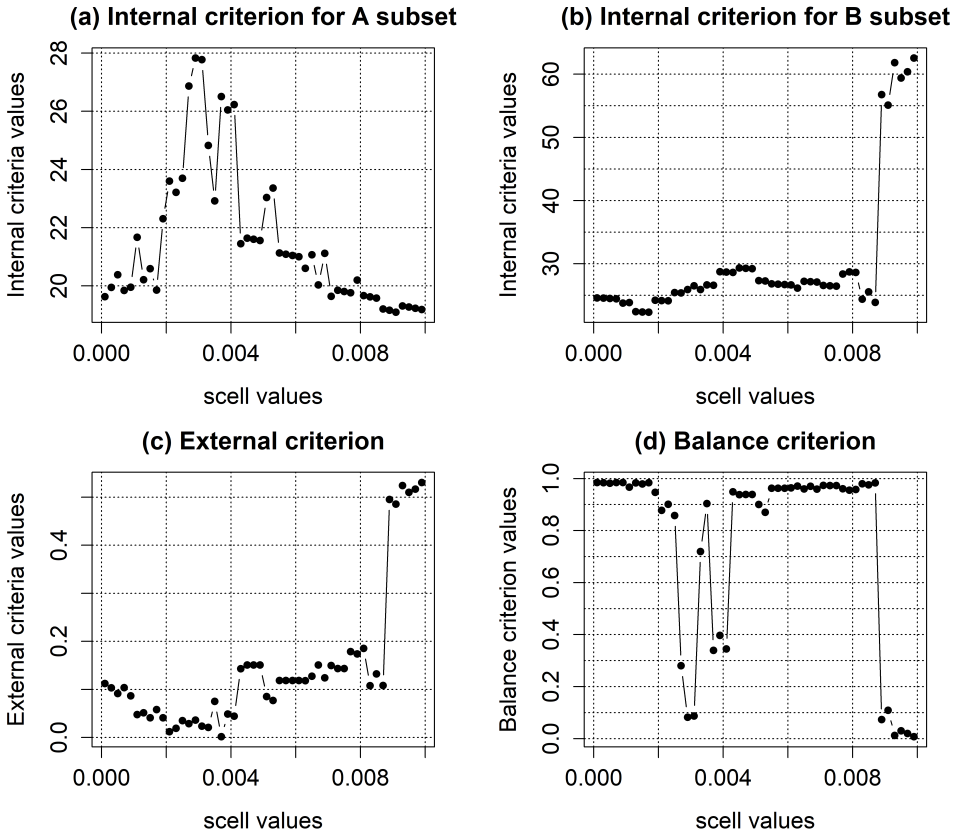

Figure 5. The simulation results for determining the optimal parameters of the SOTA clustering algorithm based on the application of inductive methods of complex systems analysis.

The value of this parameter varied within the range from 0.0001 to 0.01 with a step of 0.0002. The simulation was performed based on the use of the *R* software environment using the functions of *clValid* package [31]. The *sota()* function, in this instance, provides the possibility of implementing the SOTA clustering algorithm using both Euclidean and correlation distance metrics. Taking into account the high dimensionality of GEPs, the formation of a cluster structure during the algorithm operation was conducted using the correlation metric, and the calculation of the internal clustering quality criteria was performed using a modified hybrid metric [2].

In accordance with the first part of the algorithm, presented in Section 3.5, at the first stage, the selected subset of GEP (10,000 profiles) was divided into two equivalent subsets A and B, containing the same number of pairwise similar objects (according to the modified hybrid metric). Then, at each step of the SOTA algorithm implementation, internal clustering quality criteria were calculated on equivalent subsets of gene expression profiles (in the cases, if the number of clusters is the same). Figure 5a,b show charts of the internal criteria values versus the α_s parameter of the SOTA algorithm, while, according to these criteria, the optimal clustering corresponds to their minimum values. As can be seen in the obtained charts, the values of these criteria, calculated on the subsets A and B, to some extent disagree with each other. Figure 5c shows the chart of external criteria calculated using the appropriate internal criteria. The analysis of these charts also does

not allow us to unambiguously determine the optimal value of the α_s parameter (by the minimum value). Figure 5d shows a chart of the balance criterion, the analysis of which allows us to allocate the ranges of the parameter variation, which achieves the maximal values of the balance criterion on the one hand and stable clustering on the other hand. A more careful analysis of the obtained results showed that the balance criterion reaches its maximum value (0.985) at the fourth step of this procedure implementation. In this case, the low value of the internal criteria also confirms the fact that at this step, the nature of the GEPs distribution in the clusters corresponds to the optimal cluster structure. It should be noted that in all cases (at each iterative step), the data were divided into two clusters. The algorithm stopped when repeating the configuration of the cluster structures on two consecutive iterations. Thus, the value $\alpha_s = 0.0007$ was chosen as an optimal one for further simulation stages.

At the second stage of the model's implementation, the SOTA clustering algorithm with a determined parameter was applied to the full set of GEPs (10,000). The result of this step implementation formed two clusters: the first contained 6020 profiles, and the second contained 3980 profiles.

The classification procedure was performed using a one-dimensional two-layer convolutional neural network (CNN), the efficiency and robustness of which for gene expression data classification was proven in [28]. For the correct application of the CNN, the first cluster was supplemented with profiles with zero expression to obtain a total of 6050 profiles. In this case, filters (50×121) and (25×242) were applied on the first and second convolutional layers, respectively. The second cluster was supplemented to obtain a total of 4000 profiles, while the filters had dimensions (50×80) and (25×160) on the first and the second layers. The simulation results regarding the application of CNN to gene expression data contained in the formed clusters are shown in Figures 6 and 7.

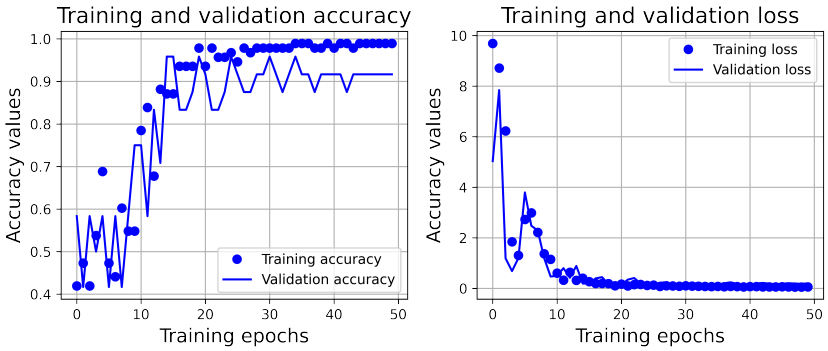

Figure 6. Charts of objects classification accuracy and the loss function value during the network training on the data in the first cluster.

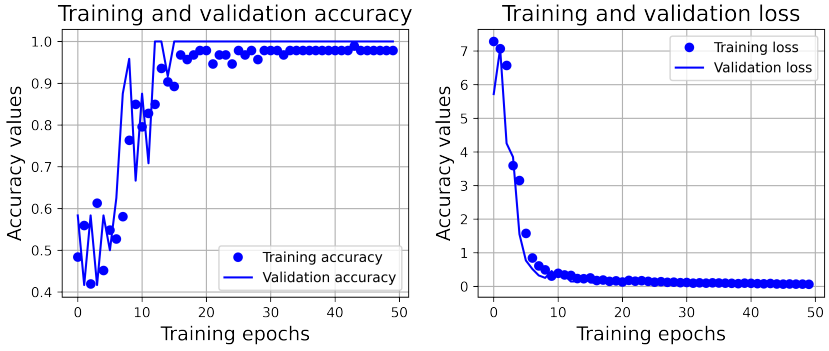

Figure 7. Charts of objects classification accuracy and the loss function value during the network training on the data in the second cluster.

The accuracy of object classification using the corresponding test data subsets was 95% for the first cluster and 97% for the objects of the second one, while from 39 objects, 37 and 38 were correctly identified in the first and second cases, respectively.

At the next stage, the SOTA algorithm was applied stepwise to the GEPs contained in the appropriate clusters, followed by the classification of the objects based on the data obtained clusters by applying CNN with adapted filters. The simulation results are presented in Table 3.

Table 3. Simulation results regarding the step-by-step application of the hybrid inductive model of GEP clustering based on the SOTA clustering algorithm and CNN.

Stage	Number of Genes	Filter	F-Measure		Accuracy %	Loss
			Cluster 1	Cluster 2		
I	6020	50 × 121 25 × 242	0.93	0.96	95	0.152
	3980	50 × 80 25 × 160	**0.96**	**0.98**	**97**	**0.146**
II	**3011**	50 × 61 25 × 122	**0.96**	**0.98**	**97**	**0.197**
	3009	50 × 61 25 × 122	**0.96**	**0.98**	**97**	**0.165**
	1934	50 × 39 25 × 78	0.92	0.96	95	0.227
	2046	50 × 41 25 × 82	0.89	0.94	92	0.205
III	**1639**	40 × 41 20 × 82	**0.96**	**0.98**	**97**	**0.156**
	1372	40 × 35 20 × 70	0.95	0.96	95	0.199
	1519	40 × 38 20 × 76	**0.96**	**0.98**	**97**	**0.123**
	1490	30 × 50 15 × 100	0.89	0.94	92	0.193

Figure 8 shows a dendrogram of the GEP distribution into clusters with values of the investigated object classification accuracy.

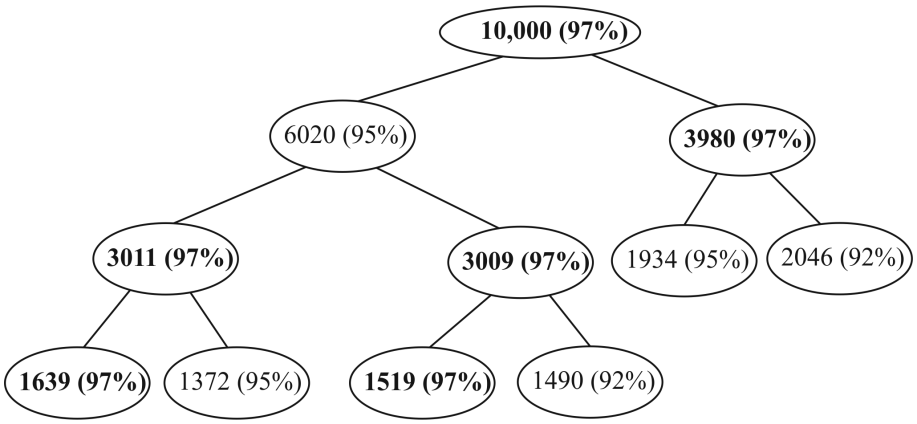

Figure 8. Dendrogram of the GEP distribution into clusters with the classification accuracy of the objects containing as attributes the values of gene expressions of the appropriate clusters.

The analysis of the obtained results allows us to conclude that the joint application of the SOTA clustering algorithm and the CNN allows us to identify clusters of GEP that can allow us to identify the objects that contain the value of gene expressions that are localized in the corresponding clusters as attributes with high accuracy. As can be seen, in the first

stage, 10,000 GEPs were divided into two clusters, while for the second (smaller cluster), a high classification accuracy was achieved on the test data subset; 38 of 39 objects were correctly identified in this case. For the larger cluster, the classification accuracy was lower in terms of various criteria, with two objects out of 39 being falsely identified.

However, it should be noted that the further division of the GEP contained in the smaller clusters into subsets consisting of 1934 and 2046 GEP worsened the classification accuracy, i.e., the division of the smaller cluster into subclusters is not reasonable in this case. Another conclusion can be reached in the case of dividing a larger cluster (6020 GEP) into subclusters. Dividing this cluster into two subsets (3011 and 3009 GEP) increases the classification accuracy of objects containing the values of gene expressions as attributes allocated in these clusters. Their further division into smaller subclusters allows us to allocate a subset of GEPs that are more informative according to the classification quality criteria for each branch.

4.4. Practical Implementation of the Hybrid Inductive Model of GEP Clustering Based on the Spectral Clustering Algorithm

Figure 9 shows the simulation results regarding the application of the proposed hybrid inductive clustering model based on the hierarchical spectral clustering algorithm. The number of clusters varied within the range from 2 to 10 during the simulation process's implementation.

As we can see, the internal and external clustering quality criteria, in some cases, contradict each other. This fact confirms that it is reasonable to apply the balance criterion which contains the internal and corresponding external criteria as the components. The balance criterion reaches its maximum value when dividing the set of GEP into three clusters. Moreover, according to the internal criterion, the optimal clustering corresponds to three clusters for subset A and two clusters for subset B. In accordance with the external criterion, the optimal clustering corresponds to the three-cluster structure.

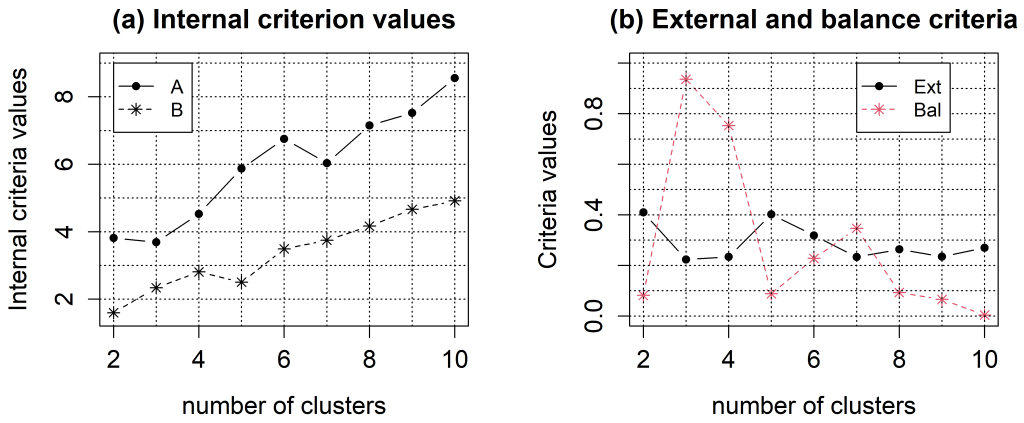

Figure 9. The simulation results regarding the practical implementation of the hierarchical spectral clustering algorithm based on inductive methods of complex systems analysis.

The next step of the model implementation is an application of a CNN to gene expression data in the allocated clusters. To confirm the previously described conclusions regarding the effectiveness of the optimal cluster structure formation based on clustering quality criteria, we studied cluster structures containing two, three, and four clusters. The simulation results are presented in Table 4. The analysis of the obtained results confirms the fact that the three-cluster structure is optimal in terms of both the classification accuracy and the value of the loss function when the CNN was training. However, it should be noted that the classification accuracy is quite high in all cases, which can be explained by the high efficiency of the CNN for this type of data and its resistance to the noise component. The classification accuracy of the studied objects was evaluated on the test data subset, which

was not used for the network's training. At the same time, in the case of a three-cluster structure, when using the third cluster containing 4964 genes, 100% classification accuracy was achieved with the minimum value of the loss function. When applying other clusters of this cluster structure, out of 39 objects that make up the test subset of gene expression data, 38 were identified correctly.

Table 4. Simulation results regarding the step-by-step application of the hybrid inductive model of GEP clustering based on the spectral clustering algorithm and CNN.

Stage and Model Parameters		Cluster 1	Cluster 2	Cluster 3	Cluster 4
Two-cluster structure	Number of genes	4074	4926	–	–
	Filter	60 × 68 30 × 136	50 × 119 25 × 238	–	–
	Accuracy, %	95	97	–	–
	Loss	0.254	0.067	–	–
Three-cluster structure	Number of genes	2487	2549	4964	–
	Filter	50 × 50 25 × 100	50 × 51 25 × 102	60 × 83 30 × 166	–
	Accuracy, %	**97**	**97**	**100**	–
	Loss	**0.141**	**0.123**	**0.058**	–
Four-cluster structure	Number of genes	1615	2779	4715	891
	Filter	30 × 54 15 × 108	50 × 56 25 × 112	50 × 95 25 × 190	50 × 18 25 × 36
	Accuracy, %	97	97	97	95
	Loss	0.169	0.142	0.189	0.295

The comparison analysis with other research in this subject area [1] allows us to draw conclusions on the performance of the proposed technique. So, in most cases presented in the review [1] a high classification accuracy is achieved when using a small number of genes. The proposed method allows us to form the subsets of differentially and co-expressed gene expression profiles, which contribute to the high value of the investigated samples' classification accuracy. Moreover, the allocated subsets of GEPs can be used as a next step in the hybrid modeling of disease diagnosis, such as various types of cancer, Alzheimer's, Parkinson's, etc., in order to take a more objective solution regarding the state of the patient using the object classification results obtained on various subsets of the allocated genes.

In our minds, the presented research creates the conditions for increasing the objectivity of the diagnosis of the complex disease by making a compromise decision based on the results of the classification of objects containing informative differentially expressed and mutually correlated gene expression data, which are allocated in different clusters.

5. Conclusions

In this manuscript, we have presented the results of research regarding the formation of subsets of differentially and co-expressed gene expression profiles based on the joint use of inductive clustering algorithms and a one-dimensional two-layer convolutional neural network. In the first stage of this procedure implementation, we removed the non-informative genes using statistical criteria and Shannon entropy with the Harrington desirability function. The number of genes decreased from 54,675 to 10,000 as a result of this step's implementation. Then, we sequentially applied the SOTA and spectral clustering algorithms implemented within the framework of objective clustering inductive technology, the application of which assumes the division of the initial data into two equivalent subsets with parallel clustering of the data allocated in these subsets within the range of the respective algorithm parameters variation. As a result, the internal, external, and balance criteria were calculated in the case if the number of clusters is the same in various clustering. The optimal clustering corresponds, in this instance, to the maximum

values of the balance clustering quality criterion. To confirm the obtained clustering results, we have applied, at the final step of this procedure's implementation, a one-dimensional two-layer convolutional neural network to identify the objects, the attributes of which are gene expressions in the allocated clusters. The simulation results have shown that in the case of the SOTA clustering application, we can allocate various clusters of differentially and co-expressed gene expression profiles which correspond to high values of both the clustering and classification quality criteria. However, the procedure of the SOTA clustering algorithm application within the framework of OCIT has larger complexity in comparison with the application of the spectral clustering algorithm. Moreover, to our mind, the application of spectral clustering within the framework of the proposed model is more effective in terms of both the studied objects classification accuracy and the complexity of the algorithm's implementation.

In our minds, the conducted research creates the conditions for increasing the objectivity of the formation of subsets of differentially and co-expressed gene expression profiles for subsequent applications. The future directions of the authors' research include the application of the proposed technique within the hybrid model of various disease diagnostics and gene regulatory network reconstruction on the basis of allocated subsets of differentially and co-expressed genes.

Author Contributions: The individual contributions of the authors are the following: Conceptualization, formal analysis, resources, writing—review and editing: S.B., L.Y.-D., I.L. and J.Š.; methodology, software (R programming), validation, statistical analysis, and investigation, writing—original draft preparation: S.B.; results, visualization: L.Y.-D. and I.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data applied during the presented research are available by the link: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi, accessed on 5 November 2022.

Acknowledgments: We thank team of the researchers from Cell Biology, Erasmus University Medical Center, Rotterdam, The Netherlands Hou J, Aerts J, den Hamer B, et al. who have performed a genome-wide gene expression analysis on a cohort of 91 patients with tumors and 65 adjacent normal lung tissue samples.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

GEP	Gene expression profile
SOTA	Self-organizing tree algorithm
CNN	Convolutional neural network
ROC	Receiver operating characteristic
AUC	Area under curve
GRN	Gene regulatory network
DLBCL	Diffuse large B-cell lymphoma
EPSO	Extended particle swarm optimization model
PSO	Particle swarm optimization model
OCIT	Objective clustering inductive technology

References

1. Almugren, N.; Alshamlan, H. A survey on hybrid feature selection methods in microarray gene expression data for cancer classification. *IEEE Access* **2019**, *7*, 78533–78548. [CrossRef]
2. Babichev. S.; Yasinska-Damri, L.; Liakh, I.; Durnyak, B. Comparison analysis of gene expression profiles proximity metrics. *Symmetry* **2021**, *13*, 1812. [CrossRef]
3. Dorazo, J.; Carazo, J.M. Phylogenetic reconstruction using an unsupervised growing neural network that adopts the topology of a phylogenetic tree. *J. Mol. Evol.* **1997**, *44*, 226–234.
4. Kohonen, T. *Self-Organizing Maps*, 3rd ed.; Springer: Berlin/Heidelberg, Germany, 2001; 502p.
5. Fritzke, B. Growing Cell Structures. A Self-Organizing Network for Unsupervised and Supervised Learning. *Neural Netw.* **1994**, *7*, 1441–1420. [CrossRef]
6. Von Luxburg, U. A tutorial on spectral clustering. *Stat. Comput.* **2007**, *17*, 395–416. [CrossRef]
7. Romero, M.; Ramírez, Ó.; Finke, J.; Rocha, C. Supervised Gene Function Prediction Using Spectral Clustering on Gene Co-expression Networks. *Stud. Comput. Intell.* **2022**, *1016*, 652–663.
8. Yu, K.; Xie, W.; Wang, L.; Zhang, S.; Li, W. Determination of biomarkers from microarray data using graph neural network and spectral clustering. *Sci. Rep.* **2021**, *11*, 23828. [CrossRef]
9. Liu, J.; Ge, S.; Cheng, Y.; Wang, X. Multi-View Spectral Clustering Based on Multi-Smooth Representation Fusion for Cancer Subtype Prediction. *Front. Genet.* **2021**, *12*, 718915. [CrossRef]
10. Taghizadeh, E.; Heydarheydari, S.; Saberi, A.; JafarpoorNesheli, S.; Rezaeijo, S.M. Breast cancer prediction with transcriptome profiling using feature selection and machine learning methods. *BMC Bioinform.* **2022**, *23*, 10. [CrossRef]
11. Chawla, S.; Rockstroh, A.; Lehman, M.; Ratther, E.; Jain, A.; Anand, A.; Gupta, A.; Bhattacharya, N.; Poonia, S.; Rai, P.; et al. Gene expression based inference of cancer drug sensitivity. *Nat. Commun.* **2022**, *13*, 5680. [CrossRef]
12. Cui, L.; Acharya, S.; Mishra, S.; Pan, Y.; Huang, J.Z. MMCo-Clus-An Evolutionary Co-clustering Algorithm for Gene Selection. *IEEE Trans. Knowl. Data Eng.* **2022**, *34*, 4371–4384. [CrossRef]
13. Zuo, C.L.; Wu, H.Y.; Zhu, M. An Improved Method of Extracting and Classifying DLBCL Information Genes. In Proceedings of the 6th International Conference on Biomedical Engineering and Applications, Hangzhou, China, 13–15 May 2022; pp.104–109.
14. Al-Shammary, D.; Albukhnefis, A.L.; Alsaeedi, A.H.; Al-Asfoor, M. Extended particle swarm optimization for feature selection of high-dimensional biomedical data. *Concurr. Comput. Pract. Exp.* **2022**, *34*, e6776. [CrossRef]
15. Alshamlan, H.; Badr, G.; Alohali, Y. A hybrid gene selection algorithm for cancer classification using microarray gene expression profiling. *Biomed. Res. Int.* **2018**, *2015*, 604910. [CrossRef]
16. Moradi, P.; Gholampour, M. A hybrid particle swarm optimization for feature subset selection by integrating a novel local search strategy. *Appl. Soft Comput.* **2016**, *43*, 117–130. [CrossRef]
17. Jain, I.; Jain, V.K.; Jain, R. Correlation feature selection based improved-binary particle swarm optimization for gene selection and cancer classification. *Appl. Soft Comput.* **2018**, *62*, 203–215. [CrossRef]
18. Pashaei, E.; Ozen, M.; Aydin, N. Gene selection and classification approach for microarray data based on random forest ranking and BBHA. In Proceedings of the 2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), Las Vegas, NV, USA, 24–27 February 2016; pp. 308–311.
19. Shreem, S.S.; Abdullah, S.; Nazri, M.Z. A Hybrid feature selection algorithm using symmetrical uncertainty and a harmony search algorithm. *Int. J. Syst. Sci.* **2016**, *47*, 1312–1329. [CrossRef]
20. Djellali, H.; Guessoum, S.; Ghoualmi-Zine, N.; Layachi, S. Fast correlation based filter combined with genetic algorithm and particle swarm on feature selection. In Proceedings of the 5th International Conference on Electrical Engineering-Boumerdes (ICEE-B), Boumerdes, Algeria, 29–31 October 2017; pp. 1–6.
21. Yu, Y.; Liang, S.; Samali, B.; Nguyen, T.N.; Zhai, C.; Li, J.; Xie, X. Torsional capacity evaluation of RC beams using an improved bird swarm algorithm optimised 2D convolutionalneural network. *Eng. Struct.* **2022**, *273*, 115066. [CrossRef]
22. Yu, Y.; Samali, B.; Rashidi, M.; Mohammadi, M.; Nguyen, T.N.; Zhang, G. Vision-based concrete crack detection using a hybrid framework considering noise effect. *J. Build. Eng.* **2022**, *61*, 105246. [CrossRef]
23. Liakh, I.; Babichev, S.; Durnyak, B.; Gado, I. Formation of Subsets of Co-expressed Gene Expression Profiles Based on Joint Use of Fuzzy Inference System,Statistical Criteria and Shannon Entropy. *Lect. Notes Data Eng. Commun. Technol.* **2023**, *149*, 25–41.
24. Midi, H.; Aziz, N.A. Augmented desirability function for multiple responses with contaminated data. *J. Eng. Appl. Sci.* **2018**, *13*, 6626–6633.
25. Iwański, M.; Mazurek, G.; Buczyński, P.; Iwański, M.M. Effects of hydraulic binder composition on the arheological characteristics of recycled mixtures with foamed bitumen for full depth reclamation. *Constr. Build. Mater.* **2022**, *330*, 127274. [CrossRef]
26. Madala, H.R.; Ivakhnenko, A.G. Clusterization and Recognition. In *Inductive Learning Algorithms for Complex Systems Modeling*; CRC Press: New York, NY, USA, 2019; 380p.
27. Babichev, S.; Durnyak, B.; Pikh, I.; Senkivskyy, V. An Evaluation of the Objective Clustering Inductive Technology Effectiveness Implemented Using Density-Based and Agglomerative Hierarchical Clustering Algorithms. *Adv. Intell. Syst. Comput.* **2020**, *1020*, 532–553.
28. Yasinska-Damri, L.; Babichev, S.; Durnyak, B.; Goncharenko, T. Application of Convolutional Neural Network for Gene Expression Data Classification. *Lect. Notes Data Eng. Commun. Technol.* **2023**, *149*, 3–24.

29. Hou, J.; Aerts, J.; den Hamer, B.; Jcken, W.; den Bakker, M.; Riegman, P.; der Leest, C.; Spek, P.; Foekens, J.A.; Hoogsteden, H.C.; et al. Gene expression-based classification of non-small cell lung carcinomas and survival prediction. *PLoS ONE* **2010**, *5*, e10312. [CrossRef] [PubMed]
30. Gene Expression Omnibus. Available online: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi (accessed on 5 November 2022).
31. Brock, G.; Pihur, V.; Datta, S.; Datta, S. clValid: An R Package for Cluster Validation. *J. Stat. Softw.* **2008**, *25*, 1–22. [CrossRef]