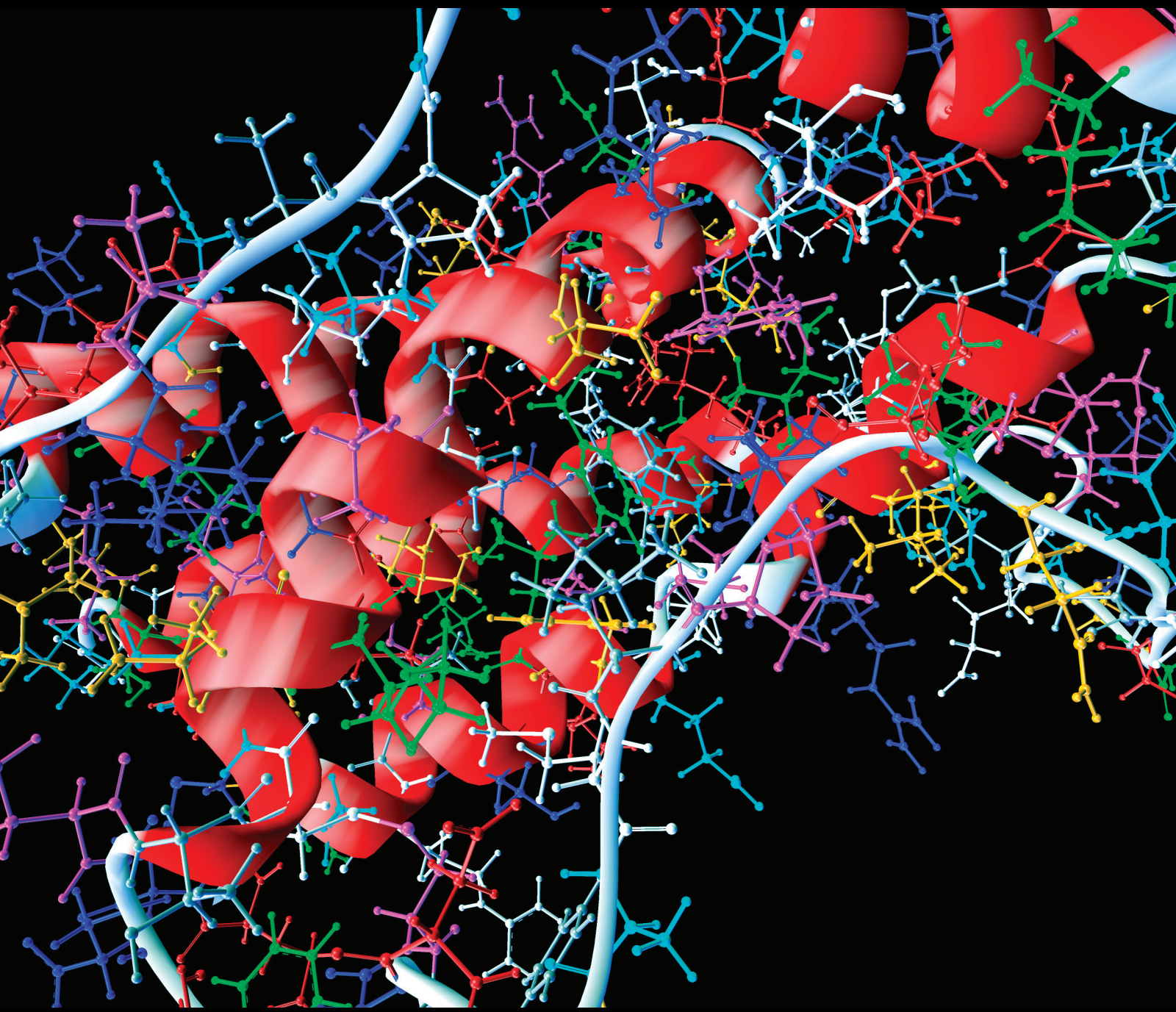
Computational and Mathematical Methods in Medicine

Soft Computing for Analysis of Biomedical Data

Lead Guest Editor: Federico Divina
Guest Editors: Miguel García-Torres, Ting Hu, and Christian E. Schaerer



Soft Computing for Analysis of Biomedical Data

Computational and Mathematical Methods in Medicine

Soft Computing for Analysis of Biomedical Data

Lead Guest Editor: Federico Divina
Guest Editors: Miguel García-Torres, Ting Hu,
and Christian E. Schaerer

Editorial Board

Contents

Hindawi
Computational and Mathematical Methods in Medicine
Volume 2018, Article ID 3902484, 2 pages
https://doi.org/10.1155/2018/3902484

Editorial
Soft Computing for Analysis of Biomedical Data

Federico Divina,[1] **Miguel García-Torres**,[1] **Ting Hu**,[2] **and Christian E. Schaerer**[3]

[1]*Universidad Pablo de Olavide, Seville, Spain*
[2]*Memorial University of Newfoundland, St. John's, Canada*
[3]*National University of Asuncion, San Lorenzo, Paraguay*

Correspondence should be addressed to Federico Divina; fdiv@upo.es

Received 10 October 2018; Accepted 10 October 2018; Published 15 November 2018

Soft computing (SC) techniques can be used to tackle problems characterized by imprecision, uncertainty, and partial truth to achieve tractability and robustness at a low computational cost.

These features represent the main differences between SC and hard computing techniques and provide SC strategies with the ability to deal with ambiguous situations like imprecision and uncertainty. For this reason, SC techniques can obtain approximate solutions to problems which have no known methods to compute an exact solution. The main SC paradigms include fuzzy systems, evolutionary computation, artificial neural computing, metaheuristics, and swarm intelligence.

Those features render SC particularly suitable for analyzing medical data, which is typically characterized by imprecision and the presence of noise. Moreover, SC techniques allow easily integrating human knowledge, which can help achieve better solutions. Biomedical data may be of different nature: texts, images, signals, and so forth, which typically contain a high presence of noise.

The overall aim of this special issue was to compile the latest research and development, up-to-date issues, and challenges in the field of SC and its applications to biomedical data.

Seventeen articles were submitted to this special issue, and finally, six original research articles were accepted and are contained in this special issue.

In the article from Y. X. Meng et al., the authors used functional gene enrichment analysis to identify genetic markers and pathways that are associated with neonatal sepsis. A case-control population based dataset was collected for the purpose of the study, and subsequent statistical tests and coexpression network analysis were employed. A set of 7 key signaling pathways and 7 hub genes were identified with high potential associated with the disease risk.

Biomedical text mining was the subject of the article from A. Onan. In particular, the author proposed an efficient multiple classifier approach to text categorization based on swarm-optimized latent Dirichlet allocation and diversity-based ensemble pruning. The proposed technique was applied to five biomedical text benchmarks. Results showed that the proposed technique outperformed other state-of-the-art classification algorithms, as well as various ensemble learning and ensemble pruning methods.

In article from Y. Li et al., different machine learning techniques were applied to the classification of diabetes follow-up data. In particular, after having applied feature selection and imbalanced processing techniques, the authors applied Support Vector Machine, Decision Trees, Adaboost, and Bagging to the resulting data. Results showed that Adaboost was the most successful technique for classifying this kind of data. Following these results, an analysis of the most relevant features was also conducted.

In their work, B. Yang and coworkers address the problem of studying the neural activity under cognitive reappraisal on simultaneous EEG (electroencephalography)-fMRI (functional magnetic resonance imaging) data. For such a purpose, the authors propose an effective fusion framework that uses a Kernel-based Canonical Correlation Analysis (KCCA). Results show that the proposed EEG-fMRI fusion approach provides an effective way to study the

neural activities of cognitive reappraisal with high spatio-temporal resolution.

A novel tool for the optimization of the structure of gene networks is proposed in the article from F. Gómez-Vela et al. In particular, the tool is called GeSOp, and it represents a new computational method for optimizing the structure of gene networks. Such a method performs a pruning of irrelevant information in the input network to facilitate the interpretation of the biological knowledge that comprises the network. To do this, GeSOp relies on a greedy heuristic algorithm that selects only the most relevant relationships and helps to identify the Hubs in the network. The performance of the method was tested in different data sets with satisfactory results in all cases.

Finally, D. Liu et al. propose a cosine similarity measure between hybrid intuitionistic fuzzy sets. In order to study this measure, the authors apply it to medical diagnosis and discuss its relevant properties. Then, based on the proposal, the authors present a decision method for medical diagnosis so that a patient can be diagnosed with the disease according to the values of the cosine similarity measure. This measure is compared with other existing similarity measures. Results show the feasibility and effectiveness of the Cosine similarity measure.

We can conclude that this special issue presents different works using different machine learning techniques applied to biomedical data. As a consequence, this issue can prove to be a valuable tool to gain insights on the state of the art of such a field.

Conflicts of Interest

The editors declare that they have no conflicts of interest.

Acknowledgments

The editors would like to thank all authors who submitted their research to this special issue, as well as all reviewers for their valuable contribution.

Federico Divina
Miguel García-Torres
Ting Hu
Christian E. Schaerer

Hindawi
Computational and Mathematical Methods in Medicine
Volume 2018, Article ID 3146873, 7 pages
https://doi.org/10.1155/2018/3146873

Research Article

Cosine Similarity Measure between Hybrid Intuitionistic Fuzzy Sets and Its Application in Medical Diagnosis

Donghai Liu,[1] **Xiaohong Chen**,[2] **and Dan Peng**[1]

[1]*Department of Mathematics, Hunan University of Science and Technology, Xiangtan, China*
[2]*Hunan University of Commerce, Changsha, China*

Correspondence should be addressed to Donghai Liu; donghailiu@126.com

Received 24 January 2018; Revised 11 June 2018; Accepted 12 September 2018; Published 17 October 2018

Guest Editor: Miguel García-Torres

In this paper, a cosine similarity measure between hybrid intuitionistic fuzzy sets is proposed. The aim of the paper is to investigate the cosine similarity measure with hybrid intuitionistic fuzzy information and apply it to medical diagnosis. Firstly, we construct the cosine similarity measure between hybrid intuitionistic fuzzy sets, and the relevant properties are also discussed. In order to obtain a reasonable evaluation in group decision, the weight of experts under different attributes is determined by the projection of individual decision information on the ideal decision information, where the ideal decision information is the average values of each expert's evaluation. Furthermore, we propose a decision method for medical diagnosis based on the cosine similarity measure between hybrid intuitionistic fuzzy sets, and the patient can be diagnosed with the disease according to the values of proposed cosine similarity measure. Finally, an example is given to illustrate feasibility and effectiveness of the proposed cosine similarity measure, which is also compared with the existing similarity measures.

1. Introduction

A similarity measure is an important tool for determining the degree of similarity between two objects in many fields, such as pattern recognition, medical diagnosis, and so on. Many similarity measures have been introduced [1–8]. Among them, some similarity measures of intuitionistic fuzzy sets (IFSs) have been proposed. For example, Li and Cheng [3] proposed a similarity measure between IFSs and applied it to pattern recognition. Huang and Yang [2] defined the similarity measure between IFSs based on the Hausdorff distance and used it to calculate the degree of similarity between IFSs. Nguuen [9] proposed a new knowledge-based similarity measure between IFSs and applied it to pattern recognition. However, due to the complexity and uncertainty of the decision-making environment, the membership degree and nonmembership degree of IFS need to be expressed by interval rather than the numerical value. Motivated by this, Atanassov and Gargov [10] introduced the concept of interval-valued intuitionistic fuzzy set (IVIFS), which is a generalization of IFS. Xu [11] proposed some distance and similarity measures between IVIFSs and applied them to pattern recognition.

On the other hand, the cosine similarity measure based on Bhattacharyya distance was first proposed in Bhattacharyya [12]. Ye [7] proposed a cosine similarity measure for IFSs (C_{IFS}) and applied it to pattern recognition. Furthermore, Ye [13] proposed the cosine similarity measure for IVIFSs (C_{IVIFS}) and applied it to group decision-making problems. However, in the complex group decision-making problem, it is difficult to use a single value to express the alternative under all attributes. Because some attributes might be represented by IFSs, but other attributes are suitable to be represented by IVIFSs. At this time, the people should use hybrid intuitionistic fuzzy set to make a decision. However, the existing methods can not deal with the hybrid fuzzy information. As far as we know, no people studied the cosine similarity measure between hybrid IFSs. Motivated by this, we will introduce the cosine similarity measure with hybrid intuitionistic fuzzy information (C_{HIFS}) in this paper. This generalization makes the C_{HIFS} measure includes C_{IFS} measure and C_{IVIFS} measure as particular case.

In addition, applying the C_{HIFS} measure to group decision-making problems is very interesting. For example, Zhou and Wahab [14] use transmissibility incorporated

with cosine similarity measure to investigate the structural damage detection. Furthermore, Zhou et al. [15] apply transmissibility function with distance measure to separate the intact patterns apart from the damaged pattern. In group decision-making problems, the weight of the experts under different attributes can be obtained by using the projection of individual decision information on the ideal decision information. Then, we aggregate all individual decisions into a collective one and apply the proposed cosine similarity measure between hybrid intuitionistic fuzzy sets to medical diagnosis.

The rest of the paper is organized as follows. In Section 2, we review the cosine similarity measure for IFSs and IVIFSs. In Section 3, we propose the C_{HIFS} measure, some properties are also analyzed. In Section 4, we propose a decision method for medical diagnosis based on the cosine similarity measure between hybrid intuitionistic fuzzy sets. In Section 5, an example is given to illustrate the feasibility and effectiveness of the proposed C_{HIFS} measure. Finally, the conclusion and further research are discussed in Section 6.

2. Preliminaries

Throughout this paper, let $X = \{x_1, x_2, ..., x_n\}$ be a finite universal set. In this section, we briefly review the IFSs and IVIFSs, the cosine similarity measure between IFSs, and the cosine similarity measure between IVIFSs.

2.1. Intuitionistic Fuzzy Set

Definition 1. Let X be a fixed set, an intuitionistic fuzzy set (IFS) A in X is defined as:

$$A = \left\{\left(x_j, \mu_A(x_j), \nu_A(x_j)\right| \ x_j \in X\right)\}, \tag{1}$$

where the functions $\mu_A(x_j)$ and $\nu_A(x_j)$ represent the membership degree and nonmembership degree of the element x_j to the set A, respectively, such that $0 \le \mu_A(x_j) + \nu_A(x_j) \le 1 \forall x_j \in X$.

The intuitionistic fuzzy index $\pi_A(x_j) = 1 - \mu_A(x_j) - \nu_A(x_j)$, and we have $0 \le \pi_A(x_j) \le 1$. For example, $A = (0.4, 0.3)$ is an intuitionistic fuzzy number, and $\pi_A = 0.3$. The space of membership degree of IFS is shown in Figure 1.

In particular, when X has only one element, the IFS $A = \{(x_j, \mu_A(x_j), \nu_A(x_j)|x_j \in X)\}$ is reduced to $A = (\mu_A(x_j), \nu_A(x_j))$, which we call it an intuitionistic fuzzy number (IFN).

For any two IFSs $A = (x_j, \mu_A(x_j), \nu_A(x_j))$ and $B = (x_j, \mu_B(x_j), \nu_B(x_j))$, the following operations are true [16]:

(1) $A + B = (\mu_A(x_j) + \mu_B(x_j) - \mu_A(x_j)\mu_B(x_j), \nu_A(x_j)\nu_B(x_j))$

(2) $\lambda A = (1 - (1 - \mu_A(x_j))^{\lambda}, (\nu_A(x_j))^{\lambda}), \lambda > 0$

(3) $A = B \quad if\ \mu_A(x_j) = \mu_B(x_j), \nu_A(x_j) = \nu_B(x_j)$

The results of the operations $A + B$ and λA are still IFSs.

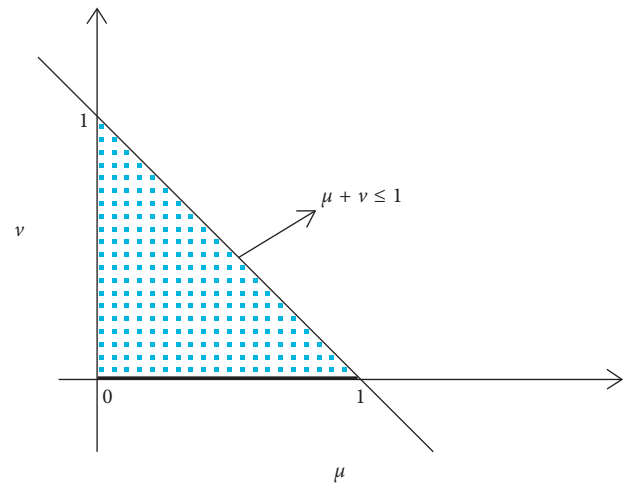


Figure 1: The membership degree of IFS.

2.2. Interval-Valued Intuitionistic Fuzzy Set

Definition 2. Let X be a fixed set, an interval-valued intuitionistic fuzzy set $\widetilde{A}$ is defined as follows:

$$\widetilde{A} = \left\{\left(x_j, \left[\mu_{\widetilde{A}L}(x_j), \mu_{\widetilde{A}U}(x_j)\right], \left[\nu_{\widetilde{A}L}(x_j), \nu_{\widetilde{A}U}(x_j)\right]\right)\right| \ x_j \in X\}, \tag{2}$$

where

$$\mu_{\widetilde{A}L}(x_j) \ge 0, \nu_{\widetilde{A}L}(x_j) \ge 0; 0 \le \mu_{\widetilde{A}U}(x_j) + \nu_{\widetilde{A}U}(x_j) \le 1, \quad \forall x_j \in X. \tag{3}$$

The interval-valued intuitionistic fuzzy index is defined as $\pi_{\widetilde{A}}(x_j) = [\pi_{\widetilde{A}L}(x_j), \pi_{\widetilde{A}U}(x_j)]$, where

$$\begin{aligned} \pi_{\widetilde{A}L}(x_j) &= 1 - \mu_{\widetilde{A}U}(x_j) - \nu_{\widetilde{A}U}(x_j), \pi_{\widetilde{A}U}(x_j) \\ &= 1 - \mu_{\widetilde{A}L}(x_j) - \nu_{\widetilde{A}L}(x_j). \end{aligned} \tag{4}$$

For example, $\widetilde{A} = ([0.3, 0.4], [0.1, 0.4])$ is an interval-valued intuitionistic fuzzy number, the fuzzy index $\pi_{\widetilde{A}}(x_j) = [0.2, 0.6]$.

Remark 1. If $\mu_{\widetilde{A}L}(x_j) = \mu_{\widetilde{A}U}(x_j), \nu_{\widetilde{A}L}(x_j) = \nu_{\widetilde{A}U}(x_j)$, then the interval-valued intuitionistic fuzzy set is reduced to intuitionistic fuzzy set.

When the set X has only one element, the IVIFS $\widetilde{A} = \{(x_j, [\mu_{\widetilde{A}L}(x_j), \mu_{\widetilde{A}U}(x_j)], [\nu_{\widetilde{A}L}(x_j), \nu_{\widetilde{A}U}(x_j)])|x_j \in X\}$ is reduced to $\widetilde{A} = ([\mu_{\widetilde{A}L}(x_j), \mu_{\widetilde{A}U}(x_j)], [\nu_{\widetilde{A}L}(x_j), \nu_{\widetilde{A}U}(x_j)])$, which we call it an interval-valued intuitionistic fuzzy number(IVIFN).

Let $\widetilde{A} = (x_j, [\mu_{\widetilde{A}L}(x_j), \mu_{\widetilde{A}U}(x_j)], [\nu_{\widetilde{A}L}(x_j), \nu_{\widetilde{A}U}(x_j)])$ and $\widetilde{B} = (x_j, [\mu_{\widetilde{B}L}(x_j), \mu_{\widetilde{B}U}(x_j)], [\nu_{\widetilde{B}L}(x_j), \nu_{\widetilde{B}U}(x_j)])$ be two IVIFSs, the following operations are true [17]:

(1) $\widetilde{A} + \widetilde{B} = ([\mu_{\widetilde{A}L}(x_j) + \mu_{\widetilde{B}L}(x_j) - \mu_{\widetilde{A}L}(x_j)\mu_{\widetilde{B}L}(x_j), \mu_{\widetilde{A}U}(x_j) + \mu_{\widetilde{B}U}(x_j) - \mu_{\widetilde{A}U}(x_j)\mu_{\widetilde{B}U}(x_j)], ([\nu_{\widetilde{A}L}(x_j)\nu_{\widetilde{B}L}(x_j), \nu_{\widetilde{A}U}(x_j)\nu_{\widetilde{B}U}(x_j)])$

(2) $\lambda\widetilde{A} = ([1 - (1 - \mu_{\widetilde{A}L}(x_j))^{\lambda}, 1 - (1 - \mu_{\widetilde{A}U}(x_j))^{\lambda}], [(\nu_{\widetilde{A}L}(x_j))^{\lambda}, (\nu_{\widetilde{A}U}(x_j))^{\lambda}]), \lambda > 0$

(3) $\widetilde{A} = \widetilde{B}$ if $\mu_{\widetilde{AL}}(x_j) = \mu_{\widetilde{BL}}(x_j), \mu_{\widetilde{AU}}(x_j) = \mu_{\widetilde{BU}}(x_j)$, $\nu_{\widetilde{AL}}(x_j) = \nu_{\widetilde{BL}}(x_j)$ and $\nu_{\widetilde{AU}}(x_j) = \nu_{\widetilde{BU}}(x_j)$

2.3. Cosine Similarity Measures for IFSs or IVIFSs.

Definition 3 (Ye [7]). Let $A = (\mu_A(x_j), \nu_A(x_j))$ and $B = (\mu_B(x_j), \nu_B(x_j))$ be two IFSs in X, the cosine similarity measure between A and B is defined as follows:

$$C_{\mathrm{IFS}}(A,B) = \frac{1}{n}\sum_{i=1}^{n}\frac{\mu_A(x_i)\mu_B(x_i) + \nu_A(x_i)\nu_B(x_i)}{\sqrt{\mu_A^2(x_i) + \nu_A^2(x_i)}\sqrt{\mu_B^2(x_i) + \nu_B^2(x_i)}}. \tag{5}$$

The cosine similarity measure between two IFSs A and B satisfies the following properties:

(1) $0 \le C_{\mathrm{IFS}}(A,B) \le 1$

(2) $C_{\mathrm{IFS}}(A,B) = C_{\mathrm{IFS}}(B,A)$

(3) $C_{\mathrm{IFS}}(A,B) = 1$ if $A = B$

Definition 4 (Ye [13]). Let $\widetilde{A} = ([\mu_{\widetilde{AL}}(x_j), \mu_{\widetilde{AU}}(x_j)], [\nu_{\widetilde{AL}}(x_j), \nu_{\widetilde{AU}}(x_j)])$ and $\widetilde{B} = ([\mu_{\widetilde{BL}}(x_j), \mu_{\widetilde{BU}}(x_j)]), [\nu_{\widetilde{BL}}(x_j), \nu_{\widetilde{BU}}(x_j)]$ be two IVIFSs in X, the cosine similarity measure between two IVIFSs $\widetilde{A}$ and $\widetilde{B}$ is defined as follows:

$$C_{\mathrm{IVIFS}}(\widetilde{A},\widetilde{B}) = \frac{1}{n}\sum_{i=1}^{n}\frac{\mu_{\widetilde{AL}}(x_i)\mu_{\widetilde{BL}}(x_i) + \mu_{\widetilde{AU}}(x_i)\mu_{\widetilde{BU}}(x_i) + \nu_{\widetilde{AL}}(x_i)\nu_{\widetilde{BL}}(x_i) + \nu_{\widetilde{AU}}(x_i)\nu_{\widetilde{BU}}(x_i) + \pi_{\widetilde{AL}}(x_i)\pi_{\widetilde{BL}}(x_i) + \pi_{\widetilde{AU}}(x_i)\pi_{\widetilde{BU}}(x_i)}{\sqrt{\mu_{\widetilde{AL}}^2(x_i) + \mu_{\widetilde{AU}}^2(x_i) + \nu_{\widetilde{AL}}^2(x_i) + \nu_{\widetilde{AU}}^2(x_i) + \pi_{\widetilde{AL}}^2(x_i) + \pi_{\widetilde{AU}}^2(x_i)} \cdot |H|}, \tag{6}$$

where

$$|H| = \sqrt{\mu_{\widetilde{BL}}^2(x_i) + \mu_{\widetilde{BU}}^2(x_i) + \nu_{\widetilde{BL}}^2(x_i) + \nu_{\widetilde{BU}}^2(x_i) + \pi_{\widetilde{BL}}^2(x_i) + \pi_{\widetilde{BU}}^2(x_i)}. \tag{7}$$

The cosine similarity measure between two IVIFs $\widetilde{A}$ and $\widetilde{B}$ satisfies the following properties:

(1) $0 \le C_{\mathrm{IVIFS}}(\widetilde{A},\widetilde{B}) \le 1$

(2) $C_{\mathrm{IVIFS}}(\widetilde{A},\widetilde{B}) = C_{\mathrm{IVIFS}}(\widetilde{B},\widetilde{A})$

(3) $C_{\mathrm{IVIFS}}(\widetilde{A},\widetilde{B}) = 1$ if $\widetilde{A} = \widetilde{B}$

3. Cosine Similarity Measure with Hybrid Intuitionistic Fuzzy Information

In this section, we will propose the cosine similarity measure with hybrid intuitionistic fuzzy information (C_{HIFS}) and some properties are also discussed.

Definition 5. Let A be fuzzy set (FS) in $X = \{x_1, x_2, \ldots, x_n\}$, I and II be two subsets of the attribute set X, such that $\mathrm{I} \cup \mathrm{II} = X, \mathrm{I} \cap \mathrm{II} = \phi$. If $x_j \in \mathrm{I}$, the value of fuzzy set A is characterized by IFSs, if $x_j \in \mathrm{II}$, the values of fuzzy set A is characterized by IVIFSs, then A is called hybrid intuitionistic fuzzy sets (HIFSs).

Definition 6. Let $A = \{(x_j, \mu_A(x_j), \nu_A(x_j)) | x_j \in X\}$ and $B = \{(x_j, \mu_B(x_j), \nu_B(x_j)) | x_j \in X\}$ be two hybrid intuitionistic fuzzy sets, such that if the same attributes $x_j \in \mathrm{I}$, $(\mu_A(x_j), \nu_A(x_j))$, and $(\mu_B(x_j), \nu_B(x_j))$ are IFSs, if the same attribute $x_j \in \mathrm{II}$, $([\mu_{AL}(x_j), \mu_{AU}(x_j)], [\nu_{AL}(x_j), \nu_{AU}(x_j)])$, and $([\mu_{BL}(x_j), \mu_{BU}(x_j)], [\nu_{BL}(x_j), \nu_{BU}(x_j)])$ are IVIFSs, which we call A and B the same type hybrid intuitionistic fuzzy sets.

Definition 7. Suppose A and B are the same type hybrid intuitionistic fuzzy sets, that is, if $x_j \in \mathrm{I}$, $(\mu_A(x_j), \nu_A(x_j))$, and $(\mu_B(x_j), \nu_B(x_j))$ are IFSs, if the same attribute $x_j \in \mathrm{II}$, $([\mu_{AL}(x_j), \mu_{AU}(x_j)], [\nu_{AL}(x_j), \nu_{AU}(x_j)])$, and $([\mu_{BL}(x_j), \mu_{BU}(x_j)], [\nu_{BL}(x_j), \nu_{BU}(x_j)])$ are IVIFSs, then the cosine similarity measure between hybrid intuitionistic fuzzy sets A and B is defined as follows:

$$\begin{aligned} C_{\mathrm{HIFS}}(A,B) = \frac{1}{n}\Bigg[& \sum_{x_i \in \mathrm{I}} \frac{\mu_A(x_i)\mu_B(x_i) + \nu_A(x_i)\nu_B(x_i) + \pi_A(x_i)\pi_B(x_i)}{\sqrt{\mu_A^2(x_i) + \nu_A^2(x_i) + \pi_A^2(x_i)}\sqrt{\mu_B^2(x_i) + \nu_B^2(x_i) + \pi_B^2(x_i)}} \\ & + \sum_{x_i \in \mathrm{II}} \frac{\mu_{AL}(x_i)\mu_{BL}(x_i) + \mu_{AU}(x_i)\mu_{BU}(x_i) + \nu_{AL}(x_i)\nu_{BL}(x_i) + \nu_{AU}(x_i)\nu_{BU}(x_i) + \pi_{AL}(x_i)\pi_{BL}(x_i) + \pi_{AU}(x_i)\pi_{BU}(x_i)}{\sqrt{\mu_{AL}^2(x_i) + \mu_{AU}^2(x_i) + \nu_{AL}^2(x_i) + \nu_{AU}^2(x_i) + \pi_{AL}^2(x_i) + \pi_{AU}^2(x_i)} \cdot \sqrt{\mu_{BL}^2(x_i) + \mu_{BU}^2(x_i) + \nu_{BL}^2(x_i) + \nu_{BU}^2(x_i) + \pi_{BL}^2(x_i) + \pi_{BU}^2(x_i)}} \Bigg]. \end{aligned} \tag{8}$$

Remark 2. If $\mathrm{I} = \phi$, then C_{HIFS} measure is reduced to C_{IVIFS} measure.

Remark 3. If $\mathrm{II} = \phi$, then C_{HIFS} measure is reduced to C_{IFS} measure.

Theorem 1. *The cosine similarity measure between two hybrid intuitionistic fuzzy sets A and B satisfies the following properties:*

(1) $0 \le C_{\mathrm{HIFS}}(A,B) \le 1$

(2) $C_{\mathrm{HIFS}}(A,B) = C_{\mathrm{HIFS}}(B,A)$

(3) $C_{\mathrm{HIFS}}(A,B) = 1$ if $A = B$

Proof

(1) It is obvious that the property (1) is true according to the cosine value in $[0,1]$

(2) Because the multiplication of numbers satisfies the commutative law, if the positions of A and B are exchanged in the computation of cosine measure, the result values will not change, so the property (3) is true.

(3) If $A = B, x_i \in \mathrm{I}$, we have $\mu_A(x_i) = \mu_B(x_i)$ and $\nu_A(x_i) = \nu_B(x_i)$.

If $A = B, x_i \in \mathrm{II}$, we have $\mu_{AL}(x_i) = \mu_{BL}(x_i), \mu_{AU}(x_i) = \mu_{BU}(x_i)$, $\nu_{AL}(x_i) = \nu_{BL}(x_i)$, and $\nu_{AU}(x_i) = \nu_{BU}(x_i)$, then $C_{\mathrm{HIFS}}(A,B) = 1$ is obvious obtained.

4. Multiple-Attribute Group Decision-Making with the Cosine Similarity Measure between Hybrid Intuitionstic Fuzzy Sets

In this section, we will apply the C_{HIFS} measure between hybrid intuitionistic fuzzy sets to medical diagnosis. The C_{HIFS} measure can be applied in many situations, such as pattern recognition, medical diagnosis, and so on. The main motivation for considering this model is that the representation of the decision information is very complex. We need several doctors correctly to evaluate the symptoms of the disease. The doctor usually provides his/her preferences for symptoms with IFSs or IVIFSs. Suppose that doctors are good at different diagnostic skills, we can obtain the weights of doctors based on the projection of individual decision on the ideal decision; then, all individual diagnosis decisions are aggregated into a collective one. At last, we apply the C_{HIFS} measure to medical diagnosis.

In a given pathology, suppose that a set of symptoms $S = (s_1, s_2, \ldots, s_n)$, a set of diagnoses $A = (A_1, A_2, \ldots, A_m)$ and a set of medical experts $E = (e_1, e_2, \ldots, e_t)$. Assume that a patient has all the symptoms, which can be represented by the hybrid intuition fuzzy set $\widetilde{B}$, our aim is to diagnose what kind of diagnoses the patient $\widetilde{B}$ belong to.

In order to solve this problem, we first introduce some relevant concepts.

Definition 8. Let $A' = (a_{ij})_{m\times n} = (\mu_{A_i}(x_j), \nu_{A_i}(x_j))_{m\times n}$ be a decision matrix, I and II be two subsets of the attribute set $X = \{x_j \mid j = 1, 2, \ldots, n\}$, such that $\mathrm{I} \cup \mathrm{II} = X$ and $\mathrm{I} \cap \mathrm{II} = \varnothing$. If the attribute $x_j \in \mathrm{I}$, then the evaluation values a_{ij} are IFSs, if the attribute $x_j \in \mathrm{II}$, then the evaluation values a_{ij} are IVIFSs. In this case, A' is called a hybrid intuitionistic fuzzy matrix.

Definition 9. Let $A' = (A_1, A_2, \ldots, A_m)^T$ and $B' = (B_1, B_2, \ldots, B_m)^T$ be two hybrid intuitionistic fuzzy matrices, where $A_i = (a_{i1}, a_{i2}, \ldots, a_{in})$ and $B_i = (b_{i1}, b_{i2}, \ldots, b_{in})$ $(i = 1, 2, \ldots m)$, if they satisfy the following conditions:

(1) $a_{ij} = (\mu_{A_i}(x_j), \nu_{A_i}(x_j))$ about x_j in A_i if and only if $b_{ij} = (\mu_{Bi}(x_j), \nu_{Bi}(x_j))$ about x_j in B_i

(2) $a_i = ([\mu_{A_iL}(x_j), \mu_{A_iU}(x_j)], [\nu_{A_iL}(x_j), \nu_{A_iU}(x_j)])$ about x_j in A_i if and only if $b_{ij} = ([\mu_{B_iL}(x_j), \mu_{B_iU}(x_j)], [\nu_{B_iL}(x_j), \nu_{B_iU}(x_j)])$ about x_j in B_i, then A_i and B_i are the same type vector, A' and B' are the same type hybrid intuitionistic fuzzy matrices.

Now, suppose that a set of diagnoses $A' = (A_1, A_2, \ldots, A_m)$, where A_i is represented by IFS $A_i = (x_j, \mu_{A_i}(x_j), \nu_{A_i}(x_j))$ or IVIFS $A_i = (x_j, [\mu_{A_iL}(x_j), \mu_{A_iU}(x_j)], [\nu_{A_iL}(x_j), \nu_{A_iU}(x_j)])$ $(i = 1, 2, \ldots, m)$. We should diagnose what kind of disease the patient $\widetilde{B}$ belongs to. Furthermore, assume that the patient $\widetilde{B}$ is represented by the same type intuitionistic fuzzy set as A_i. In the following, we will present the method for application of C_{HIFS} measure to medical diagnosis, which involves the following steps:

Step 1. Each medical expert provides his/her individual decision matrix about the relation between the diagnosis and the symptoms.

Step 2. According to the expert's diagnostic decision matrix $R_k = (r_{ij}^k)_{m\times n}$, the ideal decision information should be close to the opinions of most doctors; then, we define the ideal relation $R^* = (r_{ij}^*)_{m\times n}$ between the diagnosis A_i $(i = 1, 2, \ldots, m)$ and the symptom s_j $(j = 1, 2, \ldots, n)$ as follows:

$$R^* = \left(r_{ij}^*\right)_{m\times n} = \begin{array}{c} \\ A_1 \\ A_2 \\ \vdots \\ A_m \end{array} \begin{array}{c} \begin{array}{cccc} s_1 & s_2 & \cdots & s_n \end{array} \\ \begin{pmatrix} r_{11}^* & r_{12}^* & \cdots & r_{1n}^* \\ r_{21}^* & r_{22}^* & \cdots & r_{2n}^* \\ \vdots & \vdots & \cdots & \vdots \\ r_{m1}^* & r_{m2}^* & \cdots & r_{mn}^* \end{pmatrix} \end{array}. \tag{9}$$

$$\text{If } j \in \mathrm{I}, r_{ij}^* = \left(\mu_{ij}^*, \nu_{ij}^*\right) = \left(\frac{1}{t}\sum_{k=1}^{t}\mu_{ij}^{(k)}, \frac{1}{t}\sum_{k=1}^{t}\nu_{ij}^{(k)}\right). \tag{10}$$

$$\text{If } j \in II, r_{ij}^* = \left(\widetilde{\mu}_{ij}^*, \widetilde{\nu}_{ij}^*\right) = \left(\left[\frac{1}{t}\sum_{k=1}^{t}\mu_{ijL}^{(k)}, \frac{1}{t}\sum_{k=1}^{t}\mu_{ijU}^{(k)}\right], \left[\frac{1}{t}\sum_{k=1}^{t}\nu_{ijL}^{(k)}, \frac{1}{t}\sum_{k=1}^{t}\nu_{ijU}^{(k)}\right]\right). \tag{11}$$

Step 3. Medical experts may give unreasonable assessments when they encounter unfamiliar symptoms. So, it is not very reasonable to assume that each expert has equal weights. In order to obtain a reasonable evaluation, the weights of medical experts under different attributes are obtained by the projection of the individual evaluation on the ideal evaluation r_{ij}^*. The greater the weight of the expert is, the closer the evaluation value is to the ideal evaluation.

The projection of each decision on the ideal decision is given by

$$\text{if } j \in \text{I}, \Pr j_{r_{ij}^*}^{r_{ij}^{(k)}} = \frac{\mu_{ij}^{(k)}\mu_{ij}^* + \nu_{ij}^{(k)}\nu_{ij}^* + \pi_{ij}^{(k)}\pi_{ij}^*}{\sqrt{\mu_{ij}^{*2} + \nu_{ij}^{*2} + \pi_{ij}^{*2}}}, \tag{12}$$

$$\text{if } j \in \text{II}, \Pr j_{r_{ij}^*}^{r_{ij}^{(k)}} = \frac{\mu_{ijL}^{(k)}\mu_{ijL}^* + \mu_{ijU}^{(k)}\mu_{ijU}^* + \nu_{ijL}^{(k)}\nu_{ijL}^* + \nu_{ijU}^{(k)}\nu_{ijU}^* + \pi_{ijL}^{(k)}\pi_{ijL}^* + \pi_{ijU}^{(k)}\pi_{ijU}^*}{\sqrt{\mu_{ijL}^{*2} + \mu_{ijU}^{*2} + \nu_{ijL}^{*2} + \nu_{ijU}^{*2} + \pi_{ijL}^{*2} + \pi_{ijU}^{*2}}}. \tag{13}$$

Then the weight of medical expert's evaluation on different symptoms can be defined as

$$w_{ij}^{(k)} = \frac{\Pr j_{r_{ij}^*} r_{ij}^{(k)}}{\sum_{k=1}^{t} \Pr j_{r_{ij}^*} r_{ij}^{(k)}}, \quad k = 1, 2, \ldots, t; \quad i = 1, 2, \ldots, m;\ j = 1, 2, \ldots, n. \tag{14}$$

Step 4. According to the recognition principle of maximum degree of cosine similarity measure, the process of diagnosis $\widetilde{B}$ to A_k is derived by $k = \underset{1 \le i \le m}{\text{Max}} (C_{\text{HIFS}}(A_i, \widetilde{B}))$.

5. Numerical Example

In this section, the proposed cosine similarity measure between hybrid IFSs is applied in medical diagnosis to demonstrate its effectiveness.

5.1. Illustration of the Cosine Similarity Measures for Hybrid IFSs. Assume that a set of diagnosis $A = \{A_1 (\text{viral fever}), A_2 (\text{typhoid}), A_3 (\text{stomach problem}), A_4 (\text{chest problem})\}$ and a set of symptoms $S = \{s_1 (\text{temperature}), s_2 (\text{stomach pain}), s_3 (\text{cough}), s_4 (\text{chest pain})\}$. Suppose a patient has all the symptoms, which can be represented by the following hybrid intuitionistic fuzzy information (data obtained through a survey of doctors):

$$\widetilde{B} = \{(s_1, 0.5, 0.4), (s_2, 0.6, 0.2), (s_3, [0.5, 0.6], [0.2, 0.3]), (s_4, 0.4, 0.2)\}. \tag{15}$$

There are three medical experts evaluate each diagnosis with all the symptoms, which are represented by the hybrid IFSs, the results are shown in Tables 1–3.

By step 3 in Section 4, applying (12)–(14), we can calculate the weights of each medical expert for the diagnosis with respect to different symptoms, which are obtained in Tables 5–7.

From the previous formula $C_{\text{HIFS}}(A', B)$, we can calculate the cosine similarity measure between $\widetilde{A}_i$ $(i = 1, 2, 3, 4)$ and $\widetilde{B}$ as follows:

$$\begin{aligned} C_{\text{HIFS}}(\widetilde{A}_1, \widetilde{B}) &= 0.9674, \\ C_{HIFS}(\widetilde{A}_2, \widetilde{B}) &= 0.9477, \\ C_{\text{HIFS}}(\widetilde{A}_3, \widetilde{B}) &= 0.9140, \\ C_{HIFS}(\widetilde{A}_4, \widetilde{B}) &= 0.9511. \end{aligned} \tag{16}$$

We can conclude that the diagnosis of the patient $\widetilde{B}$ is viral fever (A_1).

5.2. Comparison Analysis. In this subsection, the existing cosine similarity measure is used to compare with the same numerical example. In the numerical example, the decision information is represented with hybrid IFS, we can transform it into a unified form. For example, the relation between the diagnosis and the symptoms under the attribute s_3 of experts is IVIFSs, and if we use the cosine similarity measure C_{IFS} proposed by Ye [7] to calculate the numerical example, we should convert the corresponding IVIFSs to IFS according to the midpoints of IVIFSs. For example, $([0.4, 0.6], [0.1, 0.3])$ can be converted to (0.5). Then using the cosine similarity measure C_{IFS} proposed by Ye [7], we can obtain the corresponding cosine similarity measure values: $C_{\text{IFS}}(A_1, \widetilde{B}) = 0.9691$, $C_{\text{IFS}}(A_2, \widetilde{B}) = 0.9546$, $C_{\text{IFS}}(A_3, \widetilde{B}) = 0.9377$, and $C_{IFS}(A_4, \widetilde{B}) = 0.9586$. That is to say, the diagnosis of the patient $\widetilde{B}$ is still the viral fever A_1. The proposed cosine similarity between hybrid IFS in this paper produces the same results as the existing methods. This means that the proposed method is feasible and effective, and it has some advantages in solving multiple criteria decision-making problems. On one hand, the method is more convenient to make decision for decision makers, who can express their preferences over the decision information by IFS or IVIFS simultaneously. On the other hand, because the information conversion will be lost in decision-making process, there are no information conversions between IFSs and IVIFSs in this model, the alternatives will be ranked directly based on the original decision information.

6. Conclusion

The paper proposed the cosine similarity measure between hybrid intuitionistic fuzzy sets, and the proposed method

Table 1: The relation between the diagnosis and the symptoms—expert 1.

	s_1	s_2	s_3	s_4
A_1	$(0.5, 0.4)$	$(0.5, 0.3)$	$([0.4, 0.6], [0.1, 0.3])$	$(0.4, 0.4)$
A_2	$(0.7, 0.3)$	$(0.7, 0.2)$	$([0.3, 0.5], [0.4, 0.5])$	$(0.6, 0.2)$
A_3	$(0.8, 0.1)$	$(0.6, 0.4)$	$([0.6, 0.7], [0.2, 0.3])$	$(0.6, 0.3)$
A_4	$(0.7, 0.2)$	$(0.5, 0.2)$	$([0.5, 0.7], [0.1, 0.2])$	$(0.5, 0.3)$

Table 2: The relation between the diagnosis and the symptoms—expert 2.

	s_1	s_2	s_3	s_4
A_1	$(0.4, 0.5)$	$(0.6, 0.2)$	$([0.5, 0.6], [0.2, 0.3])$	$(0.3, 0.4)$
A_2	$(0.5, 0.2)$	$(0.7, 0.2)$	$([0.4, 0.7], [0.1, 0.3])$	$(0.7, 0.1)$
A_3	$(0.6, 0.2)$	$(0.5, 0.1)$	$([0.5, 0.7], [0.1, 0.2])$	$(0.6, 0.2)$
A_4	$(0.7, 0.1)$	$(0.4, 0.3)$	$([0.3, 0.6], [0.2, 0.4])$	$(0.4, 0.3)$

Table 3: The relation between the diagnosis and the symptoms—expert 3.

	s_1	s_2	s_3	s_4
A_1	$(0.5, 0.3)$	$(0.6, 0.2)$	$([0.4, 0.6], [0.2, 0.3])$	$(0.5, 0.4)$
A_2	$(0.7, 0.2)$	$(0.4, 0.4)$	$([0.5, 0.7], [0.1, 0.3])$	$(0.6, 0.3)$
A_3	$(0.6, 0.3)$	$(0.7, 0.3)$	$([0.6, 0.8], [0.1, 0.2])$	$(0.7, 0.2)$
A_4	$(0.5, 0.2)$	$(0.5, 0.3)$	$([0.3, 0.6], [0.1, 0.4])$	$(0.6, 0.1)$

According to step 2 in Section 4, applying (10) and (11), respectively, the ideal relation between the diagnosis and the symptoms are shown in Table 4.

Table 4: The ideal relation between the diagnosis and the symptoms.

	s_1	s_2	s_3	s_4
A_1^*	$(0.467, 0.4)$	$(0.567, 0.233)$	$([0.433, 0.6], [0.167, 0.3])$	$(0.4, 0.4)$
A_2^*	$(0.633, 0.233)$	$(0.6, 0.267)$	$([0.4, 0.633], [0.2, 0.367])$	$(0.633, 0.2)$
A_3^*	$(0.667, 0.2)$	$(0.6, 0.267)$	$([0.567, 0.733], [0.133, 0.233])$	$(0.633, 0.233)$
A_4^*	$(0.633, 0.167)$	$(0.467, 0.267)$	$([0.367, 0.633], [0.133, 0.333])$	$(0.5, 0.233)$

Table 5: The weights of the medical expert 1 for A_i with respect to s_j.

	s_1	s_2	s_3	s_4
A_1	0.3427	0.3155	0.3391	0.3333
A_2	0.3614	0.3614	0.3052	0.3223
A_3	0.3761	0.3466	0.3293	0.3263
A_4	0.3458	0.3395	0.3314	0.3313

Table 6: The weights of the medical expert 2 for A_i with respect to s_j.

	s_1	s_2	s_3	s_4
A_1	0.3371	0.3423	0.3311	0.3148
A_2	0.2841	0.3614	0.3474	0.3531
A_3	0.3097	0.2821	0.3196	0.3193
A_4	0.3594	0.3210	0.3276	0.3106

Table 7: The weights of the medical expert 3 for A_i with respect to s_j.

	s_1	s_2	s_3	s_4
A_1	0.3202	0.3422	0.3298	0.3519
A_2	0.3545	0.2772	0.3474	0.3246
A_3	0.3142	0.3713	0.3511	0.3544
A_4	0.2948	0.3395	0.3410	0.3581

When the weight values of the experts are determined, the aggregated evaluating decision results provided by different experts are obtained in Table 8.

Table 8: The aggregated relation between the diagnosis and the symptom.

	s_1	s_2	s_3	s_4
$\mathcal{A}_1$	(0.4683, 0.3933)	(0.5708, 0.2273)	([0.4351, 0.6], [0.1581, 0.3])	(0.4093, 0.4)
$\mathcal{A}_2$	(0.5316, 0.2316)	(0.6364, 0.2424)	([0.4097, 0.6494], [0.1527, 0.3506])	(0.6386, 0.1786)
$\mathcal{A}_3$	(0.6918, 0.175)	(0.6172, 0.2431)	([0.5704, 0.7398], [0.1256, 0.2286])	(0.6388, 0.2283)
$\mathcal{A}_4$	(0.6513, 0.1559)	(0.4699, 0.2614)	([0.3739, 0.6364], [0.1255, 0.3179])	(0.5115, 0.2024)

would be quite good for some real-world applications, such as pattern recognition and medical diagnosis. Through the proposed cosine similarity measure, we can classify the patient $\widetilde{B}$ in one of the diagnosis $A_1, A_2, \ldots A_m$. Finally, a numerical example illustrated the application and efficiency of the developed approach, which is also compared to the existing methods. In future research, we expect to develop further extensions of the C_{HIFS} measure by adding the new characteristic, such as ordered weighted averaging operator, and we will also consider other applications of the proposed C_{HIFS} measure.

Abbreviations

IFS: Intuitionistic fuzzy set
IVIFS: Interval-valued intuitionistic fuzzy set
HIFS: Hybrid intuitionistic fuzzy set
C_{IFS}: Cosine similarity measure for intuitionistic fuzzy set
C_{IVIFS}: Cosine similarity measure for interval-valued intuitionistic fuzzy set
C_{HIFS}: Cosine similarity measure for hybrid intuitionistic fuzzy set.

Data Availability

No data were used to support this study.

Conflicts of Interest

The authors declare that they have no conflicts of interest regarding the publication for the paper.

Acknowledgments

This research was fully supported by the Youth Project of Hunan Education Department (17B092), the Key Project of National Nature Science Foundation of China (Nos. 71431006 and 11501191), a grant from the National Natural Science Foundation of Hunan (2017JJ2096), the National Social Science Fund of China (15BTJ028), and the Major Projects of the National Social Science Foundation of China (17ZDA046). We thank the editor and anonymous reviewers for their helpful comments on an earlier draft of this paper.

References

[1] S. M. Chen, "Measures of similarity between vague sets," *Fuzzy Sets and Systems*, vol. 74, no. 2, pp. 217–223, 1995.
[2] W. L. Hung and M. S. Yang, "Similarity measures of intuitionistic fuzzy sets based on Hausdorff distance," *Pattern Recognition Letters*, vol. 25, no. 14, pp. 1603–1611, 2004.
[3] D. Li and C. Cheng, "New similarity measures of intuitionistic fuzzy sets and application to pattern recognitions," *Pattern Recognition Letters*, vol. 23, no. 1–3, pp. 221–225, 2002.
[4] Z. Liang and P. Shi, "Similarity measures on intuitionistic fuzzy sets," *Pattern Recognition Letters*, vol. 24, no. 15, pp. 2687–2693, 2003.
[5] D. Liu, X. Chen, and D. Peng, "The intuitionistic fuzzy linguistic cosine similarity measure and its application in pattern recognition," *Complexity*, vol. 2018, Article ID 9073597, 11 pages, 2018.
[6] C. P. Pappis and N. I. Karacapilidis, "A comparative assessment of measures of similarity of fuzzy values," *Fuzzy Sets and Systems*, vol. 56, no. 2, pp. 171–174, 1993.
[7] J. Ye, "Cosine similarity measures for intuitionistic fuzzy sets and their applications," *Mathematical and Computer Modelling*, vol. 53, no. 1-2, pp. 91–97, 2011.
[8] L. Zhou, Z. Tao, H. Chen et al., "Intuitionistic fuzzy ordered weighted cosine similarity measure," *Group Decision and Negotiation*, vol. 23, no. 4, pp. 879–900, 2014.
[9] H. Nguyen, "A novel similarity/dissimilarity measure for intuitionistic fuzzy sets and its application in pattern recognition," *Expert Systems with Applications*, vol. 45, pp. 97–107, 2016.
[10] K. T. Atanassov and G. Gargov, "Interval valued intuitionistic fuzzy sets intuitionistic fuzzy sets," *Fuzzy Sets and Systems*, vol. 31, no. 3, pp. 343–349, 1989.
[11] Z. Xu, "On similarity measures of interval-valued intuitionistic fuzzy sets and their application to pattern recognitions," *Journal of Southeast University*, vol. 23, no. 1, 2007.
[12] A. Bhattacharyya, "On a measure of divergence between two multinomial populations," *Sankhya*, vol. 7, no. 4, pp. 401–406, 1946.
[13] J. Ye, "Interval-valued intuitionistic fuzzy cosine similarity measures for multiple attribute decision-making," *International Journal of General Systems*, vol. 42, no. 8, pp. 883–891, 2013.
[14] Y. L. Zhou and M. A. Wahab, "Cosine based and extended transmissibility damage indicators for structural damage detection," *Engineering Structures*, vol. 141, pp. 175–183, 2017.
[15] Y. L. Zhou, N. M. Maia, and M. A. Wahab, "Damage detection using transmissibility compressed by principal component analysis enhanced with distance measure," *Journal of Vibration and Control*, vol. 24, no. 10, 2017.
[16] K.T. Atanassov, "Intuitionistic fuzzy sets," *Fuzzy Sets and Systems*, vol. 20, no. 1, pp. 87–96, 1986.
[17] Z.S. Xu and J. Chen, "On geometric aggregation over interval-valued intuitionistic fuzzy information," in *Proceedings of International Conference on Fuzzy Systems and Knowledge Discovery*, pp. 466–471, IEEE, Haikou, Hainan, China, 2007.

Hindawi
Computational and Mathematical Methods in Medicine
Volume 2018, Article ID 6708520, 10 pages
https://doi.org/10.1155/2018/6708520

Research Article

Potential Genes and Pathways of Neonatal Sepsis Based on Functional Gene Set Enrichment Analyses

YuXiu Meng,[1] Xue Hong Cai,[2] and LiPei Wang[1]

[1]*Department of Neonatology, First People's Hospital of Jining, Jining, Shandong 272000, China*
[2]*Department of Pediatrics, Traditional Chinese Medicine Hospital of Yanzhou, Jining, Shandong 272100, China*

Correspondence should be addressed to LiPei Wang; wanglipeijining@yeah.net

Received 18 January 2018; Revised 4 June 2018; Accepted 27 June 2018; Published 30 July 2018

Academic Editor: Ting Hu

Background. Neonatal sepsis (NS) is considered as the most common cause of neonatal deaths that newborns suffer from. Although numerous studies focus on gene biomarkers of NS, the predictive value of the gene biomarkers is low. NS pathogenesis is still needed to be investigated. *Methods*. After data preprocessing, we used KEGG enrichment method to identify the differentially expressed pathways between NS and normal controls. Then, functional principal component analysis (FPCA) was adopted to calculate gene values in NS. In order to further study the key signaling pathway of the NS, elastic-net regression model, Mann–Whitney *U* test, and coexpression network were used to estimate the weights of signaling pathway and hub genes. *Results*. A total of 115 different pathways between NS and controls were first identified. FPCA made full use of time-series gene expression information and estimated *F* values of genes in the different pathways. The top 1000 genes were considered as the different genes and were further analyzed by elastic-net regression and MWU test. There were 7 key signaling pathways between the NS and controls, according to different sources. Among those genes involved in key pathways, 7 hub genes, PIK3CA, TGFBR2, CDKN1B, KRAS, E2F3, TRAF6, and CHUK, were determined based on the coexpression network. Most of them were cancer-related genes. PIK3CA was considered as the common marker, which is highly expressed in the lymphocyte group. Little was known about the correlation of PIK3CA with NS, which gives us a new enlightenment for NS study. *Conclusion*. This research might provide the perspective information to explore the potential novel genes and pathways as NS therapy targets.

1. Introduction

Neonatal sepsis is the most prevalent cause of death of the neonates with few certainly reported biomarkers for many years. At least 35% neonatal deaths were caused by infections each year. The neonates have usually suffered from early-onset NS, which occurs within the first 72 hours after birth [1, 2]. According to the recent report, the diagnosis of sepsis is humbled by the nonspecific and highly variable human inflammatory and anti-inflammatory processes [3]. The main risk factor that causes the neonatal death is infections, which include respiratory infections, drug-resistant infections, and neonatal tetanus [4]. In order to manage the infections, the research to develop primary and secondary prevention strategies based on different kinds of infections has been a hot field for NS study in recent decades [5, 6]. Future medical research should be based on reducing the application and duration of antibiotics for NS. In view of the side effects caused by the treatment of NS, it is important to make the division for using the right standard of practice for the vulnerable group. The current classification criteria for susceptible populations are crucial to future research and to improve the development of neonatal management strategies. Medzhitov et al. found that the Toll-like receptor-2 and Toll-like receptor-4 are involved in the recognition process of the bacteria in neonates [7]. Septic neonates have a significant upregulation and obvious decline of several genes, which involved in innate immunity [8, 9]. The neonatal innate immune response to sepsis is driven by innate immunity genes (IL1R2, ILRN, and SOCS3) [10]. Current studies also investigated the relationship between the cytokine pattern and onset of NS and proved that the increased expression of proinflammatory cytokines, such as TNF-alpha, IL-6, and IL-10,

was associated with the acute and post-acute phase of NS, respectively [11].

Despite the numerous studies of NS pathogenesis based on genes, the valuable predictors have remained unveiled and contributed to being a major challenge to the research of NS. Gene transcriptomic profiles can be used to identify diagnostic and prognostic gene signatures in complex diseases and to reveal the pathogenesis of NS [9, 10]. Several systems biology approaches were built to dissect the physiological mechanism of sepsis. In particular, methods for discovering the context-specific activations of pathways [12, 13] were merged. However, for the time-course studies, it will be difficult to do clinical trials if the role of the gene of choice is not specific to the biological process of interest. In other words, a temporally differentially expressed gene should show a significant nonconstant expression pattern across time points. To address this, weighting methods were needed to be carried out to assess the functional similarities between a given gene and the sets in different time points [14].

In this report, a method based on the functional principal component analysis (FPCA) was proposed to discover arbitrary nonconstant trends in time-course data analysis [15]. After estimating the impact of the gene, an elastic-net regression model was used to analyze the weights of genes. Besides, a generalized Mann–Whitney U (MWU) test was also applied for gene set-level inferences. Finally, hub genes were determined by the topological feature of coexpression networks [16]. Using the proposed analysis method, susceptible pathways and crucial genes will be revealed. And they will facilitate the future investigation of NS.

2. Methods

2.1. Data Recruitment and Preprocess. Gene expression profiles of human peripheral blood cells at various time points from samples of meningococcal sepsis were deposited at Gene Expression Omnibus database with the data accession no. GSE11755, including NS patients and normal controls. These datasets were processed on Affymetrix Human Genome U133 Plus 2.0 Array platform. Totally, forty-one samples, which were drawn at four time points ($t = 0$, $t = 8$, $t = 24$, and $t = 72$ h after admission to the paediatric intensive care unit), were studied, and key pathways and hub genes were also identified. Next, based on the RNA microarray, gene expressions isolated from whole blood, lymphocytes, and monocytes were also analyzed, respectively. According to the different sources of microarray data, we adopted different groups: The first, we named All Sources, contained all the 41 samples (10 controls and 31 patients). The second, we named Blood Source, contained the microarray data derived from blood (3 controls and 8 patients). The third, we named Lymphocyte Source, contained the microarray data derived from lymphocytes (4 controls and 12 patients). The fourth, we named Monocyte Source, contained the microarray data derived from monocytes (3 controls and 11 patients). In the following analyses, we conducted four parallel analyses based on different groups. The study was approved by the local medical ethics committee.

For data preprocessing, a freely available R platform (http://cran.r-project.org/) was applied. GraphPad Prism 7.0 software was used to create images. And data preprocess of dataset was commenced with reading the data by the standard method carried out by Affy. Expressions of genes were normalized using the robust multiarray average (RMA) method, in order to eliminate the influence of nonspecific hybridization [17]. And then, genes were further filtered by quartile-based algorithm [18]. A total of 15144 genes were reported for each subject.

2.2. Pathway Enrichment Analysis. Pathway analysis was used to find the significant pathways of the NS and control groups according to Kyoto Encyclopedia of Genes and Genomes (KEGG) [19]. Fisher's exact test was adopted to select the significant pathways, and the threshold of significance was defined by FDR and p value. Significant pathways were extracted according to the thresholds of $p < 0.05$ and intersection gene count >1.

2.3. A Gene-Level Summary Statistic by the Functional Principal Component Analysis. In the present research, the FPCA model was used to identify temporally differentially expressed genes [20]. The gene expression profile obtained was assumed to be the scattered members from the true profile of gene expression. And the true profile will be further interfered by noisy signals. After subtracting the average expression value of genes, FPCA was used to center all the gene values. The gene expression profile of preprocessed data was weighted according to their corresponding mean expression and FPCA score across all the gene expression values.

The observed expression using the FPCA model is as follows:

$$\widehat{X}_i(t) = \widehat{\mu}_i + \sum_{l=1}^{L} \widehat{\xi}_{il} \widehat{\Phi}_l(t), \tag{1}$$

where $\widehat{\mu}_i$ is the average expression of the temporal sample, $\widehat{\Phi}_l(t)$ is the lth eigenfunction, and $\widehat{\xi}_{il}$ is the FPC value that quantifies how much $\widehat{X}_i(t)$ can be explained by $\widehat{\Phi}_l(t)$.

When it applied to the time-course gene expression, we used functional F-statistic to summarize the gene pattern information for each gene in the time points:

$$F_i = \frac{\mathrm{RSS}_i^0 - \mathrm{RSS}_i^1}{\mathrm{RSS}_i^1}, \tag{2}$$

where RSS_i^0 is the residual sum of squares of null hypotheses and RSS_i^1 is the residual sum of squares of alternative hypotheses. F_i can be viewed as a "signal-to-noise" ratio and revealed the importance of genes.

2.4. Estimating the Weights of Signaling Pathway Using the Elastic-Net Regression Model. In this study, we also took an approach with computationally efficient and highly flexible methods on the basis of an equivalent influence between the penalty function regression and a standard multivariate

Table 1: Top 6 differentially expressed pathways according to the KEGG analysis.

Pathway_name	p value	FDR	Gene count
Hsa04740: Olfactory transduction	$1.25E-138$	$3.59E-136$	59
Hsa05206: MicroRNAs in cancer	$1.07E-23$	$1.53E-21$	133
Hsa04080: Neuroactive ligand-receptor interaction	$2.84E-10$	$2.03E-08$	179
Hsa04110: Cell cycle	$2.42E-10$	$2.03E-08$	117
Hsa04380: Osteoclast differentiation	$1.45E-09$	$8.27E-08$	122
Hsa00830: Retinol metabolism	$2.47E-09$	$1.0E-07$	23

regression, in order to minimize optimization problem, which is known as the functional elastic-net regression problem [14]. This problem occurs because of the model selection methods in a functional linear regression model that is needless for the concurrent function regression.

The main function of the model is as follows:

$$\widehat{\beta}_i = \min_{\beta_i} \mathrm{OBJ}\left(\beta_i \middle| x_i(t), \widehat{\Phi}_i(t)\right)$$
$$\mathrm{OBJ}\left(\beta_i \middle| x_i(t), \widehat{\Phi}_i(t)\right) = \left\|x_i(t) - \widehat{\Phi}_i(t)^T \beta_i\right\|^2 + \lambda_1 \left\|\beta_i\right\|_1 + \lambda_2 \left\|\beta_i\right\|^2, \tag{3}$$

where λ is the penalty coefficient and β_i is the vector of the set of linear coefficients. When $\widehat{\beta}_i$ is calculated and estimated, then the weights of the pathways can be obtained by

$$\widehat{\mathscr{W}}_{i,k} := \frac{\sum_{l=1}^{L}\left(\widehat{\beta}_{l,i}^{k}\right)^2}{\sum_{k\in K_i}\sum_{l=1}^{L}\left(\widehat{\beta}_{l,i}^{k}\right)^2}. \tag{4}$$

A similar approach can be used to estimate the weights of genes.

2.5. Weighted Mann–Whitney U (MWU) Test with Correlation Using Gene Set Enrichment Analysis (GSEA). MWU test is used to compare two independent samples. Given that two samples were exactly from the same groups, the mean was different. The aim of the MWU test was to analyze whether there was a significant difference between the means of the two groups. Recent reports showed that MWU test plays an important role in gene set enrichment analysis (GSEA) [21, 22]. The pathway enrichment analysis was carried out based on the genome-wide background and was applied to identify the biological functions of the significant clusters. KEGG pathway enrichment was also performed. Categories with more than 5 genes were presented, and p value < 0.01 were considered significant in pathway enrichment analysis [23].

2.6. Identification of Hub Genes Based on the Coexpression Networks. Adjacency matrixes were firstly constructed based on the intergenomic relationships evaluated by Spearman correlation coefficient [24]. Topological features were further studied to find key nodes in the network. Genes whose degree was greater than the average degree values and whose Spearman correlation coefficient was greater than 0.6 were considered as hub genes.

3. Results

3.1. Pathway Enrichment Analysis. Gene expression profile of human NS with the series of GSE11755 was downloaded from Gene Expression Omnibus. After preprocessing the expression profile data of the dataset, we collected data from a total of 41 samples, including six children with meningococcal sepsis. Blood was drawn at four time points and matched with controls. Pathway enrichment analysis of NS and controls was conducted on the basis of the KEGG pathway database. A total of 286 pathways covering 6893 genes were obtained. After Fisher's exact test, 115 differential pathways covering 3532 genes met the thresholds of $p < 0.05$ and intersection gene count >1. Table 1 shows the top 6 differential signaling pathways in ascending order based on p value.

3.2. Integrated Analysis of Gene Signatures Using the FPCA Model. In the present research, the FPCA model was used to identify temporally differentially expressed genes and each gene would get an F value. Based on the 115 differential pathways (covering 3532 genes), we identified top 1000 gene signatures of NS using FPCA model, which were defined as dysregulated genes. FPCA narrowed the gene search range from 3532 to 1000. Greater F value means that the expression level differed greatly with others. Figures 1(a)–1(d) show the curve of gene signatures with F value. Among the dysregulated genes, the top 12 genes from All Sources, Blood Source, Lymphocyte Source, and Monocyte Source were CDC37, NCOA2, P2RY12, RXRB, EDEM2, ACTN4, STX12, PPM1A, PRKACB, DUSP10, VEGFA, and SLC44A2. Since NS is mainly caused by infections, the dysregulated genes in NS should be immune response related. However, there were few genes in the list that were immune related. Activation of the cytokines in a specific infection might not be derived from all the regulated genes that can activate those genes. Therefore, it is important to find the pathways which are activated by infections in NS. FPCA could effectively utilize the time-series information and overcome the traditional control design deficiencies [14]. F values would be used for the MWU test.

3.3. Estimating the Weights of Genes Using the Elastic-Net Regression Model. Genes that exist in multiple pathways were considered as overlapping genes. These genes are thought to play multiple roles in hypothesis testing, where the weight coefficients were overestimated. In the present study, elastic-net regression model was used to decompose

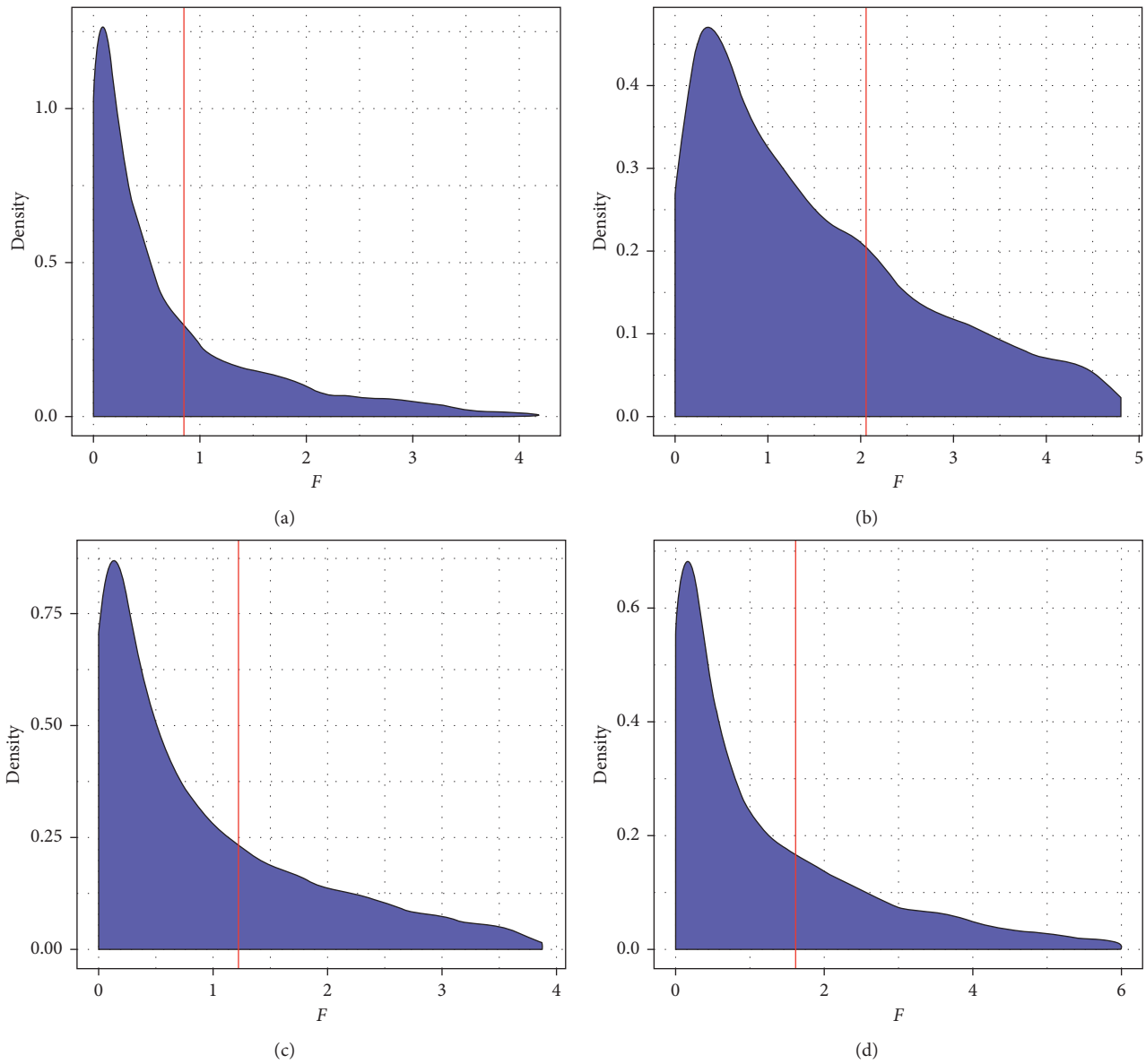


Figure 1: The distribution of F value of pathway genes. Time-series gene signatures data were analyzed by FPCA and each gene obtained an F value (x-coordinate, F value). Y-axis represents gene density. The genes were ranked in the order of F value, and the top 1000 of them were selected. The red line represents the threshold of top 1000 genes. (a) All Sources, (b) Blood Source, (c) Lymphocyte Source, and (d) Monocyte Source.

an overlapping gene between gene sets and eliminate the overlapping effects. After calculating the weight value of each gene and adding the weight values of the pathway genes, the total weight value of the pathways was obtained. Figure 2 shows the sum weight of each pathway. The weight value (w) of each gene would be used for the MWU test.

3.4. Functional Enrichment Analysis Using GSEA and MWU Model Test. Based on the KEGG pathway enrichment, 115 differential pathways were obtained. In order to more accurately find key pathways and molecules, FPCA and elastic-net regression were performed to eliminate overlapping gene effects. Combined with the MWU test, key molecular pathways in the gene transcription data of NS and controls were identified. Based on the t-test, pathways were ranked in the descending order. After the pathway data were tested by the MWU model, a total of 7 pathway terms met the condition p values < 0.05. There was no pathway met the conditions in the monocyte group. The resulting pathways are presented in Table 2.

According to the MWU test, there were 7 pathways were screened based on the p values < 0.05. We selected the top 3 significant pathways: hsa05220: Chronic myeloid leukemia; hsa04380: Osteoclast differentiation; and hsa05222: Small-cell lung cancer for further analysis. Besides, pathways including the proinflammatory cytokine genes were also studied, such as hsa05164: Influenza A (TNF, IL-6, IL-18,

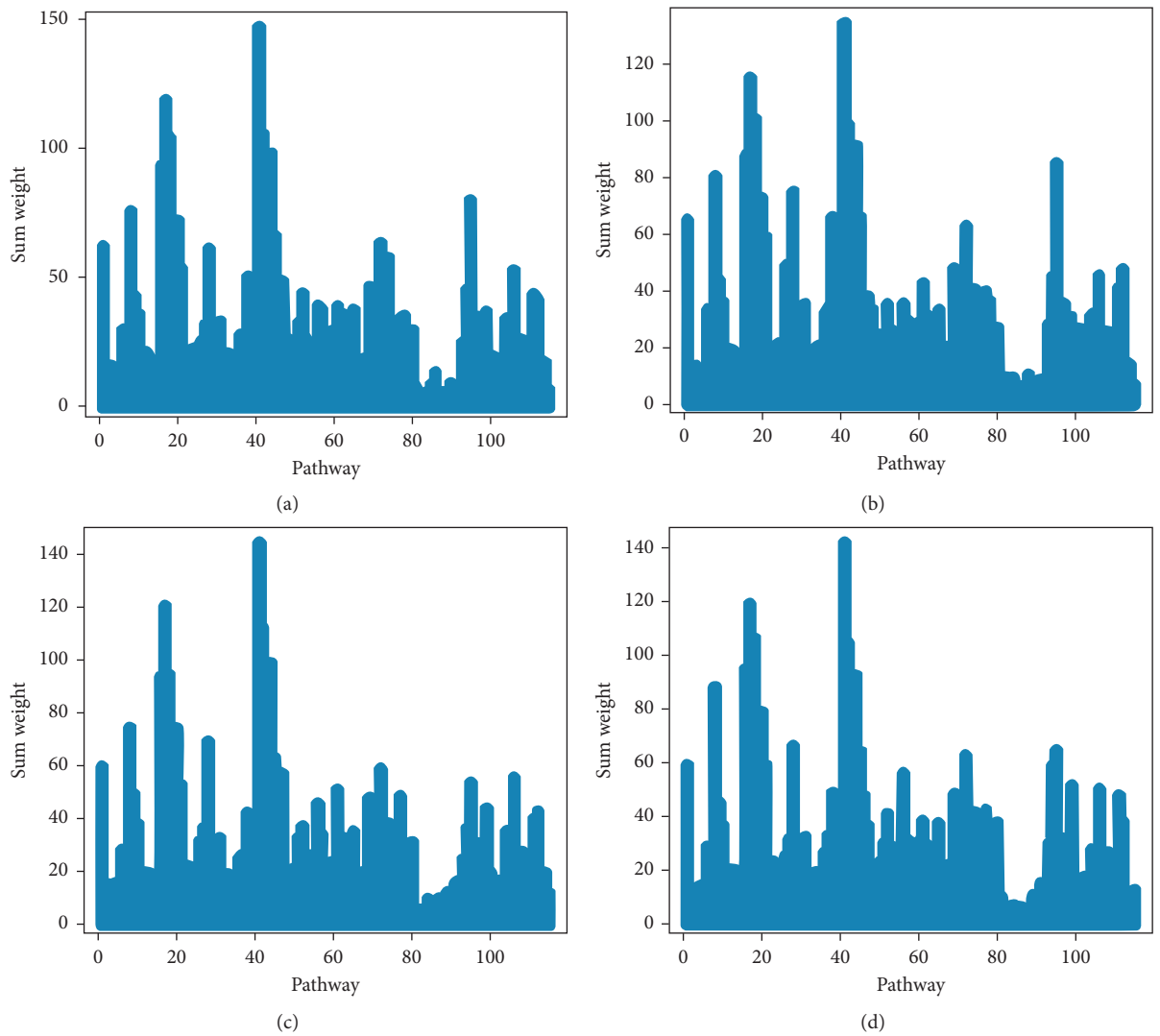


Figure 2: Sum weights of 115 differential pathways. *Y*-axis represents the sum weights of pathways. *X*-axis represents the number of pathways. (a) All Sources, (b) Blood Source, (c) Lymphocyte Source, and (d) Monocyte Source.

Table 2: p values of MWU test of sample groups.

Groups	KEGG pathways	p values
All Sources	hsa05220: Chronic myeloid leukemia	0.0457024
Blood Source	hsa05120: Epithelial cell signaling in *Helicobacter pylori* infection	0.0357933
	hsa04380: Osteoclast differentiation	0.0380088
	hsa04666: Fc gamma R-mediated phagocytosis	0.0415344
Lymphocyte Source	hsa05222: Small-cell lung cancer	0.0150380
	hsa04660: T cell receptor signaling pathway	0.0412070
	hsa05219: Bladder cancer	0.0463346
Monocyte Source	None	

hsa: *Homo sapiens* (human); p values < 0.05 significant difference.

and IFNA1; p value 0.3256); hsa04620: Toll-like receptor signaling pathway (TNF, IL-6, and IFNA1; p value 0.2185); hsa05168: Herpes simplex infection (TNF, IL-6, IFNA1, and IL-15; p value 0.4868). Unfortunately, the MWU test showed that there was no difference between the controls and patients in those proinflammatory cytokines included pathways. For the obtained genes in the top 3 pathways, Figure 3(a) reveals that the expression change of hsa05220: Chronic myeloid leukemia from All Sources was not obvious. The levels of hsa05120: Epithelial cell signaling in *Helicobacter pylori* infection from Blood Source after admission to the paediatric intensive care unit were significantly higher than control. Besides, the levels of hsa05222: Small-cell lung cancer from Lymphocyte Source were up at 72 h after admission to the paediatric intensive care unit.

3.5. Identification and Estimation of the Weights of the Hub Genes in the Pathway. Networks provide effective models to study complex biological systems, such as gene and protein interaction networks. A weighted gene coexpression network was constructed using adjacency matrix based on superman coefficient. We further studied the topological

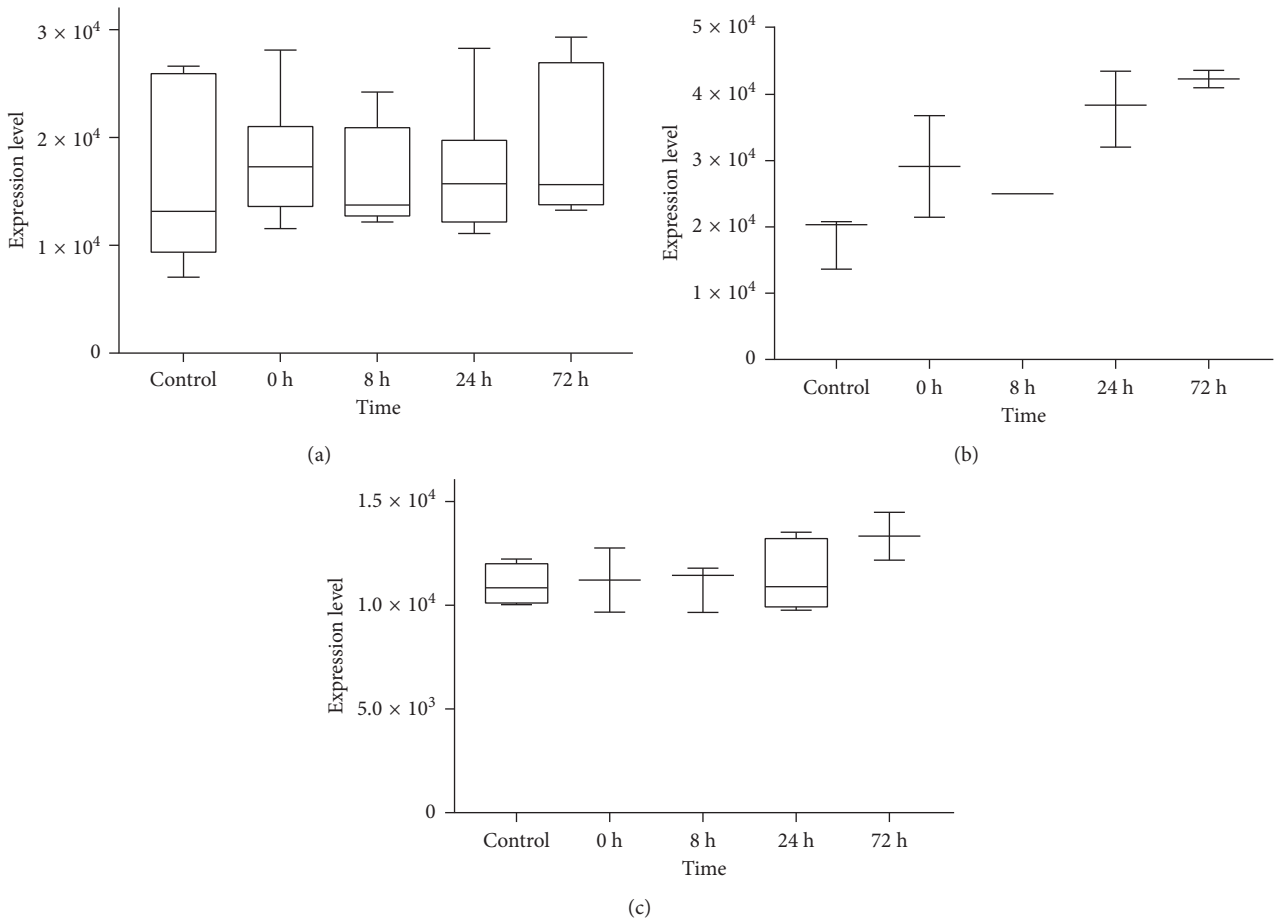


Figure 3: Expression levels of the top 3 significant signaling pathways. (a) hsa05220: Chronic myeloid leukemia from All Sources, (b) hsa05120: Epithelial cell signaling in *Helicobacter pylori* infection from Blood Source, and (c) hsa05222: Small-cell lung cancer from Lymphocyte Source. *Y*-axis represents expression levels of pathways. *X*-axis represents control and several time points after admission to the paediatric intensive care unit. The graphs were made with GraphPad Prism 7.0.

features to find key nodes in the networks. Genes whose degree was greater than the average degree values were considered as hub genes. Based on the three networks of 7 pathways from All Sources, Blood Source, and Lymphocyte Source, we mapped a Venn diagram. Figure 4 shows that the intersection of these three sets contained only one gene, PIK3CA. We defined PIK3CA as the common marker of NS. The intersection of All Sources and Blood Source had two genes, namely, PIK3CA and TGFBR2. A total of 4 genes (PIK3CA, CDKN1B, KRAS, and E2F3) existed in All Sources and Lymphocyte Source sets, simultaneously. There were 3 genes that shared in both Blood Sources and Lymphocyte Source: PIK3CA, TRAF6, and CHUK.

3.6. Expression Levels of Hub Genes and Common Inflammatory Factors. After analyzing the topological features of networks based on 7 pathways, 7 genes were considered as the hub genes which were described above. In order to investigate the relevance of the hub genes and NS, the expression levels of PIK3CA, TGFBR2, CDKN1B, KRAS, E2F3, TRAF6, and CHUK were further analyzed. As we all know that NS is mainly caused by infections, the levels of proinflammatory genes were also observed, such as tumor necrosis factor alpha (TNF-α), interleukin-2 (IL-2), interleukin-6 (IL-6), interleukin-7 (IL-7), interleukin-10 (IL-10), and interferon alpha-1 (IFNA1). Besides, we examined expressions of housekeeping genes GAPDH and beta-catenin (not shown here) aiming at objectively reflecting the changes in hub genes. Figure 5 shows the expression levels of these genes in Box-whisker plot. We can easily found that PIK3CA levels from common and blood groups of patients after admission to the paediatric intensive care unit had no obvious changes compared with controls, while expression of PIK3CA from Lymphocyte Source significantly decreased. According to the reports, NS is mainly caused by infections; however, there was no significant difference between controls and patients in immune response-related gene expression levels (Figures 5(d)–5(i)). The expressions of CDKN1B, KRAS, E2F3, TRAF6, and CHUK were not displayed here.

4. Discussion

In recent years, many new mathematical model methods such as high-dimensional differential equations [25, 26], dynamic Bayesian network [27, 28], and Granger's model [29]

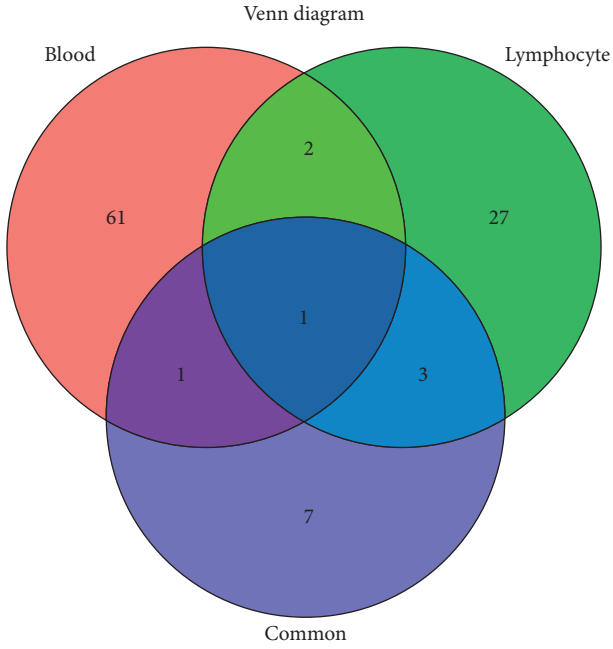


FIGURE 4: Venn diagram of hub genes based on the coexpression networks. Venn diagram showing the number of hub genes obtained from All Sources (Blue), Blood Source (Red), and Lymphocyte Source (Green).

were widely used in molecular biology and bioinformatics. According to reports, inchoate changes in gene expression underlying diseases or infections could be calculated by mathematical models. Low et al. [27, 28] used them to analyze the temporal causality between genes on account of changes expressed at many time points. Time-series gene expression experiments are getting more and more popular. This method plays an important role in studying translation and gene regulation. We provide a flexible way to detect common expression patterns in the individual subjects. Elastic-net regression model combined with the MWU test was used in this study. According to this method, both individual gene and gene set changes, which are induced by infection in a subject-specific way, will be detected.

In the classic MWU test, each variable is independent and there is no relationship between them. However, genes are interrelated, in particular within the related signaling pathway. Therefore, we must make some amendments to the classic MWU test, which can be used to accommodate with gene correlation. In our method, we assume that genes in the relative signaling pathway share a common pairwise correlation q and the irrelevant genes maintain independence. In the current study based on KEGG enrichment of gene signatures, the results showed that, among several KEGG pathways, the top 3 significant pathways were hsa05220: Chronic myeloid leukemia, hsa04380: Osteoclast differentiation, and hsa05222: Small-cell lung cancer, respectively. Besides, pathways including the proinflammatory cytokine genes were also studied, such as hsa05164: Influenza A, hsa04620: Toll-like receptor signaling pathway, and hsa05168: Herpes simplex infection. Unfortunately, the MWU test showed that there was no difference between the control and common groups in those proinflammatory cytokines included pathways. In order to determine the hub genes, based on the topological characteristics of coexpression networks, PIK3CA was defined as the common marker of NS. Then, TGFBR2, CDKN1B, KRAS, E2F3, TRAF6, and CHUK were also selected as our target molecules.

PIK3CA, an oncogene, encodes the p110 catalytic subunit of class I phosphatidylinositol 3-kinases (PI3Ks), namely, PI3Kp110a. Approximately 4/5 of the mutations in PIK3CA occur in the two hot spots, exon 9 and exon 20. Its mutation not only can reduce the apoptosis of cells but also can promote the infiltration of tumors and increase the activity of its downstream kinase PI3Ks [30]. Under physiological conditions, PIK3CA is expressed in brain, lung, mammary gland, gastrointestinal tract, cervix, and other tissues and has many important physiological functions such as regulation of somatic cell proliferation, differentiation, and survival. PIK3CA is often inactive and usually not easily detected. However, PIK3CA was overexpressed after mutation, which could increase the catalytic activity of PI3Ks and promote cell canceration in tissues. PIK3CA mutation has become the molecular biomarker of many tumors [31–34]. PI3K-Akt-mTOR signaling is associated with the balance between cell proliferation and survival and plays a major role not only in tumor growth but also in the potential response of cancer treatment, such as wortmannin and LY294002 [35, 36]. Unfortunately, it seems that there is no direct correlation between PIK3CA and NS in the existing literature.

TGFBR2, transforming growth factor, beta receptor II, is a tumor suppressor gene. The encoded protein is a transmembrane protein that has a protein kinase domain, forms a heterodimeric complex with another receptor protein, and binds TGF-beta. Heterozygous mutations in TGFBR2 play an important role in Marfan syndrome, which is an extracellular matrix disorder with cardinal manifestations in the eye, skeleton, and cardiovascular systems [37]. Several recent reports showed that inducible ablation of TGF-β receptor type 2 signaling was able to limit hepatic stellate cells and fibrosis and attenuates tumor-associated inflammation [38]. TGF-β acts as a key regulator of immune cells, epithelium, in inflammatory bowel disease [39]. Many studies have shown that TGFBR2 signaling was associated with inflammatory-related diseases. But, whether TGFBR2 and NS are related is still a mystery. CDKN1B, a cyclin-dependent kinase inhibitor 1B, can bind to and prevent the activation of cyclin E-CDK2 or cyclin D-CDK4 complexes and thus controls the cell cycle progression at G1. KRAS is a gene that acts as an on/off switch in cell signaling and controls cell proliferation. Most of the target molecules we selected were related to the cell proliferation and tumorigenesis. While very few literature studies report the correlation between them and NS. This study gave us a new enlightenment for neonatal sepsis research. However, there were some limitations to our study. Firstly, PIK3CA and other molecular biomarkers were predictive biomarkers of n, and further experimental verification should be conducted to verify our results. Besides, whether the workflow was suitable for other analysis or another database is a question.

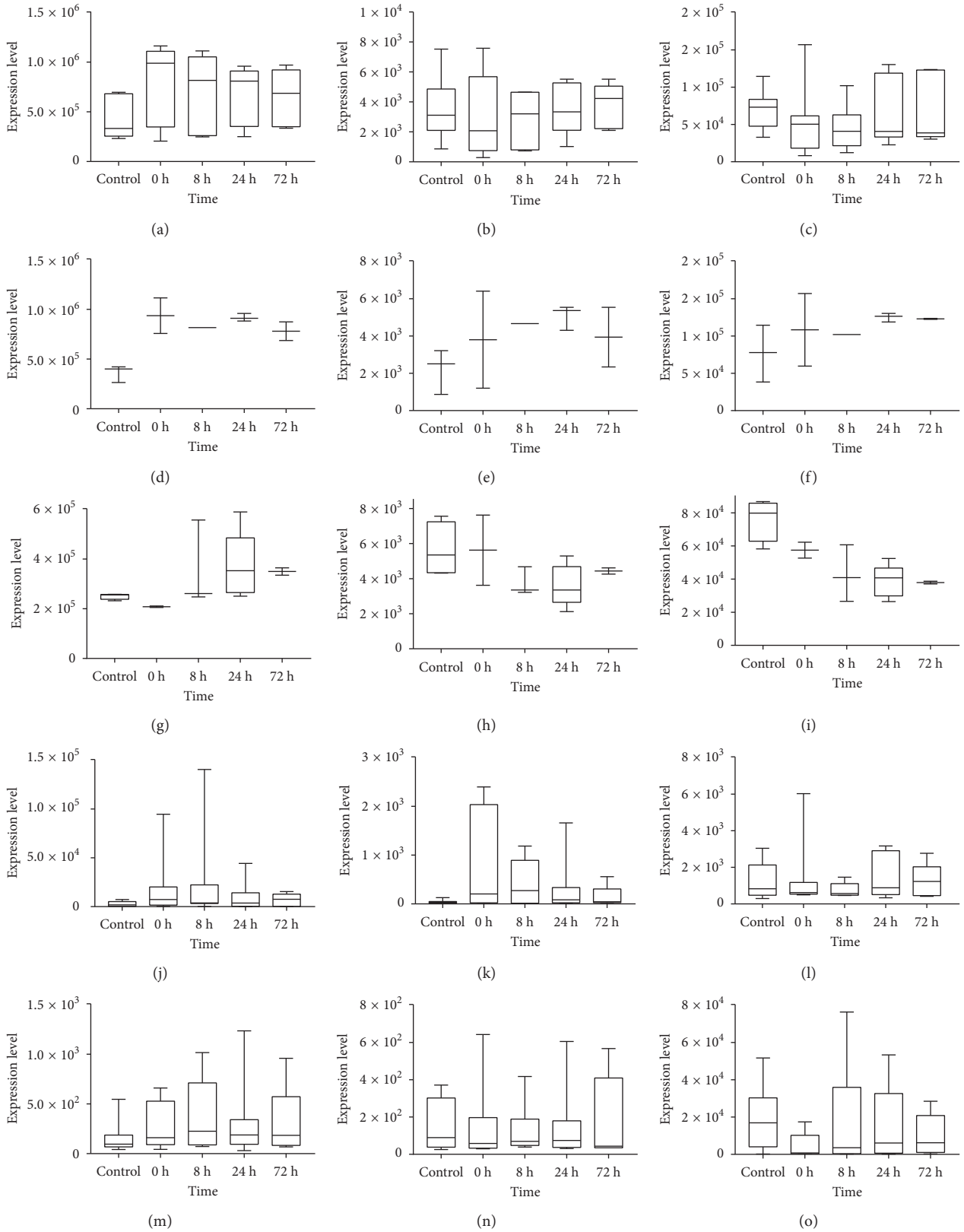


Figure 5: Box-whisker plot of expression levels of genes from GSE11755. (a) GAPDH (internal reference) from All Sources; (b) PIK3CA from All Sources; (c) TGFBR2 from All Sources. (d) GAPDH from Blood Source; (e) PIK3CA from Blood Source; (f) TGFBR2 from Blood Source. (g) GAPDH from Lymphocyte Source; (h) PIK3CA from Lymphocyte Source; (i) TGFBR2 from Lymphocyte Source. Levels of (j) IL-6, (k) IL-10, (l) TNF-α, (m) IL-18, (n) IL-7, and (o) IFNA1 from All Sources. *Y*-axis represents the expression levels of genes. *X*-axis represents control and several time points after admission to the paediatric intensive care unit. The box represents the express range and the central line was the median of the data. All graphs were made with GraphPad Prism 7.0.

5. Conclusion

In conclusion, a comprehensive process of data in datasets of NS was conducted in our research. Then, the function and signaling pathways of NS were presented systematically by the cutting edge models. Finally, based on the potential pathways and their topological characteristics of coexpression networks, several critical genes for NS were identified. PIK3CA was defined as the common marker of NS. However, the current study was based on the previous reports and more clinical evidence results were needed.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Disclosure

YuXiu Meng and Xue Hong Cai are co-first authors.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] S. Skibsted, A. E. Jones, M. A. Puskarich et al., "Biomarkers of endothelial cell activation in early sepsis," *Shock*, vol. 39, no. 5, pp. 427–432, 2013.

[2] J. Song, D. Hu, C. He et al., "Novel biomarkers for early prediction of sepsis-induced disseminated intravascular coagulation in a mouse cecal ligation and puncture model," *Journal of Inflammation*, vol. 10, no. 1, p. 7, 2013.

[3] D. Jrovsky, I. C. Marchetti, M. A. da Silva Mori et al., "Early-onset neonatal pneumococcal sepsis: a fatal case report and brief literature review," *Pediatric Infectious Disease Journal*, vol. 37, no. 4, pp. e111–e112, 2017.

[4] A. Abbas and I. Ahmad, "First report of neonatal early-onset sepsis caused by multi-drug-resistant Raoultella ornithinolytica," *Infection*, vol. 46, no. 2, pp. 275–277, 2017.

[5] W. van Herk, S. el Helou, J. Janota et al., "Variation in current management of term and late-preterm neonates at risk for early-onset sepsis: an international survey and review of guidelines," *Pediatric Infectious Disease Journal*, vol. 35, no. 5, pp. 494–500, 2016.

[6] S. Vergnano, M. Sharland, P. Kazembe, C. Mwansambo, and P. T. Heath, "Neonatal sepsis: an international perspective," *Archives of Disease in Childhood–Fetal and Neonatal Edition*, vol. 90, no. 3, pp. F220–F224, 2005.

[7] R. Medzhitov, "Toll-like receptors and innate immunity," *Nature Reviews Immunology*, vol. 1, no. 2, pp. 135–145, 2001.

[8] J. L. Wynn and S. Guthrie, "Postnatal age is a critical determinant of the neonatal host response to sepsis," *Molecular Medicine*, vol. 21, pp. 496–504, 2015.

[9] M. Cernada, E. Serna, C. Bauerl, M. C. Collado, G. Perez-Martinez, and M. Vento, "Genome-wide expression profiles in very low birth weight infants with neonatal sepsis," *Pediatrics*, vol. 133, no. 5, pp. e1203–e1211, 2014.

[10] C. L. Smith, P. Dickinson, T. Forster et al., "Identification of a human neonatal immune-metabolic network associated with bacterial infection," *Nature Communications*, vol. 5, p. 4649, 2014.

[11] K. S. Khaertynov, S. V. Boichuk, S. F. Khaiboullina et al., "Comparative assessment of cytokine pattern in early and late onset of neonatal sepsis," *Journal of Immunology Research*, vol. 2017, Article ID 8601063, 8 pages, 2017.

[12] B. M. Hartmann, J. Thakar, R. A. Albrecht et al., "Human dendritic cell response signatures distinguish 1918, pandemic, and seasonal H1N1 influenza viruses," *Journal of Virology*, vol. 89, no. 20, pp. 10190–10205, 2015.

[13] D. Katanic, A. Khan, and J. Thakar, "PathCellNet: cell-type specific pathogen-response network explorer," *Journal of Immunological Methods*, vol. 439, pp. 15–22, 2016.

[14] Y. Zhang, D. J. Topham, J. Thakar, and X. Qiu, "FUNNEL-GSEA: FUNctioNal ELastic-net regression in time-course gene set enrichment analysis," *Bioinformatics*, vol. 33, no. 13, pp. 1944–1952, 2017.

[15] I. Sohn, K. Owzar, S. L. George, S. Kim, and S. H. Jung, "A permutation-based multiple testing method for time-course microarray experiments," *BMC Bioinformatics*, vol. 10, p. 336, 2009.

[16] S. Das, P. K. Meher, A. Rai, L. Mohan Bhar, and B. Nath Mandal, "Statistical approaches for gene selection, hub gene identification and module interaction in gene co-expression network analysis: an application to aluminum stress in soybean (Glycine max L.)," *PLoS One*, vol. 12, no. 1, Article ID e0169605, 2017.

[17] Y. Kim, B. Q. Doan, P. Duggal, and J. E. Bailey-Wilson, "Normalization of microarray expression data using within-pedigree pool and its effect on linkage analysis," *BMC Proceedings*, vol. 1, no. S1, p. S152, 2007.

[18] R. C. Gehrau, V. R. Mas, C. I. Dumur et al., "Donor hepatic steatosis induce exacerbated ischemia-reperfusion injury through activation of innate immune response molecular pathways," *Transplantation*, vol. 99, no. 12, pp. 2523–2533, 2015.

[19] J. Zhu and X. Yao, "Use of DNA methylation for cancer detection: promises and challenges," *International Journal of Biochemistry and Cell Biology*, vol. 41, no. 1, pp. 147–154, 2009.

[20] S. Hug, A. Raue, J. Hasenauer et al., "High-dimensional Bayesian parameter estimation: case study for a model of JAK2/STAT5 signaling," *Mathematical Biosciences*, vol. 246, no. 2, pp. 293–304, 2013.

[21] R. Lin, S. Dai, R. D. Irwin, A. N. Heinloth, G. A. Boorman, and L. Li, "Gene set enrichment analysis for non-monotone association and multiple experimental categories," *BMC Bioinformatics*, vol. 9, p. 481, 2008.

[22] A. P. Oron, Z. Jiang, and R. Gentleman, "Gene set enrichment analysis using linear models and diagnostics," *Bioinformatics*, vol. 24, no. 22, pp. 2586–2591, 2008.

[23] X. Chen, L. Wang, J. D. Smith, and B. Zhang, "Supervised principal component analysis for gene set enrichment of microarray data with continuous or survival outcomes," *Bioinformatics*, vol. 24, no. 21, pp. 2474–2481, 2008.

[24] J. Petereit, S. Smith, F. C. Harris, and K. A. Schlauch, "Petal: Co-expression network modelling in R," *BMC Systems Biology*, vol. 10, no. 2, p. 51, 2016.

[25] T. T. Cai and A. Zhang, "Inference for high-dimensional differential correlation matrices," *Journal of Multivariate Analysis*, vol. 143, pp. 107–126, 2016.

[26] N. C. Chung and J. D. Storey, "Statistical significance of variables driving systematic variation in high-dimensional data," *Bioinformatics*, vol. 31, no. 4, pp. 545–554, 2015.

[27] S. T. Low, M. S. Mohamad, S. Omatu, L. En Chai, S. Deris, and M. Yoshioka, "Inferring gene regulatory networks from perturbed gene expression data using a dynamic Bayesian

network with a Markov Chain Monte Carlo algorithm," in *Proceedings of the IEEE International Conference on Granular Computing*, Noboribetsu, Hokkaido, Japan, October 2014.

[28] W. C. Young, A. E. Raftery, and K. Y. Yeung, "Fast Bayesian inference for gene regulatory networks using ScanBMA," *BMC Systems Biology*, vol. 8, no. 1, p. 47, 2014.

[29] S. Basu, A. Shojaie, and G. Michailidis, "Network granger causality with inherent grouping structure," *Journal of Machine Learning Research*, vol. 16, pp. 417–453, 2012.

[30] Y. Samuels, Z. Wang, A. Bardelli et al., "High frequency of mutations of the PIK3CA gene in human cancers," *Science*, vol. 304, no. 5670, p. 554, 2004.

[31] N. A. Lockney, X. Pei, L. E. Blumberg et al., "PIK3CA activating mutations are associated with decreased local control in lung cancer brain metastases treated with radiation," *International Journal of Radiation Oncology Biology Physics*, vol. 96, no. 2, pp. S178–S179, 2016.

[32] C. D. Young, A. D. Pfefferle, P. Owens et al., "Conditional loss of ErbB3 delays mammary gland hyperplasia induced by mutant PIK3CA without affecting mammary tumor latency, gene expression, or signaling," *Cancer Research*, vol. 73, no. 13, pp. 4075–4085, 2013.

[33] Z. Xu, X. Huo, C. Tang et al., "Frequent mutations in MLH1, MET, KIT, PDGFRA, and PIK3CA genes in human gastrointestinal stromal tumors," *Pensar-Revista de Ciências Jurídicas*, vol. 11, no. 1, 2013.

[34] L. Xiang, W. Jiang, J. Li et al., "PIK3CA mutation analysis in Chinese patients with surgically resected cervical cancer," *Scientific Reports*, vol. 5, article 14035, 2015.

[35] I. A. Mayer and C. L. Arteaga, "The PI3K/AKT pathway as a target for cancer treatment," *Annual Review of Medicine*, vol. 67, no. 1, pp. 11–28, 2015.

[36] F. Atif, S. Yousuf, and D. G. Stein, "Anti-tumor effects of progesterone in human glioblastoma multiforme: role of PI3K/Akt/mTOR signaling," *Journal of Steroid Biochemistry and Molecular Biology*, vol. 146, pp. 62–73, 2015.

[37] R. D. Cario, E. Sticchi, S. Nistri, G. Pepe, and B. Giusti, "Role of TGFBR1 and TGFBR2 genetic variants in determining or modulating Marfan syndrome," *Nutrition, Metabolism and Cardiovascular Diseases*, vol. 27, no. 1, p. e17, 2017.

[38] D. R. Principe, B. DeCant, J. Staudacher et al., "Loss of TGFβ signaling promotes colon cancer progression and tumor-associated inflammation," *Oncotarget*, vol. 8, no. 3, p. 3826, 2017.

[39] S. Ihara, Y. Hirata, and K. Koike, "TGF-β in inflammatory bowel disease: a key regulator of immune cells, epithelium, and the intestinal microbiota," *Journal of Gastroenterology*, vol. 52, no. 7, pp. 1–11, 2017.

Hindawi
Computational and Mathematical Methods in Medicine
Volume 2018, Article ID 2497471, 22 pages
https://doi.org/10.1155/2018/2497471

Research Article

Biomedical Text Categorization Based on Ensemble Pruning and Optimized Topic Modelling

Aytuğ Onan

Celal Bayar University, Department of Software Engineering, 45400 Turgutlu, Manisa, Turkey

Correspondence should be addressed to Aytuğ Onan; aytugonan@gmail.com

Received 20 January 2018; Revised 29 May 2018; Accepted 31 May 2018; Published 22 July 2018

Academic Editor: Federico Divina

Text mining is an important research direction, which involves several fields, such as information retrieval, information extraction, and text categorization. In this paper, we propose an efficient multiple classifier approach to text categorization based on swarm-optimized topic modelling. The Latent Dirichlet allocation (LDA) can overcome the high dimensionality problem of vector space model, but identifying appropriate parameter values is critical to performance of LDA. Swarm-optimized approach estimates the parameters of LDA, including the number of topics and all the other parameters involved in LDA. The hybrid ensemble pruning approach based on combined diversity measures and clustering aims to obtain a multiple classifier system with high predictive performance and better diversity. In this scheme, four different diversity measures (namely, disagreement measure, Q-statistics, the correlation coefficient, and the double fault measure) among classifiers of the ensemble are combined. Based on the combined diversity matrix, a swarm intelligence based clustering algorithm is employed to partition the classifiers into a number of disjoint groups and one classifier (with the highest predictive performance) from each cluster is selected to build the final multiple classifier system. The experimental results based on five biomedical text benchmarks have been conducted. In the swarm-optimized LDA, different metaheuristic algorithms (such as genetic algorithms, particle swarm optimization, firefly algorithm, cuckoo search algorithm, and bat algorithm) are considered. In the ensemble pruning, five metaheuristic clustering algorithms are evaluated. The experimental results on biomedical text benchmarks indicate that swarm-optimized LDA yields better predictive performance compared to the conventional LDA. In addition, the proposed multiple classifier system outperforms the conventional classification algorithms, ensemble learning, and ensemble pruning methods.

1. Introduction

The immense quantity of biomedical text documents can serve as an essential source of information for biomedical research. Biomedical text documents are characterized by an immense quantity of unstructured and sparse information in a wide range of forms, such as scientific articles, biomedical datasets, and case reports. Text mining aims to identify valuable information from unstructured text documents with the use of tools and techniques from several disciplines, such as machine learning, information retrieval, and computational linguistics. The use of text mining is one of the most promising tools in the biomedical domain that has attracted a lot of research interest. Text mining in biomedical domain can be successfully applied in a wide range of applications, including identification of disease-specific knowledge [1], diagnosis, treatment, and prevention of cancer [2], identification of obesity status of patients [3], identification of risk factors for heart disease [4], annotation of gene expression [5], and identification of drug targets and candidates [6].

Biomedical text mining follows the same stages (namely, format conversation, tokenization, stop word removal, normalization, stemming, dictionary construction, and vector space construction) utilized in the text processing from other domains [7]. To build accurate classification schemes on text documents, one pivotal issue is to identify an appropriate representation model for the documents [8]. The vector space model (also known as term vector model) is one of the most commonly employed representation schemes to process text documents, owing to its simple structure [9]. In this model, each text document is represented as vectors of identifiers (index terms). The vector space model suffers from

high dimensional feature space, irrelevancy, and sparsity of features. Since each document is represented as a bag of words with the corresponding frequencies, words are regarded as statistically independent. Hence, word order is not taken into consideration [10].

Considering the limitations of the vector space model and the high dimensional unstructured nature of biomedical text documents, there are a number of representation schemes (such as the latent semantic analysis, the probabilistic latent semantic analysis, and the latent Dirichlet allocation) employed to process biomedical text documents [7]. The latent semantic analysis (LSA) is a scheme to extract and represent the contextual meaning of words with the use of statistical computations utilized on a large amount of text [11]. LSA can represent the semantic relations within the text. It can find the latent classes, while reducing the dimensionality of vector space model [12]. However, LSA has no strong statistical foundation and can suffer from high mathematical complexity [13]. The probabilistic latent semantic analysis (PLSA) is a statistical method for analysis of data which is based on a latent class model. PLSA has a strong statistical foundation. It can find latent topics and it can yield better performance compared to LSA [13].

The latent Dirichlet allocation (LDA) is an efficient generative probabilistic topic model, where each document is represented as a random mixture of latent topics. LDA can find latent topics, reduce the high dimensionality of vector space model, and can outperform other linguistic representation schemes, such as latent semantic analysis and probabilistic latent semantic analysis [14]. LDA involves several parameter values, such as the number of topics, the number of iterations for Gibbs sampling, α parameter to control the topic distribution per document, and β parameter to model distributions of terms per topic (Panichella et al., 2003). For unstructured text documents, information about the document-wise content and number of relevant topics is not known in advance (Zhao et al., 2005). Hence, the identification of an appropriate value for the number of topics is a challenging problem for unstructured text documents. An insufficient or excessive number of topics can degrade the predictive performance of machine learning algorithms built on LDA-based topic modelling. In addition to the number of topics, LDA requires several other parameters. Therefore, finding an optimal configuration for LDA-based topic modelling involves extensive empirical analysis with different configurations.

In order to build robust classification schemes, multiple classifier systems (also known as ensemble classifiers) have been widely employed in the field of pattern recognition, owing to its remarkable improvement in generalization ability and predictive performance [15]. There are three main stages of the ensemble learning process, namely, ensemble generation, ensemble pruning, and ensemble combination [16, 17]. The ensemble generation stage is the phase, in which base learning algorithms to be utilized in the multiple classifier system are generated. The base learning algorithms can be generated either homogeneously or heterogeneously. The ensemble combination stage seeks to integrate the individual predictions of base learning algorithms. The ensemble pruning stage aims to identify an optimal subset of base learning algorithms from the ensemble to enhance the predictive performance and computational efficiency. It has been empirically validated that ensemble pruning can yield more robust classification schemes [18].

Considering these issues, we propose a multiple classifier approach to biomedical text categorization based on swarm-optimized topic modelling and ensemble pruning. In the presented scheme, swarm-optimized approach is employed to estimate the parameters of LDA, including the number of topics and all the other parameters involved in LDA. Motivated by the success of hybrid ensemble pruning schemes [19–21], the proposed approach combines diversity measures and clustering. In this scheme, four different diversity measures (namely, disagreement measure, Q-statistics, the correlation coefficient, and the double fault measure) are computed to capture the diversities within the ensemble. Based on these diversity measures, a combined diversity matrix is obtained. Based on this matrix, a swarm intelligence based clustering algorithm partitions the classification algorithms into a number of disjoint groups and one algorithm (with the highest predictive performance) from each cluster is selected to build the multiple classifier system. In the empirical analysis, five biomedical text benchmarks have been utilized. In the swarm-optimized LDA, different metaheuristic algorithms (such as genetic algorithms, particle swarm optimization, firefly algorithm, cuckoo search algorithm, and bat algorithm) are considered. In addition, five different metaheuristic clustering algorithms are considered in the ensemble pruning stage. The empirical analysis on biomedical text benchmarks indicates that swarm-optimized LDA yields better predictive performance compared to the conventional LDA. In addition, the proposed hybrid ensemble pruning scheme outperforms the conventional classification algorithms and ensemble learning methods.

The main contributions of our proposed categorization scheme can be summarized as follows:

(i) We introduced a metaheuristic approach to optimize the set of parameters utilized in LDA-based topic modelling. In this regard, the number of topics (k), the number of Gibbs iterations (n), α parameter to control the topic distribution per document, and β parameter to model distributions of terms per topic are considered. We conducted several experiments on swarm-optimized LDA with different metaheuristic algorithms (namely, genetic algorithms, particle swarm optimization, firefly algorithm, cuckoo search algorithm, and bat algorithm). To the best of our knowledge, this is the first comprehensive empirical analysis devoted to metaheuristic algorithms on LDA-based topic modelling.

(ii) We introduced an ensemble pruning approach based on combined diversity measures and metaheuristic clustering. To the best of our knowledge, this is the first study in ensemble pruning, which utilizes metaheuristic clustering algorithms to obtain diversified base learning algorithms.

(iii) The presented classification scheme, which integrates swarm-optimized LDA-based modelling with the hybrid ensemble pruning scheme, is employed on biomedical text categorization. To the best of our knowledge, this is the first comprehensive study on LDA-based topic modelling and ensemble pruning on biomedical text categorization.

The rest of this paper is structured as follows. In Section 2, related work on topic modelling and multiple classifier systems have been presented. Section 3 presents the theoretical foundations, Section 4 presents the proposed text categorization framework, Section 5 presents the experimental results, and Section 6 presents the concluding remarks.

2. Related Work

This section presents the related work on topic modelling and multiple classifier systems in biomedical text categorization.

2.1. Related Work on Topic Modelling. Topic modelling models have been successfully employed to summarize large-scale collections of text documents. Probabilistic topic modelling methods can be utilized to identify the core topics of text collections. In addition, topic modelling schemes can be utilized in a variety of tasks in computational linguistics, such as analysis of source code documents [23], summarizing opinions of product reviews [24], identification of topic evolution [25], aspect detection in review documents [26], analysis of Twitter messages [27], and sentiment analysis [28, 29].

Probabilistic topic modelling has attracted the attention of researchers on biomedical domain. Biomedical text collections suffer from high dimensionality and topic modelling methods are effective tools to handle with large-scale collections of documents. Hence, topic modelling can yield promising results on biological and biomedical text mining [30]. For instance, Wang et al. [31] presented a probabilistic topic modelling scheme to identify protein-protein interactions from the biological literature. In this scheme, the correlation between different methods and related words is modelled in a probabilistic way to extract the detection methods. In another study, Arnold et al. [32] utilized the latent Dirichlet allocation method to identify relevant clinical topics and to structure clinical text reports. Song and Kim [33] employed the latent Dirichlet allocation method to conduct bibliometric analysis on bioinformatics from full-text text collections of PubMed Central articles. In another study, Sarioglu et al. [34] utilized topic modelling to represent clinical reports in a compact way, so that these collections can be efficiently processed. In another study, Bisgin et al. [35] applied topic modelling to drug labelling, which is a human-intensive task with many ambiguous semantic descriptions. In this way, manual annotation challenges can be eliminated. Likewise, Wang et al. [36] introduced a topic modelling based scheme to identify literature-driven annotations for gene sets. In this scheme, the number of topics to be utilized in topic modelling is empirically inferred through the analysis with various parameter values (5, 10, 15, 20, etc.) for the number of topics. In another study, Bisgin et al. [37] employed the latent Dirichlet allocation based topic modelling to identify interdependencies between cellular endpoints. The experimental analysis indicated that LDA can substantially enhance the understanding of systems biology. Probabilistic topic modelling has also been employed to identify drug repositioning strategies [38]. Wang et al. [39] utilized topic modelling to analyze 17,723 abstracts from PubMed publications related to adolescent substance use and depression. In this study, topic modelling was employed to identify the literature and to capture other relevant topics. In another study, Wang et al. [40] presented a topic modelling based scheme to mine biomedical text collections. In this scheme, topic modelling was employed as a fine-grained preprocessing model. Recently, Sullivan et al. [41] utilized topic modelling to identify unsafe nutritional supplements from review documents. In another study, Chen et al. [42] employed probabilistic topic modelling to represent hospital admission processes in a compact way.

2.2. Related Work on Multiple Classifier Systems. Multiple classifier systems have been successfully employed in a wide range of applications in pattern recognition, including biomedical domain. Empirical analysis on multiple classifier systems indicates that ensemble pruning can enhance the predictive performance of multiple classifier systems [18]. Ensemble pruning approaches can be mainly divided into five groups, as exponential search, randomized search, sequential search, ranking-based, and clustering based methods [16]. Exponential approaches to ensemble pruning seek to examine all possible subsets of base learning algorithms within the multiple classifier system. For instance, Aksela [43] examined the predictive performance of several evaluation metrics (namely, correlation between errors, Q-statistics, and mutual information) in ensemble pruning. Randomized approaches to ensemble pruning aim to explore the search space of candidate classifiers with the use of metaheuristic algorithms. A wide range of metaheuristics, such as genetic algorithms, tabu search, and population based incremental learning, have been successfully utilized for ensemble pruning [44, 45]. For instance, Sheen and Sirisha [46] introduced an ensemble pruning scheme for malware detection based on harmony search. Likewise, Mendialdua et al. [47] utilized the estimation of distribution algorithm for ensemble pruning. In sequential search based methods, the search space of candidate classifiers has been explored in forward, backward, or forward-backward direction. For instance, Margineantu and Dietterich [48] introduced a sequential approach for ensemble pruning based on reduced error pruning with back-fitting. Similarly, Caruana et al. [49] presented a forward stepwise selection based approach for ensemble pruning. Recently, Dai et al. [50] introduced a reverse reduced error-based ensemble pruning algorithm based on subtraction operation. Ranking-based approaches to ensemble pruning aim to identify an optimal subset of classifiers based on a ranking obtained by a particular evaluation measure. For instance, Kotsiantis and Pintelas [51] presented a t-test based ranking scheme for ensemble pruning. More recently, Galar et al. [52] presented an ordering-based metric for

ensemble pruning. Clustering based approaches to ensemble pruning partition the base learning algorithms of ensemble into clusters. For instance, Zhang and Cao [53] presented a spectral clustering based algorithm for ensemble pruning. In this scheme, the base learning algorithms were grouped into two clusters based on predictive performance and diversity. Then, one cluster of ensemble was pruned and one cluster of ensemble was retained as the pruned subset of classifiers.

2.3. Motivation and Contribution of the Study. As outlined in advance, probabilistic topic modelling methods are essential tools to identify hidden topics in large-scale collections of text documents. In order to enhance the performance of LDA, there are a number of extensions on the basic model. For instance, Griffiths and Tenenbaum [54] introduced a hierarchical latent Dirichlet allocation model. In this model, topic distributions are identified from hierarchies of topics, where each hierarchy is modelled by a nested Chinese restaurant process. Each node of tree corresponds to a particular topic, where each topic is associated with a distribution. In another study, Teh et al. [55] presented a hierarchical latent Dirichlet allocation scheme, in which parameter value for the number of topics is inferred through the use of posterior inference. Grant and Cordy [56] introduced a heuristic approach to estimate the number of topics in source code analysis. In another study, Panichella et al. [57] presented a genetic algorithm based scheme to identify optimal configurations for latent Dirichlet allocation. In this scheme, parameter set for topic modelling was estimated with the use of genetic algorithm. The presented scheme was employed on three different tasks of software engineering, namely, traceability link recovery, feature location, and software artifact labelling. Likewise, Zhao et al. [58] introduced a heuristic approach to estimate the appropriate number of topics for latent Dirichlet allocation. In this scheme, the appropriate number of topics is identified through the use of ratio for perplexity change. Recently, Karami et al. [59] presented a fuzzy approach to topic modelling. In this scheme, fuzzy clustering was employed to identify optimal number of topics.

In addition to the aforementioned five ensemble pruning approaches, hybrid methods have attracted research attention in the pattern recognition. Hybrid approaches to ensemble pruning seek to integrate several ensemble pruning paradigms. For instance, Lin et al. (2014) introduced a hybrid ensemble pruning algorithm which integrates k-means clustering and dynamic selection. Similarly, Mousavi and Eftekhari [60] presented a hybrid ensemble pruning scheme which integrates static and dynamic ensemble selection with NSGA-II multiobjective genetic algorithm. In another study, Cavalcanti et al. [21] presented a hybrid ensemble pruning algorithm based on genetic algorithm and graph coloring. In this scheme, several different diversity measures (such as Q-statistics, correlation coefficient, Kappa statistics, and double fault measure) are combined via a genetic algorithm. Similarly, Onan et al. [19, 20] introduced a hybrid ensemble pruning algorithm based on consensus clustering and multiobjective evolutionary algorithm. In this scheme, classifiers are assigned into clusters based on their predictive performance and the set of candidate classifiers are explored through the use of evolutionary algorithm.

Recent studies on topic modelling indicate that the identification of an appropriate parameter value for the number of topics is an essential task to build robust classification schemes. In addition, hybrid ensemble pruning schemes can outperform conventional classifiers, ensemble learning methods, and ensemble pruning methods. Through their potential use on text classification, the number of works that utilize metaheuristic algorithms to optimize parameters of LDA and the number of works that utilize ensemble pruning schemes are very limited. To fill this gap, this paper presents a classification scheme based on swarm-optimized topic modelling and hybrid ensemble pruning for text categorization.

3. Theoretical Foundations

This section summarizes the theoretical foundations of the study. Namely, the latent Dirichlet allocation method, swarm-based optimization algorithms, ensemble learning methods, ensemble pruning methods, cluster validity indices, and pairwise diversity measures are presented.

3.1. The Latent Dirichlet Allocation. The latent Dirichlet allocation model (LDA) is a widely employed generative probabilistic model to identify the latent topics in text documents [22]. In LDA, each document is represented as a random mixture of latent topics and each topic is represented as a mixture of words. The mixture distributions are Dirichlet-distributed random variables to be inferred. In this scheme, each document exhibits the topics in different proportions, each word in each document is drawn among the topics, and topics are chosen based on per-document distribution over topics [61]. LDA attempts to determine the underlying latent topic structure based on the observed data. In LDA, the words of each document correspond to the observed data. For each document in the corpus, words are obtained by following a two-staged procedure. Initially, a distribution over topics is randomly chosen for each word of the document [22]. In LDA, a word is a discrete data from a vocabulary indexed by $\{1, \ldots, V\}$, a sequence of N words $w=(w_1, w_2, \ldots, w_n)$, and a corpus is a collection of M documents denoted by $D=\{\mathrm{w}_1, \mathrm{w}_2, \ldots, \mathrm{w}_M\}$. The generative process of LDA is summarized in Box 1.

LDA process can be modelled by a three-level Bayesian graphical model, as given in Figure 1. In this graphical model, nodes are used to represent random variables and edges are used to denote the possible dependencies between the variables. In this notation, α refers to Dirichlet parameter, Θ refers to document-level topic variables, z refers to per-word topic assignment, w refers to the observed word, and β indicates the topics [61].

Based on this notation, the generative process of LDA corresponds to a joint distribution of the hidden and observed variables. The probability density function of a k-dimensional Dirichlet random variable is computed as given by (1), the joint distribution of a topic mixture is computed as given by

For each document w in a corpus D:
(1) Choose $N \sim$ Poisson (ξ).
(2) Choose $\Theta \sim$ Dir (α).
(3) For each of the N words w_n:
(a) Choose a topic $z_n \sim$ Multinomial (Θ).
Choose a word w_n from $p(w_n \mid z_n, \beta)$, a multinomial probability conditioned on the topic z_n.

Box 1: The generative process of LDA (Blei et al., 2013; [19, 20]).

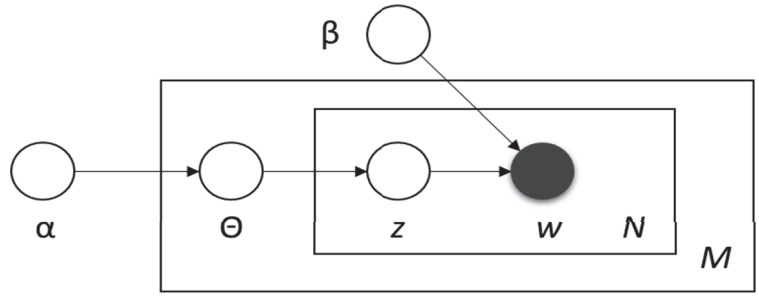


Figure 1: The graphical representation of LDA [22].

(2), and the probability of a corpus is computed as given by (3) [22]:

$$p(\Theta \mid \alpha) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \Theta_1^{\alpha_1 - 1} \cdots \Theta_k^{\alpha_k - 1} \tag{1}$$

$$p(\Theta, z, w \mid \alpha, \beta) = p(\Theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \Theta)\, p(w_n \mid z_n, \beta) \tag{2}$$

$$p(D \mid \alpha, \beta) = \prod_{d=1}^{M} \int p(\Theta_d \mid \alpha) \cdot \left(\prod_{n=1}^{N_d} \sum_{Z_{dn}} p(z_{dn} \mid \Theta_d)\, p(w_{dn} \mid z_{dn}, \beta) \right) d\Theta_d \tag{3}$$

In LDA, the computation of the posterior distribution of the hidden variables is an important inferential task. The exact inference of hidden variables is exponentially large. Hence, approximation algorithms (such as Laplace approximation, variational approximation, and Gibbs sampling) have been utilized in LDA process [61].

3.2. Ensemble Learning Methods. Ensemble learning methods aim to combine the predictions of multiple classification algorithms so that a classification model with higher predictive performance can be achieved [62]. In dependent methods, the outputs of former classifiers determine the outputs of following classifiers. In contrast, the outputs of classifiers are individually identified and combined to produce the final prediction in independent methods. Dependent ensemble methods include Boosting (e.g., AdaBoost algorithm) and independent methods include Bagging, Dagging, and Random Subspace. To examine the predictive performance of the proposed scheme, four well-known ensemble learning methods (namely, AdaBoost [63], Bagging [64], Random Subspace [65], and Stacking [66]) are considered.

3.3. Ensemble Pruning Methods. The ensemble pruning methods aim to identify optimal subset of classification algorithms to improve the predictive performance and computational efficiency of multiple classifier systems. To examine the predictive performance of proposed ensemble pruning algorithm, we have employed four ensemble pruning algorithms. These methods are the ensemble pruning methods from libraries of models [49], Bagging ensemble selection [67], LibD3C algorithm [68], and ensemble pruning based on combined diversity measures [21].

3.4. Swarm-Based Optimization Algorithms. Swarm-based optimization algorithms, including genetic algorithms, particle swarm optimization, firefly algorithm, cuckoo search algorithm, and bat algorithm, have been successfully employed on applications of data science, such as data clustering and data categorization [68]. In the proposed scheme, swarm-based optimization algorithms have been utilized to optimize the set of parameters of LDA-based topic modelling. In addition, the proposed ensemble pruning algorithm employs swarm-based optimization algorithms to group classifiers into clusters. In the empirical analysis, genetic algorithms [69], particle swarm optimization algorithm [70], firefly algorithm [71], cuckoo search algorithm [72], and bat algorithm [73] are utilized.

3.5. Cluster Validity Indices. This section briefly introduces four cluster validity indices (namely, the Bayesian information criterion, Calinski-Harabasz index, Davies-Bouldin index, and Silhouette index), which are utilized to evaluate the clustering quality of different configurations of LDA.

The Bayesian information criterion (BIC) is computed as given below:

$$\mathrm{BIC} = -\ln(\mathrm{L}) + v\ln(\mathrm{n}) \tag{4}$$

where n denotes the number of topics, L denotes the likelihood of parameters to generate data in the model, and v denotes the number of free parameters in Gaussian model [74]. The smaller the Bayesian information criterion, the better the generated model.

The Calinski-Harabasz index (CH) is the ratio of the traces of between cluster scatter matrix and the internal scatter matrix, which is computed as given below [74]:

$$\mathrm{CH}(K) = \frac{[trace\ B/K-1]}{[trace\ W/N-K]} \tag{5}$$

$$\text{trace B} = \sum_{k=1}^{K} |C_k| \left\| \overline{c_k} - \overline{x} \right\|^2 \tag{6}$$

$$\text{trace C} = \sum_{k=1}^{K}\sum_{i=1}^{N} w_{k,i} \left\| x_i - \overline{C_k} \right\|^2 \tag{7}$$

where K denotes the number of clusters, N denotes the number of data instances, $|C_k|$ denotes the number of elements in cluster C_k, x_i denotes a point within cluster C_k, B denotes the between-cluster scatter matrix, which represents the error sum of squares between different clusters, and W denotes the internal scatter matrix, which represents the squared differences of instances in a cluster. Here, trace of an n-by-n square matrix corresponds to the sum of the elements on the main diagonal [75].

The Davies-Bouldin index (DB) is a cluster validity index, which aims to maximize between-cluster distance and to minimize the distance between centroids of clusters and the other data points, that is defined as given by the following equation:

$$\mathrm{BD} = \frac{1}{c}\sum_{i=1}^{c} \max_{i\neq j} \left\{ \frac{d(X_i) + d(X_j)}{d(c_i, c_j)} \right\} \tag{8}$$

where c denotes the number of clusters, i and j correspond to cluster labels, $d(c_i, c_j)$ corresponds to distance between centroids of clusters, and X_i corresponds to a data point within cluster C_i. The smaller the DB criterion, the better the generated model.

The Silhouette index (SI) is defined as given by (9):

$$\mathrm{SI} = \frac{1}{N}\sum_{i} \left(\frac{1}{n_i} \sum_{x\in C_i} \frac{b(x) - a(x)}{\max[b(x), a(x)]} \right) \tag{9}$$

$$a(x) = \frac{1}{n_i - 1} \sum_{y\in C_{i, y\neq x}} d(x, y) \tag{10}$$

$$b(x) = \min_{j\neq i} \left[\frac{1}{n_i} \sum_{y\in C_j} d(x, y) \right] \tag{11}$$

where N denotes the number of clusters, n_i denotes the size of cluster C_i, $a(x)$ denotes the average distance between the ith instance and all instances in X_j, $b(x)$ denotes the minimum distance from i to the centroids of clusters not containing i.

3.6. Pairwise Diversity Measures. This section briefly introduces four diversity measures (namely, disagreement measure, Q-statistics, the correlation coefficient, and the double fault measure) which are utilized in the proposed ensemble classification scheme.

Q-statistics, the correlation coefficient ($p_{i,k}$), the disagreement measure (Dis), and the double fault measure (DF) among two classifiers D_i and D_k are computed using (12), (13), (14), and (15), respectively [76]:

$$Q_{i,k} = \frac{N^{11}N^{00} - N^{01}N^{10}}{N^{11}N^{00} + N^{01}N^{10}} \tag{12}$$

$$\rho_{i,k} = \frac{N^{11}N^{00} - N^{01}N^{10}}{\sqrt{\left(N^{11} + N^{10}\right)\left(N^{01} + N^{00}\right)\left(N^{11} + N^{01}\right)\left(N^{10} + N^{00}\right)}} \tag{13}$$

$$Dis_{i,k} = \frac{N^{01} + N^{10}}{N^{11} + N^{10} + N^{01} + N^{00}} \tag{14}$$

$$DF_{i,k} = \frac{N^{00}}{N^{11} + N^{10} + N^{01} + N^{00}} \tag{15}$$

where N^{11}, N^{00}, N^{10}, and N^{01} denote the number of correctly classified instances by the two classifiers, the number of incorrectly classified instances by the two classifiers, the number of instances correctly classified by D_i and incorrectly classified by D_k, and the number of instances correctly classified by D_k and incorrectly classified by D_i, respectively.

4. The Proposed Text Categorization Framework

The proposed text categorization framework combines the swarm-optimized Latent Dirichlet allocation and diversity-based hybrid ensemble pruning scheme. The rest of this section explains the methods utilized in the proposed biomedical text categorization framework.

4.1. Swarm-Optimized Latent Dirichlet Allocation. The latent Dirichlet allocation (LDA) is an efficient generative probabilistic model that can be employed to represent unstructured text documents in an efficient way. In general, LDA-based topic modelling involves the calibration of several parameters, summarized as follows:

(i) Number of topics in LDA-based topic modelling (k).

(ii) α parameter to control the topic distribution per document. A higher value for α parameter denotes better smoothing of topics for each document.

(iii) β parameter to model distributions of terms per topic.

In order to improve the computational complexity of LDA, LDA is usually employed in conjunction with an approximation method. In this work, we utilized Gibbs sampling

method in conjunction with LDA. In this way, the number of iterations (N) for sampling is also involved as an additional parameter value. Identifying appropriate parameter values of LDA with the optimal configuration is a challenging task. Without setting appropriate parameter values, LDA-based representation may degrade the predictive performance of classification schemes. Too low or too much number of topics can result in a poor predictive performance. Hence, finding an optimal configuration for LDA-based topic modelling involves extensive empirical analysis. Exhaustively enumerating possible parameter values for LDA to identify an optimal configuration involves high computational analysis with a wide range of parameter values.

In this paper, five metaheuristic algorithms (namely, genetic algorithms, particle swarm optimization, firefly algorithm, cuckoo search algorithm, and bat algorithm) are utilized to calibrate the parameters of LDA. In this scheme, values of all parameters of LDA are taken into consideration. Hence, various values for each parameter are evaluated to find an optimal configuration. In the presented problem, the first issue is to examine the merit of a particular LDA-based configuration. In order to evaluate the merit of a particular configuration of LDA before employing on a particular task, we have employed four internal cluster validity indices, namely, the Bayesian information criterion, Calinski-Harabasz index, Davies-Bouldin index, and Silhouette index. Higher clustering quality of a particular LDA-based configuration tends to yield higher predictive performance on LDA-based categorization tasks [19, 20]. For this reason, we seek to identify an LDA configuration which maximizes the overall clustering quality of LDA configuration.

Since exhaustively enumerating possible configurations for LDA can be computationally infeasible task, the identification of a parameter set which maximizes the overall clustering quality can be modelled as an optimization problem. In the presented scheme, five swarm-based optimization algorithms (namely, genetic algorithms, particle swarm optimization, firefly algorithm, cuckoo search algorithm, and bat algorithm) have been considered. The presented approach seeks to find an LDA configuration [k, α, β, N] which maximizes the clustering quality in terms of internal cluster validity indices (Bayesian information criterion, Calinski-Harabasz index, Davies-Bouldin index, and Silhouette index). The presented scheme starts with a randomly generated population of initial configuration. Then, randomly generated LDA configurations are utilized to cluster text documents. The merit of clusters is evaluated using four internal clustering validity indices and the swarm-based optimization algorithms have been utilized to optimize the parameter values. In Figure 2, the general structure of swarm-optimized LDA is summarized.

4.2. Diversity-Based Ensemble Pruning. Diversity-based ensemble pruning approach is a hybrid ensemble pruning scheme, which integrates combined pairwise diversity measures and swarm-based clustering algorithms. The presented ensemble pruning method consists of two main stages, namely, computation of pairwise diversity matrices among the base learning algorithms of the ensemble and swarm-based clustering on combined pairwise diversity matrix to obtain final base learning algorithms of the pruned ensemble.

The general structure of diversity-based ensemble pruning algorithm is presented in Figure 3. Initially, many different base learning algorithms (classification algorithms) from the model library with varying parameter values have been taken as the initial set of classifiers. The model library contains classification algorithms from five groups, namely, five Bayesian classifiers, fourteen function based classifiers, ten instance based classifiers, three rule based classifiers, and eight decision tree classifier which have been considered. The detailed description regarding the classification algorithms of the model library is presented in Table 2. Classification algorithms of the model library have been trained on the training set. In this way, the predictive characteristics of different learning algorithms have been obtained.

After training classification algorithms, pairwise diversity matrices are computed. The diversity and accuracy are two essential factors to build multiple classifier systems with high predictive performance. There are many pairwise and nonpairwise diversity measures presented in the literature. Different diversity measures concentrate on different aspects of the diversity and there is not a widely accepted definition for the term. Motivated by the success of the combined diversity measures in the ensemble pruning [21], we seek to find an appropriate subset of diversity measures. In this regard, we have conducted an experimental analysis with five widely utilized diversity measures (namely, Q-statistics, correlation coefficient, disagreement measure, double fault measure, and kappa statistics). Since there are five diversity measures, we have evaluated 2^5-1=31 different subset cases. The values obtained for each measure are normalized. Since the highest predictive performance is obtained by averaging the four diversity measures (Q-statistics, correlation coefficient, disagreement measure, and double fault measure), this configuration is utilized in the proposed ensemble pruning. For four pairwise diversity measures mentioned above, the diversity values of each pair of classifiers are computed using the validation set. Then, the combined pairwise diversity matrix is obtained from the four pairwise diversity matrices by averaging the diversity values of the individual diversity matrices.

After computation of the combined pairwise diversity matrix, clustering has been employed on the combined diversity matrix. Clustering has been widely employed technique for ensemble pruning, which aims to group classification algorithms into clusters such that the classifiers with the similar characteristics are assigned into the same cluster. By obtaining classifiers from the different clusters, a multiple classifier system with high diversity can be achieved. In this study, five metaheuristic clustering algorithms (namely, genetic algorithm based clustering, particle swarm clustering, firefly clustering, cuckoo clustering, and bat clustering) have been employed on the combined diversity matrix. Based on the clustering results, the classification algorithms have been assigned into a number of clusters.

On the empirical analysis with five metaheuristic clustering algorithms, the highest predictive performance is

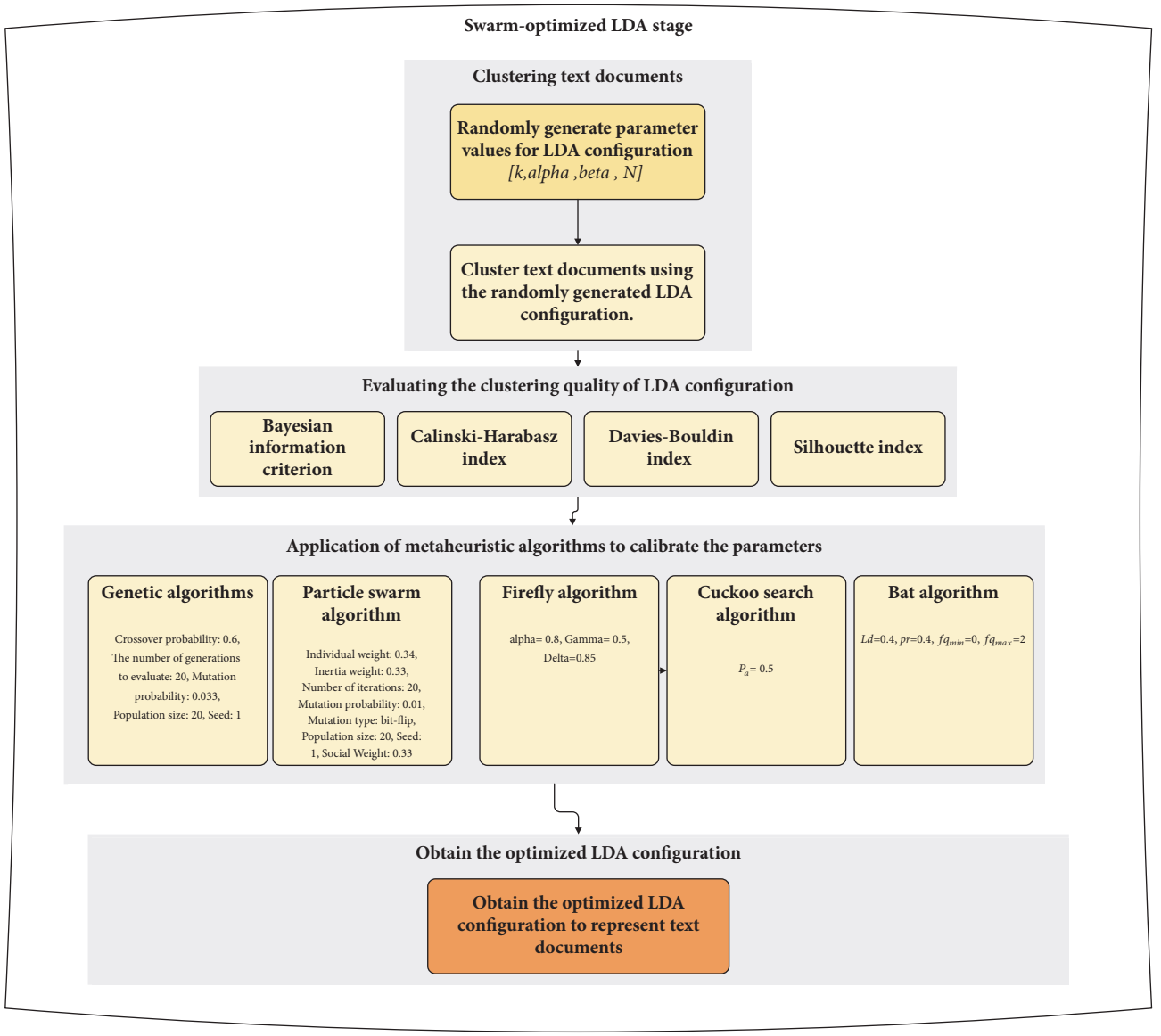


Figure 2: Swarm-optimized latent Dirichlet allocation.

achieved by firefly clustering algorithm. Hence, we utilized firefly clustering scheme to cluster classification algorithms on the combined diversity matrix based on their predictive characteristics. Let A denote an agent that consists of m n-dimensional points, a_i denote n-dimensional points in A, P denote a set containing of l n-dimensional points, p_i denote n-dimensional point contained in P, and $Dist(A,P)$ denote the distance between A and p; the general structure of firefly clustering algorithm utilized in the proposed scheme is outlined in Box 2.

After applying clustering algorithm on the combined pairwise diversity matrix, clustering results are utilized to select the classifiers of the pruned ensemble. In order to do so, classifiers of each cluster are ranked based on their predictive performance (in terms of classification accuracy). Then, one classifier with the highest predictive performance is selected from each cluster. Let N denote the number of clusters obtained at the end of firefly clustering algorithms, and one classifier has been selected from each classifier. In this way, N classifiers constitute the pruned ensemble. In order to combine the predictions of the selected classifiers, majority voting scheme is employed.

5. Experimental Analysis

In order to examine the predictive performance of the proposed biomedical text categorization scheme, an extensive empirical analysis has been performed. This section presents the datasets utilized in the analysis, the experimental procedure, and the experimental results.

5.1. Dataset. The experimental analysis has been conducted on five public biomedical text categorization datasets. These datasets are Oh5 collection, Oh10 collection, Oh15 collection,

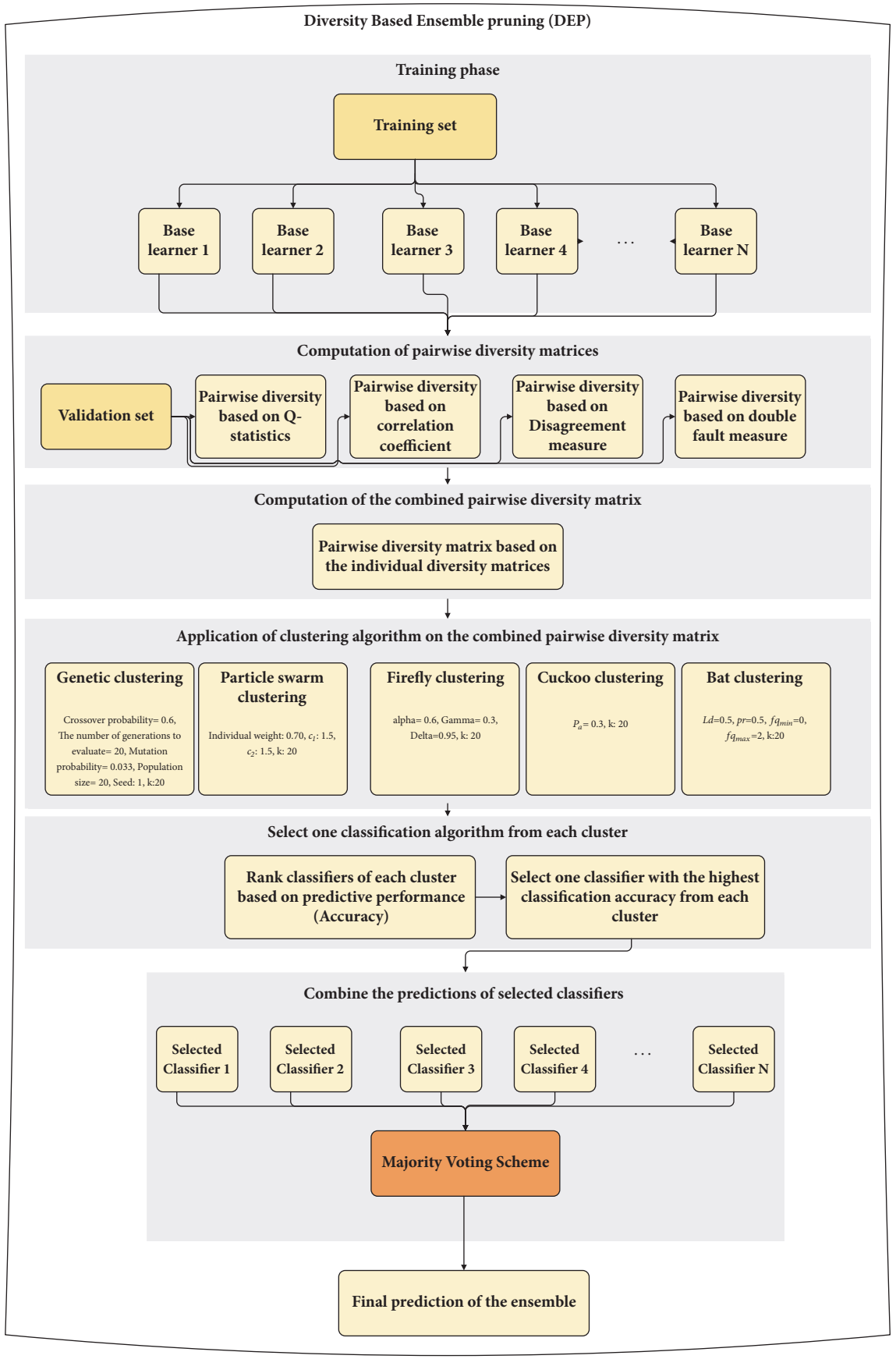


Figure 3: Diversity-based ensemble pruning approach.

Input: Data points: $P=\{p_1, p_2, \ldots, p_l\}$, α, δ and γ parameters.
Output: An agent A with the highest fitness value.
Initialize $A=\{A_1, A_2, \ldots, A_k\}$ agents,
for before stopping criterion has been met do
(i) For each A_i agent, calculate fitness function value F_i based on the following equation:

$$F(A) = \sum_{i=1}^{l} Distance(A, p_i)$$

$$Distance(A, p) = \min(\|a_1 - p\|, \|a_2 - p\|, \ldots, \|a_m - p\|).$$

(ii) For each A_i agent, compare fitness value of A_i with fitness value of A_j agent. If $F_i > F_j$ then, Update A_i agent based on the following equations:

$$a_j^i = a_j^i + d * e^{-\gamma * d^2} + \alpha * r$$

$$d = (a_j^x - a_j^i)$$

(iii) Update $\alpha=\alpha * \delta$.

Box 2: The general structure of firefly clustering algorithm.

Table 1: Descriptive information for the datasets.

Dataset	Number of documents	Number of terms	Average occurrence of terms	Number of classes
Oh5	918	3013	54.43	10
Oh10	1050	3239	55.63	10
Oh15	3101	54142	17.46	10
Ohscal	11162	11466	60.38	10
Ohsumed-400	9200	13512	55.14	12

Ohscal collection, and Ohsumed-400 collection [77]. Oh5, Oh10, Oh15, Ohscal, and Ohsumed-400 collections are part of OHSUMED collection. Each collection contains biomedical text collections. The basic descriptive information about biomedical text collections utilized in the empirical analysis has been summarized in Table 1, and the number of terms extracted after preprocessing is given.

5.2. Evaluation Metrics. In order to evaluate the predictive performance of the presented biomedical text categorization scheme, classification accuracy (ACC) and F-measure have been employed as the evaluation measure.

Classification accuracy is one of the most widely utilized measures in performance evaluation of classification algorithms. It is the proportion of the number of true positives and true negatives obtained by the classifiers in the total number of instances as given by the following equation:

$$ACC = \frac{TN + TP}{TP + FP + FN + TN} \tag{16}$$

where *TN, TP, FP,* and *FN* represent the number of true negatives, true positives, false positives, and false negatives, respectively.

F-measure is another common measure in performance evaluation of classification algorithms. F-measure is the harmonic mean of the precision and recall of a classification algorithm. It can take values between 0 and 1 and the higher values of F-measure indicate a better predictive performance. Based on the characteristics of datasets utilized in the empirical analysis, there are two variants of F-measure, namely, micro-averaged F-measure and macro-averaged F-measure. The micro-averaged F-measure extends F-measure to multiclass problems by averaging precision and recall values across all classes. However, F-measure and micro-averaged F-measure cannot focus entirely on rare classes [78]. Since some of the datasets utilized in the empirical analysis are imbalanced dataset, the macro-averaged F-measure is also utilized as another evaluation measure. The macro-averaged F-measure, which determines the average F-measure across all one-versus-all classes, is computed as given by (17):

$$\text{Macro} - \text{averaged F} - \text{measure} = \frac{1}{n}\sum_{i=1}^{n} \frac{2 \times Precision_i \times Recall_i}{Precision_i + Recall_i} \tag{17}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{18}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{19}$$

where *TP, FP,* and *FN* represent the number of true positives, false positives, and false negatives, respectively.

5.3. Experimental Procedure. In the experimental analysis, dataset is divided into tenfold (parts). In this scheme, sixfold is utilized for training, twofold is utilized for validation, and twofold is utilized for test. The experimental analysis is performed with the machine learning toolkit WEKA

Table 2: Classification algorithms used to build the model library.

Classifier Group	Classification Algorithms
Bayesian Classifiers (5)	Bayesian logistic regression (with Norm-based hyper-parameter selection), Bayesian logistic regression (with Cross-validated hyper-parameter selection), Bayesian logistic regression (with Specific value based hyper-parameter selection), Naive Bayes, Naive Bayes Multinomial
Function based classifiers (14)	FLDA, Kernel Logistic Regression (with Poly Kernel), Kernel Logistic Regression (with Normalized Poly Kernel), LibLINEAR (with L2-regularized logistic regression), LibLINEAR (with L2-regularized L2-loss support vector classification), LibLINEAR (with L1-regularized logistic regression), LibSVM (with radial basis function), LibSVM (with linear kernel), LibSVM (with polynomial kernel), LibSVM (with sigmoid kernel), Multi-layer perceptron, radial basis function networks, Logistic regression, Gaussian radial basis function networks
Instance based classifiers (10)	KNN (with K: 1), KNN (with K:2), KNN (with K:3), KNN (with K: 4), KNN (with K:5), KNN (with K:6), KNN (with K:7), KNN (with K:8), KNN (with K:9), KNN (with K:10)
Rule based classifiers (3)	FURIA (with Product T-norm), FURIA (with Minimum T-norm), RIPPER
Decision tree classifiers (8)	BFTree (Unpruned), BFTree (Post-pruning), BFTree (Pre-pruning), Functional Tree, C4.5 (J48), NBTree, Random Forest, Random Tree

Table 2 is reproduced from ONAN et al. [19, 20] (under the Creative Commons Attribution License/public domain).

(Waikato Environment for Knowledge Analysis) version 3.9, which is an open-source platform with many machine learning algorithms implemented in Java [79]. The presented classification scheme is also implemented in Java. In the empirical analysis on swarm-based latent Dirichlet allocation, Naïve Bayes algorithm and support vector machines are utilized as the base learning algorithms. In order to compare the presented multiple classifier system, four well-known ensemble methods (namely, AdaBoost, Bagging, Random Subspace, and Stacking) have been considered. For AdaBoost, Bagging, and Random Subspace algorithms, Naïve Bayes and support vector machines are utilized as the base learners. In the Stacking (stacked generalization), the classifier ensemble consisted of five base learners (namely, Naïve Bayes, support vector machines, logistic regression, Bayesian logistic regression, and linear discriminant analysis). For ensemble selection from libraries of models (ESM) and Bagging ensemble selection (BES), the same model library presented in Table 2 has been utilized [19, 20].

For evaluating ensemble pruning schemes, we have adopted the scheme outlined in [19, 20]. In the experimental analysis, ESM, BES, and LibD3C algorithms are considered with different parameter values. For ESM algorithm, four different schemes (namely, forward selection, backward elimination, forward-backward selection, and the best model scheme) have been considered. In ESM algorithm, root mean squared error (RMSE), classification accuracy (ACC), ROC area, precision, recall, and F-measure are considered as the evaluation measures. For BES algorithm, different bag sizes ranging from 10 to 100 are considered. In this algorithm, root mean squared error (RMSE), accuracy (ACC), ROC area, precision, recall, F-measure, and the combination of all metrics are employed as the evaluation measures. For LibD3C algorithm, five different ensemble combination rules (namely, average of probabilities, product of probabilities, majority voting, minimum probability, and maximum probability) are considered. In the experimental analysis, the highest predictive performances obtained from these algorithms are reported. In Table 3, the parameter values of metaheuristic algorithms utilized in swarm-based LDA are presented. In Table 4, parameters of metaheuristic clustering algorithms utilized in the ensemble pruning stage are given. The parameters of the metaheuristic algorithms utilized in the swarm-based LDA stage and the parameters of the metaheuristic algorithms utilized in the ensemble pruning stage are determined based on the benchmark empirical results for the algorithms [80, 81].

5.4. Experimental Results and Discussion. The presented biomedical text categorization framework consists of two main stages, namely, swarm-optimized latent Dirichlet allocation stage and diversity-based ensemble pruning stage.

Swarm-optimized latent Dirichlet allocation stage aims to estimate the parameters of LDA. In the empirical analysis on LDA, five different metaheuristic algorithms (namely, genetic algorithms, particle swarm optimization, firefly algorithm, cuckoo search algorithm, and bat algorithm) are considered. To evaluate the clustering quality of different configurations of LDA, four internal cluster validity indices (namely, the Bayesian information criterion, Calinski-Harabasz index, Davies-Bouldin index, and Silhouette index) are considered. In addition, the proposed scheme presents an ensemble pruning based on combined diversity measures and metaheuristic clustering. In the tables, the highest (the best) results achieved by a particular configuration are indicated as both boldface and underline and the second best results are indicated as both boldface and italics.

In order to evaluate the merit of swarm-optimized topic modelling in LDA, Table 5 presents the classification accuracies obtained by different LDA-based configurations with Naïve Bayes and support vector machine classifiers. To verify the impact of ensemble pruning method in the presented scheme, Table 6 presents the classification results obtained by conventional algorithms, ensemble learning methods, conventional ensemble pruning methods, and the proposed diversity-based ensemble pruning method. For the

Table 3: Parameters of the metaheuristics algorithms utilized in swarm-based LDA.

Metaheuristic algorithm	Parameter Values
Genetic algorithms	Crossover probability: 0.6, The number of generations to evaluate: 20, Mutation probability: 0.033, Population size: 20, Seed: 1
Particle swarm optimization	Individual weight: 0.34, Inertia weight: 0.33, Number of iterations: 20, Mutation probability: 0.01, Mutation type: bit-flip, Population size: 20, Seed: 1, Social Weight: 0.33
Firefly algorithm	α= 0.8, γ= 0.5, δ=0.85
Cuckoo search algorithm	P_a= 0.5
Bat algorithm	Ld=0.4, pr=0.4, fq_{min}=0, fq_{max}=2

Table 4: Parameters of the metaheuristics algorithms utilized in ensemble pruning.

Metaheuristic algorithm	Parameter Values
Genetic clustering	Crossover probability= 0.6, The number of generations to evaluate= 20, Mutation probability= 0.033, Population size= 20, Seed: 1, k:20
Particle swarm clustering	Individual weight: 0.70, c_1: 1.5, c_2: 1.5, k: 20
Firefly clustering	α= 0.6, γ= 0.3, δ=0.95, k: 20
Cuckoo clustering	P_a= 0.3, k: 20
Bat clustering	Ld=0.5, pr=0.5, fq_{min}=0, fq_{max}=2, k:20

Table 5: Classification accuracies obtained with different LDA-based configurations.

	Naive Bayes (NB)					Support Vector Machines (SVM)				
Configuration	oh5	oh10	oh15	ohscal	Ohsu-med	oh5	oh10	oh15	ohscal	Ohsu-med
LDA (k=50)	74.38	66.66	69.40	59.27	28.35	76.24	78.73	83.17	70.62	34.64
LDA (k=100)	70.85	63.64	67.44	60.05	29.56	78.28	78.25	83.23	73.23	38.82
LDA (k=150)	69.02	65.24	65.51	59.01	29.43	76.72	79.09	84.74	73.8	41.27
LDA (k=200)	66.17	64.01	63.61	58.93	27.99	77.33	77.93	84	74.19	41.82
GA-LDA (BIC)	75.16	67.24	74.70	71.66	35.45	77.98	69.03	75.12	73.62	35.83
PSO-LDA (BIC)	75.40	68.60	76.90	72.43	35.46	78.22	72.56	75.17	75.89	36.23
FA-LDA (BIC)	75.48	71.26	77.48	72.80	35.60	79.50	74.73	76.63	76.90	37.69
CSA-LDA (BIC)	76.66	71.96	78.77	72.94	35.65	79.56	75.97	77.96	77.02	37.94
BA-LDA (BIC)	78.82	72.21	79.77	73.02	36.58	79.85	76.53	78.89	77.34	38.89
GA-LDA (CH)	79.02	72.88	80.11	74.53	36.85	80.62	77.72	80.31	78.17	38.96
PSO-LDA (CH)	80.20	72.93	80.66	74.76	37.03	81.50	77.91	80.50	78.99	39.03
FA-LDA (CH)	81.20	72.99	80.72	75.13	37.75	81.80	77.99	80.55	79.09	39.03
CSA-LDA (CH)	81.40	73.12	81.71	76.02	38.34	82.61	78.01	80.78	79.82	39.03
BA-LDA (CH)	81.46	73.49	81.82	76.21	39.24	82.87	78.93	81.01	79.89	39.52
GA-LDA (DB)	84.46	76.22	84.13	78.71	40.50	84.73	80.95	85.88	82.46	43.02
PSO-LDA (DB)	84.60	80.07	85.14	79.21	42.57	85.13	81.11	86.17	84.22	43.51
FA-LDA (DB)	85.89	80.82	85.17	80.83	44.60	86.22	81.88	86.73	84.62	44.61
CSA-LDA (DB)	***86.42***	***80.97***	***86.10***	***81.69***	***45.21***	***86.79***	***82.00***	***86.96***	***85.07***	***46.67***
BA-LDA (DB)	**87.60**	**81.36**	**87.32**	**83.56**	**47.00**	**88.86**	**82.09**	**88.05**	**85.24**	**50.08**
GA-LDA (SI)	81.57	73.57	82.03	76.48	39.36	83.21	79.00	82.24	79.93	40.58
PSO-LDA (SI)	82.61	73.76	82.50	76.61	39.66	83.58	79.33	83.03	80.36	40.87
FA-LDA (SI)	83.19	74.18	82.88	77.47	39.68	83.69	79.41	83.11	80.95	40.95
CSA-LDA (SI)	83.78	75.11	83.01	78.06	39.69	83.84	80.83	84.47	81.82	41.12
BA-LDA (SI)	84.11	76.08	83.03	78.13	40.08	84.49	80.90	85.52	81.99	42.65

LDA: latent Dirichlet allocation, GA-LDA: genetic algorithm based LDA, PSO-LDA: particle swarm optimization based LDA, FA-LDA: firefly algorithm based LDA, CSA-LDA: cuckoo search algorithm based LDA, BA-LDA: bat algorithm based LDA, BIC: Bayesian information criterion, CH: Calinski-Harabasz index, DB: Davies-Bouldin index, and SI: Silhouette index.

Table 6: Classification results obtained by conventional algorithms and the proposed diversity-based ensemble pruning (with LDA (k=50) based representation).

Classification algorithm	oh5	oh10	oh15	ohscal	ohsumed
NB	75.19	67.43	70.77	60.24	29.41
SVM	77.59	80.29	84.47	71.58	34.72
Bagging+NB	76.08	69.77	70.94	60.21	29.21
Bagging+SVM	84.36	77.20	79.07	71.92	35.98
AdaBoost+NB	73.53	68.07	70.26	60.09	29.60
AdaBoost+SVM	84.06	77.19	78.88	72.08	35.03
RandomSubspace+NB	74.75	67.29	68.51	57.58	28.60
RandomSubspace+SVM	78.02	69.89	71.22	67.65	31.80
Stacking	83.78	81.32	81.69	60.02	40.76
ESM	79.25	79.07	78.91	72.52	37.84
BES	80.11	80.61	81.08	73.02	40.04
LibD3C	82.86	82.93	84.51	74.86	41.17
CDM	84.77	84.13	85.32	76.45	43.55
DEP (Genetic clustering)	81.61	81.96	84.64	74.21	43.27
DEP (PSO clustering)	80.91	81.41	83.31	73.98	***45.73***
DEP (Firefly clustering)	**86.52**	**86.08**	**86.29**	**77.47**	**47.48**
DEP (Cuckoo clustering)	***85.06***	83.00	***85.84***	***76.81***	45.43
DEP (Bat clustering)	84.47	***84.18***	82.11	72.70	44.13

NB: Naïve Bayes algorithm, SVM: support vector machines, ESM: ensemble selection from libraries of models, BES: Bagging ensemble selection, LibD3C: hybrid ensemble pruning based on k-means and dynamic selection, CDM: ensemble pruning based on combined diversity measures, and DEP: the proposed diversity-based ensemble pruning.

Table 7: Comparison of the proposed text categorization scheme with conventional classifiers, ensemble learners, and ensemble pruning method (with BA-LDA (DB) based representation).

Classification algorithm	oh5	oh10	oh15	ohscal	ohsumed
NB	87.67	81.42	87.44	83.64	47.09
SVM	88.97	82.22	88.16	85.32	50.08
Bagging+NB	89.32	83.35	88.87	83.47	48.52
Bagging+SVM	88.03	84.84	87.86	83.92	50.73
AdaBoost+NB	89.77	83.60	87.48	86.18	51.18
AdaBoost+SVM	88.18	84.95	87.35	86.29	51.85
RandomSubspace+NB	88.32	83.96	86.66	88.09	50.70
RandomSubspace+SVM	88.56	84.11	89.58	88.29	50.29
Stacking	88.28	86.87	88.93	84.90	53.84
ESM	88.58	86.66	90.25	88.48	51.94
BES	89.29	86.00	90.98	89.12	52.47
LibD3C	90.35	87.95	91.27	90.48	53.41
CDM	***91.51***	***89.61***	***93.17***	***91.33***	***54.47***
Proposed scheme	**93.14**	**91.29**	**93.76**	**92.14**	**58.17**

NB: Naïve Bayes algorithm, SVM: support vector machines, ESM: ensemble selection from libraries of models, BES: Bagging ensemble selection, LibD3C: hybrid ensemble pruning based on k-means and dynamic selection, and CDM: ensemble pruning based on combined diversity measures.

results reported in Table 6, the biomedical text categorization datasets are represented with LDA (k=50); i.e., swarm-optimized latent Dirichlet allocation stage has not been applied for the results presented in Table 6 to examine the predictive performance of the proposed ensemble pruning scheme. Finally, Table 7 compares the predictive performance of conventional algorithms, ensemble learning methods, conventional ensemble pruning methods, and the proposed diversity-based ensemble pruning method when swarm-optimized latent Dirichlet allocation stage has been applied to represent the dataset.

As can be observed from the classification accuracies presented in Table 5, the performance of LDA-based representation schemes generally enhances with the use

of metaheuristic algorithms in conjunction with LDA to estimate the parameters of it. Among the different metaheuristic algorithms, the highest predictive performance is obtained by bat algorithm based LDA with Davies-Bouldin index based evaluation. The second highest predictive performance is obtained by cuckoo search algorithm based LDA with Davies-Bouldin index based evaluation. Regarding the performance of different evaluation measures, the highest performance is achieved by Davies-Bouldin index based configurations. The second predictive performance is achieved by Silhouette index based configurations, which is followed by Calinski-Harabasz index based configurations. Regarding the performance of conventional LDA-based representation schemes, the highest predictive performance is generally achieved when k=50. The predictive performance patterns obtained by different LDA-based configurations with Naïve Bayes algorithm are valid for LDA-based configurations with support vector machines algorithm.

In the empirical analysis on the ensemble pruning, five swarm-based clustering algorithms (namely, genetic clustering, particle swarm-based clustering, firefly clustering, cuckoo clustering, and bat clustering) have been considered. Regarding the predictive performance obtained by conventional classification algorithms, support vector machines algorithm outperforms Naïve Bayes algorithm for the compared datasets. In addition, Bagging ensemble of Naïve Bayes algorithm yields better predictive performance compared to Naïve Bayes algorithm. In general, the predictive performance is enhanced with the use of conventional ensemble learning methods (namely, Bagging, AdaBoost, and Random Subspace algorithm). As can be seen from the results reported in Table 6, conventional ensemble pruning methods outperform the conventional classification algorithms and ensemble learning schemes. In addition, hybrid ensemble pruning schemes (the proposed diversity-based ensemble pruning method, LibD3C algorithm, and ensemble pruning based on combined diversity measures) outperform the other ensemble pruning schemes (ensemble selection from libraries of models and Bagging ensemble selection). The highest predictive performance is obtained by the proposed diversity-based ensemble pruning scheme with firefly clustering. The second highest predictive performance is generally obtained by the proposed diversity-based ensemble pruning scheme with cuckoo clustering.

Based on the extensive empirical analysis with different metaheuristic algorithms in swarm-based LDA and with different clustering algorithms in diversity-based ensemble pruning algorithm, the highest predictive performance is obtained by bat algorithm based LDA with Davies-Bouldin index and diversity-based ensemble pruning with firefly clustering. In Table 7, the predictive performance of the proposed biomedical text categorization scheme is compared with two classification algorithms (namely, Naïve Bayes algorithm and support vector machines), four ensemble methods (namely, Bagging, AdaBoost, Random Subspace, and Stacking), and four ensemble pruning methods (namely, ensemble selection from libraries of models, Bagging ensemble selection, LibD3C algorithm, and ensemble pruning based on combined diversity measures). For the results reported in Table 7, the biomedical text categorization datasets are represented with bat algorithm based LDA with Davies-Bouldin index (BA-LDA (DB)). As can be observed from the results outlined in Table 7, the proposed scheme outperforms the conventional classifiers, ensemble learning methods, and ensemble pruning methods.

In addition to classification accuracy, the predictive performances of classification algorithms, ensemble learning methods, and ensemble pruning methods have been also examined in terms of the macro-averaged F-measure. In Table 8, the macro-averaged F-measure results obtained by different LDA-based configurations with Naïve Bayes and support vector machine classifiers are presented. Regarding the macro-averaged F-measure results presented in Table 8, the highest predictive performance is obtained by bat algorithm based LDA with Davies-Bouldin index based representation. The same patterns obtained in terms of classification accuracies presented in Table 5 are also valid for F-measure based results. Hence, the utilization of metaheuristic optimization algorithms in conjunction with LDA to calibrate its hyper-parameters enhances the predictive model.

To examine the performance improvement achieved by the proposed ensemble pruning scheme, Table 9 presents the macro-averaged F-measure values obtained by conventional algorithms, ensemble learning methods, conventional ensemble pruning methods, and the proposed diversity-based ensemble pruning method. For the results reported in Table 9, the biomedical text categorization datasets are represented with LDA (k=50); i.e., swarm-optimized latent Dirichlet allocation stage has not been applied for the results presented in Table 9. Regarding the macro-averaged F-measure results presented in Table 9, the highest predictive performance is obtained by the proposed diversity-based ensemble pruning scheme with firefly clustering. The second highest predictive performance is generally obtained by the proposed diversity-based ensemble pruning scheme with cuckoo clustering and ensemble pruning based on combined diversity.

In Table 10, the macro-averaged F-measure results obtained by classification algorithms, ensemble learning methods, and ensemble pruning methods are presented. For the results reported in Table 10, the biomedical text categorization datasets are represented with bat algorithm based LDA with Davies-Bouldin index (BA-LDA (DB)). Regarding the macro-averaged F-measure results, the proposed scheme outperforms the conventional classifiers, ensemble learning methods, and ensemble pruning methods.

To statistically validate the results obtained in the empirical analysis, we have performed the two-way ANOVA (analysis of variance) test in the Minitab statistical program. The two-way ANOVA test is an extension of the one-way ANOVA test, which aims to evaluate the effect of two different categorical independent variables on one dependent variable. In two-way ANOVA test, both the main effect of each independent variable and their interactions are taken into assessment. The results for the two-way ANOVA test of overall results (in terms of classification accuracy) are presented in Table 11, where DF, SS, MS, F, and P denote degrees of freedom,

Table 8: The macro-averaged F-measure results obtained with different LDA-based configurations.

	Naive Bayes (NB)					Support Vector Machines (SVM)				
Configuration	oh5	oh10	oh15	ohscal	Ohsu-med	oh5	oh10	oh15	ohscal	Ohsu-med
LDA (k=50)	0.75	0.68	0.71	0.61	0.30	0.77	0.80	0.85	0.73	0.36
LDA (k=100)	0.72	0.65	0.69	0.62	0.31	0.79	0.80	0.85	0.75	0.40
LDA (k=150)	0.70	0.67	0.67	0.61	0.31	0.77	0.81	0.86	0.76	0.43
LDA (k=200)	0.67	0.65	0.65	0.61	0.29	0.78	0.80	0.86	0.76	0.44
GA-LDA (BIC)	0.76	0.69	0.76	0.74	0.37	0.79	0.70	0.77	0.76	0.37
PSO-LDA (BIC)	0.76	0.70	0.78	0.75	0.37	0.79	0.74	0.77	0.78	0.38
FA-LDA (BIC)	0.76	0.73	0.79	0.75	0.37	0.80	0.76	0.78	0.79	0.39
CSA-LDA (BIC)	0.77	0.73	0.80	0.75	0.37	0.80	0.78	0.80	0.79	0.40
BA-LDA (BIC)	0.80	0.74	0.81	0.75	0.38	0.81	0.78	0.81	0.80	0.41
GA-LDA (CH)	0.80	0.74	0.82	0.77	0.38	0.81	0.79	0.82	0.81	0.41
PSO-LDA (CH)	0.81	0.74	0.82	0.77	0.39	0.82	0.79	0.82	0.81	0.41
FA-LDA (CH)	0.82	0.74	0.82	0.77	0.39	0.83	0.80	0.82	0.82	0.41
CSA-LDA (CH)	0.82	0.75	0.83	0.78	0.40	0.83	0.80	0.82	0.82	0.41
BA-LDA (CH)	0.82	0.75	0.83	0.79	0.41	0.84	0.81	0.83	0.82	0.41
GA-LDA (DB)	0.85	0.78	0.86	0.81	0.42	0.86	0.83	0.88	0.85	0.45
PSO-LDA (DB)	0.85	***0.82***	0.87	0.82	0.44	0.86	0.83	0.88	0.87	0.45
FA-LDA (DB)	***0.87***	***0.82***	0.87	0.83	0.46	0.87	***0.84***	***0.89***	0.87	0.46
CSA-LDA (DB)	***0.87***	**0.83**	***0.88***	***0.84***	***0.47***	***0.88***	***0.84***	***0.89***	***0.88***	***0.49***
BA-LDA (DB)	**0.88**	**0.83**	**0.89**	**0.86**	**0.49**	**0.90**	**0.84**	**0.90**	**0.88**	**0.52**
GA-LDA (SI)	0.82	0.75	0.84	0.79	0.41	0.84	0.81	0.84	0.82	0.42
PSO-LDA (SI)	0.83	0.75	0.84	0.79	0.41	0.84	0.81	0.85	0.83	0.43
FA-LDA (SI)	0.84	0.76	0.85	0.80	0.41	0.85	0.81	0.85	0.83	0.43
CSA-LDA (SI)	0.85	0.77	0.85	0.80	0.41	0.85	0.82	0.86	0.84	0.43
BA-LDA (SI)	0.85	0.78	0.85	0.81	0.42	0.85	0.83	0.87	0.85	0.44

LDA: latent Dirichlet allocation, GA-LDA: genetic algorithm based LDA, PSO-LDA: particle swarm optimization based LDA, FA-LDA: firefly algorithm based LDA, CSA-LDA: cuckoo search algorithm based LDA, BA-LDA: bat algorithm based LDA, BIC: Bayesian information criterion, CH: Calinski-Harabasz index, DB: Davies-Bouldin index, and SI: Silhouette index.

adjusted sum of squares, adjusted mean square, F-Value, and probability value, respectively. Degrees of freedom are the amount of information in the data. The adjusted sum of squares term (SS) denotes the amount of variation in the response data that is explained by each term of the model. F-statistics (F) is the test statistic to identify whether a term is associated with the response and the probability value (P) is used to determine the statistical significance of the terms and model. The results presented in Table 11 are divided into three parts. The upper part of the table denotes the statistical analysis of results on the different LDA-based configurations, the middle part of the table denotes the statistical analysis of results on ensemble pruning, and the lower part of the table denotes the statistical analysis of results on conventional classifiers, ensemble learning methods, and ensemble pruning methods. For two-way ANOVA test, two different factors (different datasets and different algorithmic configurations) are taken as categorical independent variables. In addition, the interaction among these factors is also taken into consideration. According to the results presented in Table 11, probability value is $P<0.001$ for different factors and their interactions. Hence, there are statistically meaningful differences between the predictive performances of compared methods. The performance gain obtained by swarm-optimized LDA is statistically meaningful. Similarly, the performance gain obtained by the proposed ensemble pruning method is also statistically meaningful ($P<0.001$).

The results for the two-way ANOVA test of overall results (in terms of the macro-averaged F-measure values) are presented in Table 12. According to the results presented in Table 12, there are statistically meaningful differences between the predictive performances of compared methods ($P<0.001$).

In Figure 4, the confidence intervals for the mean values of classification accuracies obtained by the different LDA-based configuration schemes are presented. Similarly, in Figure 5, the confidence intervals for the mean values of classification accuracies obtained by the conventional classifiers, ensemble learners, and ensemble pruning methods are presented. For results depicted in Figure 5, the biomedical text categorization datasets are represented with LDA (k=50); i.e., swarm-optimized latent Dirichlet allocation stage has not been applied. In contrast, in Figure 6, the confidence intervals for the mean values of classification accuracies obtained by the conventional classifiers, ensemble learners, and ensemble pruning methods are given. In Figure 6, swarm-optimized latent Dirichlet allocation stage has been applied to represent the dataset. For the statistical significance of results,

Table 9: The macro-averaged F-measure results obtained by conventional algorithms and the proposed diversity-based ensemble pruning (with LDA (k=50) based representation).

Classification algorithm	oh5	oh10	oh15	ohscal	ohsumed
NB	0.76	0.68	0.72	0.61	0.30
SVM	0.78	0.81	0.86	0.73	0.35
Bagging+NB	0.77	0.70	0.72	0.61	0.30
Bagging+SVM	0.85	0.78	0.81	0.73	0.37
AdaBoost+NB	0.74	0.69	0.72	0.61	0.31
AdaBoost+SVM	0.85	0.78	0.80	0.74	0.36
RandomSubspace+NB	0.76	0.68	0.70	0.59	0.29
RandomSubspace+SVM	0.79	0.71	0.73	0.69	0.33
Stacking	0.84	0.80	0.81	0.72	0.38
ESM	0.80	0.81	0.81	0.74	0.39
BES	0.81	0.82	0.83	0.75	0.41
LibD3C	0.84	0.85	0.86	0.76	0.42
CDM	***0.86***	***0.86***	***0.87***	***0.78***	0.45
DEP (Genetic clustering)	0.82	0.84	0.86	0.76	0.45
DEP (PSO clustering)	0.82	0.83	0.85	0.75	***0.47***
DEP (Firefly clustering)	**0.87**	**0.88**	**0.88**	**0.79**	**0.49**
DEP (Cuckoo clustering)	***0.86***	0.85	**0.88**	***0.78***	***0.47***
DEP (Bat clustering)	0.85	***0.86***	0.84	0.74	0.45

NB: Naïve Bayes algorithm, SVM: support vector machines, ESM: ensemble selection from libraries of models, BES: Bagging ensemble selection, LibD3C: hybrid ensemble pruning based on k-means and dynamic selection, CDM: ensemble pruning based on combined diversity measures, and DEP: the proposed diversity-based ensemble pruning.

Table 10: The macro-averaged F-measure results of methods (with BA-LDA (DB) based representation).

Classification algorithm	oh5	oh10	oh15	ohscal	ohsumed
NB	0.89	0.82	0.88	0.84	0.48
SVM	0.90	0.83	0.89	0.86	0.51
Bagging+NB	0.90	0.84	0.90	0.84	0.49
Bagging+SVM	0.89	0.86	0.89	0.85	0.51
AdaBoost+NB	0.91	0.84	0.88	0.87	0.52
AdaBoost+SVM	0.89	0.86	0.88	0.87	0.52
RandomSubspace+NB	0.90	0.86	0.88	0.90	0.52
RandomSubspace+SVM	0.90	0.86	0.91	0.90	0.51
Stacking	0.90	0.87	0.91	0.88	0.54
ESM	0.90	0.88	0.92	0.90	0.53
BES	0.93	0.90	0.95	0.93	0.55
LibD3C	0.94	0.92	0.95	0.94	0.56
CDM	***0.95***	***0.93***	***0.97***	***0.95***	***0.57***
Proposed scheme	**0.97**	**0.95**	**0.98**	**0.96**	**0.61**

NB: Naïve Bayes algorithm, SVM: support vector machines, ESM: ensemble selection from libraries of models, BES: Bagging ensemble selection, LibD3C: hybrid ensemble pruning based on k-means and dynamic selection, and CDM: ensemble pruning based on combined diversity measures.

confidence intervals are divided into regions denoted by red dashed lines. As the interval plots indicate, the predictive performances obtained by the swarm-optimized LDA (BA-LDA (DB)) and DEP (firefly clustering) are statistically significant.

In Figure 7, average execution times of compared algorithms have been presented in seconds. As can be observed from Figure 7, average execution times on base learning algorithms (Naïve Bayes and support vector machines) are the lowest. Conventional ensemble learning methods generally enhance the predictive performance of the conventional base learning algorithms. However, ensemble learning methods involve more execution times. Compared to the ensemble learning methods, ensemble pruning schemes have more execution time. The highest execution time is involved in ensemble pruning based on combined diversity measures (CDM) and the second highest execution time is required in the proposed classification scheme (DEP-firefly clustering).

Table 11: Two-way ANOVA test results of classification accuracy values.

Statistical analysis of results on different LDA-based configurations					
Source	**DF**	**SS**	**MS**	**F**	**P**
Configuration	23	4073.9	177.1	90.50	P<0.001
Dataset	4	60336.7	15084.2	7707.50	P<0.001
Classifier	1	881.0	881.0	450.15	P<0.001
Configuration∗Dataset	92	334.0	3.6	1.85	P<0.001
Configuration∗Classifier	23	932.9	40.6	20.73	P<0.001
Dataset∗Classifier	4	106.3	26.6	13.57	P<0.001
Error	92	180.1	2.0		
Total	239	66844.8			
Statistical analysis of results on classifiers and ensemble pruning methods (with LDA (k=50) based representation).					
Source	**DF**	**SS**	**MS**	**F**	**P**
Configuration	17	2691.7	158.34	25.86	P<0.001
Dataset	4	23128.7	5782.17	944.48	P<0.001
Error	68	416.3	6.12		
Total	89				
Statistical analysis of results on conventional classifiers, ensemble learners, and ensemble pruning methods (with BA-LDA (DB) based representation).					
Source	**DF**	**SS**	**MS**	**F**	**P**
Configuration	13	324.5	24.96	17.81	P<0.001
Dataset	4	14736.0	3684.00	2628.98	P<0.001
Error	52	72.9	1.40		
Total	69	15133.4			

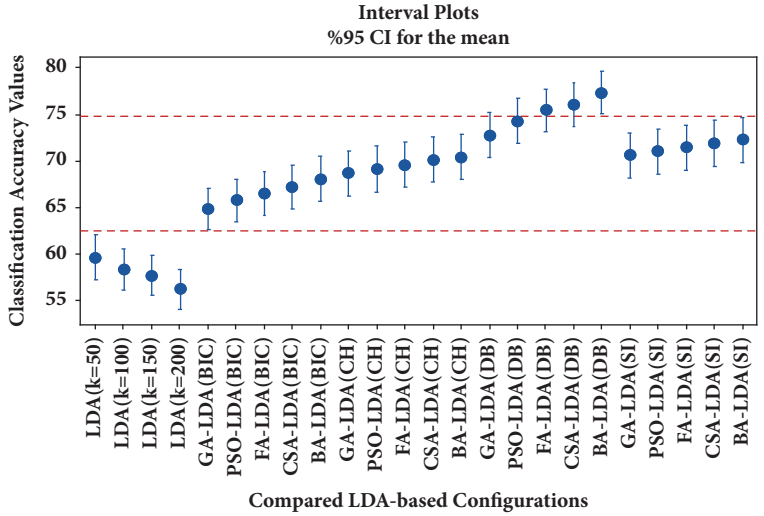


Figure 4: Interval plots for compared LDA-based configurations.

Table 12: Two-way ANOVA test results of the macro-averaged F-measure.

Statistical analysis of results on different LDA-based configurations					
Source	**DF**	**SS**	**MS**	**F**	**P**
Configuration	23	0.42777	0.01860	91.27	P<0.001
Dataset	4	5.99867	1.49967	7359.42	P<0.001
Classifier	1	0.09263	0.09263	454.58	P<0.001
Configuration∗Dataset	92	0.03536	0.00038	1.89	P<0.001
Configuration∗Classifier	23	0.09800	0.00426	20.91	P<0.001
Dataset∗Classifier	4	0.01123	0.00281	13.78	P<0.001
Error	92	0.01875	0.00020		
Total	239	6.68241			
Statistical analysis of results on classifiers and ensemble pruning methods (with LDA (k=50) based representation).					
Source	**DF**	**SS**	**MS**	**F**	**P**
Configuration	17	0.27733	0.016314	23.26	P<0.001
Dataset	4	2.41143	0.692858	859.46	P<0.001
Error	68	0.04770	0.000701		
Total	89	2.73646			
Statistical analysis of results on conventional classifiers, ensemble learners, and ensemble pruning methods (with BA-LDA (DB) based representation).					
Source	**DF**	**SS**	**MS**	**F**	**P**
Configuration	13	0.03613	0.002780	14.68	P<0.001
Dataset	4	1.53718	0.384296	2029.89	P<0.001
Error	52	0.00984	0.000189		
Total	69	1.58316			

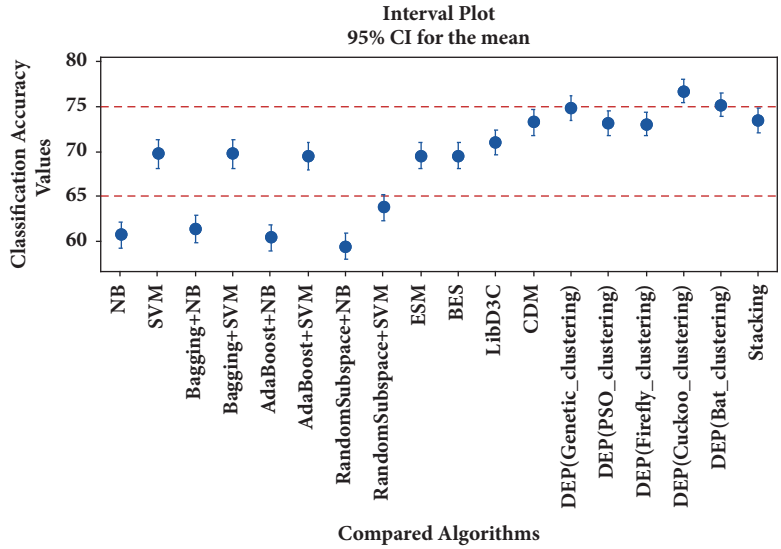


Figure 5: Interval plots for classifiers and ensemble pruning methods.

Metaheuristic optimization methods are well-established techniques on tuning the parameters. Hence, there is a trade-off between predictive performance and execution times.

6. Conclusion

In this work, we propose a novel biomedical text classification scheme based on swarm-optimized latent Dirichlet allocation and diversity-based ensemble pruning. Biomedical text categorization is an important research direction due to the immense quantity of unstructured information available. The latent Dirichlet allocation (LDA) is a popular representation scheme for text documents, which can yield better performance than other linguistic representation schemes, such as latent semantic analysis and probabilistic latent semantic analysis. We found out that the identification of appropriate parameter values is very important to the performance of LDA. In addition, it has been experimentally validated that the use of metaheuristic optimization algorithms to calibrate the parameters of LDA yields promising results on biomedical text categorization. The presented text classification scheme also employs an ensemble pruning approach

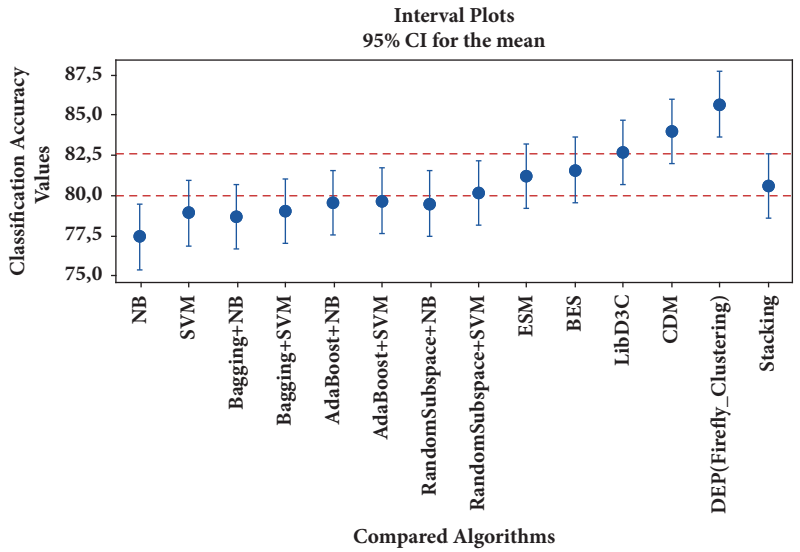


Figure 6: Interval plots for compared algorithms.

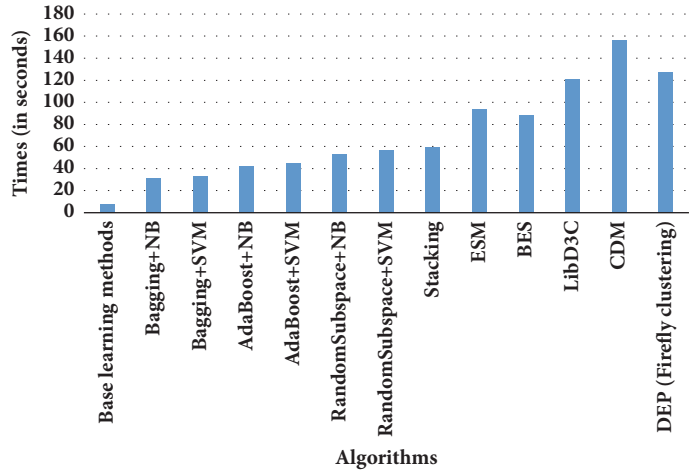


Figure 7: Average execution times (in seconds) for compared algorithms.

based on combined diversity measures to identify a robust multiple classifier system with high predictive performance. The presented ensemble pruning approach combines four different diversity measures (namely, disagreement measure, Q-statistics, the correlation coefficient, and the double fault measure). In addition, the scheme employs the swarm-based clustering algorithm. The experimental results indicate that the proposed multiple classifier system outperforms the conventional classification algorithms, ensemble learning, and ensemble pruning methods.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] E. S. Chen, G. Hripcsak, H. Xu, M. Markatou, and C. Friedman, "Automated acquisition of disease–drug knowledge from biomedical and clinical documents: an initial study," *Journal of the American Medical Informatics Association*, vol. 15, no. 1, pp. 87–98, 2008.

[2] R. Rodriguez-Esteban, "Biomedical text mining and its applications," *PLoS Computational Biology*, vol. 5, no. 12, Article ID e1000597, 2009.

[3] R. L. Figueroa and C. A. Flores, "Extracting Information from Electronic Medical Records to Identify the Obesity Status of

a Patient Based on Comorbidities and Bodyweight Measures," *Journal of Medical Systems*, vol. 40, no. 8, pp. 1–9, 2016.

[4] J. Urbain, "Mining heart disease risk factors in clinical text with named entity recognition and distributional semantic models," *Journal of Biomedical Informatics*, vol. 58, pp. S143–S149, 2015.

[5] T. G. Soldatos, S. I. O'Donoghue, V. P. Satagopam et al., "Martini: using literature keywords to compare gene sets," *Nucleic Acids Research*, vol. 38, no. 1, pp. 26–38, 2010.

[6] C. A. Trugenberger, C. Wälti, D. Peregrim, M. E. Sharp, and S. Bureeva, "Discovery of novel biomarkers and phenotypes by semantic technologies," *BMC Bioinformatics*, vol. 14, no. 1, article 51, 2013.

[7] A. Holzinger, J. Schantl, M. Schroettner, C. Seifert, and K. Verspoor, "Biomedical text mining: state-of-the-art, open problems and future challenges," in *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics*, pp. 271–300, Springer, Berlin, Germany, 2014.

[8] A. Onan and S. Korukoğlu, "A feature selection model based on genetic rank aggregation for text sentiment classification," *Journal of Information Science*, 2017.

[9] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval*, vol. 1, No. 1, p. 496, Cambridge University Press, Cambridge, UK, 2008.

[10] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," *Machine Learning: ECML-98*, pp. 137–142, 1998.

[11] T. K. Landauer, D. Laham, B. Rehder, and M. E. Schreiner, "How well can passage meaning be derived without using word order? A comparison of Latent Semantic Analysis and humans," in *Proceedings of the 19th annual meeting of the Cognitive Science Society*, pp. 412–417, 1997.

[12] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, article 391, 1990.

[13] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 1999*, pp. 50–57, ACM, August 1999.

[14] M. Girolami and A. Kabán, "Sequential activity profiling: Latent dirichlet allocation of Markov chains," *Data Mining and Knowledge Discovery*, vol. 10, no. 3, pp. 175–196, 2005.

[15] T. G. Dietterich, "Ensemble methods in machine learning," in *International workshop on multiple classifier systems*, pp. 1–15, Springer, Berlin, Germany, 2000.

[16] J. Mendes-Moreira, C. Soares, A. M. Jorge, and J. F. D. Sousa, "Ensemble approaches for regression: A survey," *ACM Computing Surveys (CSUR)*, vol. 45, no. 1, article 10, 2012.

[17] F. Roli, G. Giacinto, and G. Vernazza, "Methods for designing multiple classifier systems," *Lecture Notes in Computer Science*, vol. 2096, pp. 78–87, 2001.

[18] Z. Zhou, J. Wu, and W. Tang, "Ensembling neural networks: many could be better than all," *Artificial Intelligence*, vol. 137, no. 1-2, pp. 239–263, 2002.

[19] A. Onan, H. Bulut, and S. Korukoglu, "An improved ant algorithm with LDA-based representation for text document clustering," *Journal of Information Science*, vol. 43, no. 2, pp. 275–292, 2017.

[20] A. Onan, S. Korukoğlu, and H. Bulut, "A hybrid ensemble pruning approach based on consensus clustering and multi-objective evolutionary algorithm for sentiment classification," *Information Processing & Management*, vol. 53, no. 4, pp. 814–833, 2017.

[21] G. D. C. Cavalcanti, L. S. Oliveira, T. J. M. Moura, and G. V. Carvalho, "Combining diversity measures for ensemble pruning," *Pattern Recognition Letters*, vol. 74, pp. 38–45, 2016.

[22] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, no. 4-5, pp. 993–1022, 2003.

[23] K. Tian, M. Revelle, and D. Poshyvanyk, "Using latent dirichlet allocation for automatic categorization of software," in *Proceedings of the 6th IEEE International Working Conference on Mining Software Repositories, 2009. MSR'09*, pp. 163–166, IEEE, 2009.

[24] Z. Zhai, B. Liu, H. Xu, and P. Jia, "Constrained LDA for grouping product features in opinion mining," *Advances in Knowledge Discovery and Data Mining*, pp. 448–459, 2011.

[25] Q. Wu, C. Zhang, Q. Hong, and L. Chen, "Topic evolution based on LDA and HMM and its application in stem cell research," *Journal of Information Science*, vol. 40, no. 5, pp. 611–620, 2014.

[26] A. Bagheri, M. Saraee, and F. De Jong, "ADM-LDA: An aspect detection model based on topic modelling using the structure of review sentences," *Journal of Information Science*, vol. 40, no. 5, pp. 621–636, 2014.

[27] L. Hong and B. D. Davison, "Empirical study of topic modeling in twitter," in *Proceedings of the first workshop on social media analytics*, pp. 80–88, ACM, 2010.

[28] Z. Chen, Y. Huang, J. Tian, X. Liu, K. Fu, and T. Huang, "Joint model for subsentence-level sentiment analysis with Markov logic," *Journal of the Association for Information Science and Technology*, vol. 66, no. 9, pp. 1913–1922, 2015.

[29] A. Onan, S. Korukoglu, and H. Bulut, "LDA-based Topic Modelling in Text Sentiment Classification: An Empirical Analysis," *International Journal of Computational Linguistics and Applications*, vol. 7, no. 1, pp. 101–119, 2016.

[30] L. Liu, L. Tang, W. Dong, S. Yao, and W. Zhou, "An overview of topic modeling and its current applications in bioinformatics," *SpringerPlus*, vol. 5, no. 1, article 1608, 2016.

[31] H. Wang, M. Huang, and X. Zhu, "Extract interaction detection methods from the biological literature," *BMC Bioinformatics*, vol. 10, no. 1, article S55, 2009.

[32] C. W. Arnold, S. M. El-Saden, A. A. Bui, and R. Taira, "Clinical case-based retrieval using latent topic analysis," in *AMIA annual symposium proceedings*, vol. 2010, p. 26, American Medical Informatics Association, 2010.

[33] M. Song and S. Y. Kim, "Detecting the knowledge structure of bioinformatics by mining full-text collections," *Scientometrics*, vol. 96, no. 1, pp. 183–201, 2013.

[34] E. Sarioglu, K. Yadav, and H. A. Choi, "Topic Modeling Based Classification of Clinical Reports," in *ACL (Student Research Workshop)*, pp. 67–73, 2013.

[35] H. Bisgin, Z. Liu, H. Fang, X. Xu, and W. Tong, "Mining FDA drug labels using an unsupervised learning technique-topic modeling," *BMC Bioinformatics*, vol. 12, no. 10, article no. S11, 2011.

[36] V. Wang, L. Xi, A. Enayetallah, E. Fauman, and D. Ziemek, "GeneTopics - interpretation of gene sets via literature-driven topic models," *BMC Systems Biology*, vol. 7, no. 5, article no. S10, 2013.

[37] H. Bisgin, M. Chen, Y. Wang et al., "A systems approach for analysis of high content screening assay data with topic modeling," *BMC Bioinformatics*, vol. 14, no. 14, article no. S11, 2013.

[38] H. Bisgin, Z. Liu, R. Kelly, H. Fang, X. Xu, and W. Tong, "Investigating drug repositioning opportunities in FDA drug labels through topic modeling," *BMC Bioinformatics*, vol. 13, no. 15, article S6, 2012.

[39] S.-H. Wang, Y. Ding, W. Zhao et al., "Text mining for identifying topics in the literatures about adolescent substance use and depression," *BMC Public Health*, vol. 16, no. 1, article no. 279, 2016.

[40] X. Wang, P. Zhu, T. Liu, and K. Xu, "BioTopic: A topic-driven biological literature mining system," *International Journal of Data Mining and Bioinformatics*, vol. 14, no. 4, pp. 373–386, 2016.

[41] R. Sullivan, A. B. E. E. D. Sarker, OK. A. R. E. N. Connor, A. M. A. N. D. A. Goodin, M. A. R. K. Karlsrud, and G. R. A. C. I. E. L. A. Gonzalez, "Finding potentially unsafe nutritional supplements from user reviews with topic modeling," in *Pacific Symposium on Biocomputing*, vol. 21, pp. 528–539, World Scientific, Kohala Coast, Hawaii, 2016.

[42] J. H. Chen, M. K. Goldstein, S. M. Asch, L. Mackey, and R. B. Altman, "redicting inpatient clinical order patterns with probabilistic topic models vs conventional order sets," *Journal of the American Medical Informatics Association*, ocw136, 2016.

[43] M. Aksela, "Comparison of classifier selection methods for improving committee performance," in *International Workshop on Multiple Classifier Systems*, pp. 84–93, Springer, Berlin, Germany, 2003.

[44] D. Ruta and B. Gabrys, "Application of the evolutionary algorithms for classifier selection in multiple classifier systems with majority voting," in *International Workshop on Multiple Classifier Systems*, pp. 399–408, Springer, Berlin, Germany.

[45] Z. H. Zhou and W. Tang, "Selective ensemble of decision trees," *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing*, pp. 589-589, 2003.

[46] S. Sheen and A. P. Sirisha, "Malware detection by pruning of parallel ensembles using harmony search," *Pattern Recognition Letters*, vol. 34, pp. 1679–1686, 2013.

[47] I. Mendialdua, A. Arruti, E. Jauregi, E. Lazkano, and B. Sierra, "Classifier Subset Selection to construct multi-classifiers by means of estimation of distribution algorithms," *Neurocomputing*, vol. 157, pp. 46–60, 2015.

[48] D. D. Margineantu and T. G. Dietterich, "Pruning adaptive boosting," in *Proceedings of the Fourteenth International Conference on Machine Learning*, pp. 211–218, San Francisco, Calf, USA, 1997.

[49] R. Caruana, A. Niculescu-Mizil, G. Crew, and A. Ksikes, "Ensemble selection from libraries of models," in *Proceedings of the 21st International Conference on Machine Learning (ICML '04)*, pp. 18–39, Banff, Canada, July 2004.

[50] Q. Dai, T. Zhang, and N. Liu, "A new reverse reduce-error ensemble pruning algorithm," *Applied Soft Computing*, vol. 28, pp. 237–249, 2015.

[51] S. B. Kotsiantis and P. E. Pintelas, "Selective averaging of regression models," *Annals of Mathematics, Computing & Teleinformatics*, vol. 1, no. 3, pp. 65–74, 2005.

[52] M. Galar, A. Fernández, E. Barrenechea, H. Bustince, and F. Herrera, "Ordering-based pruning for improving the performance of ensembles of classifiers in the framework of imbalanced datasets," *Information Sciences*, vol. 354, pp. 178–196, 2016.

[53] H. Zhang and L. Cao, "A spectral clustering based ensemble pruning approach," *Neurocomputing*, vol. 139, pp. 289–297, 2014.

[54] T. L. Griffiths, M. I. Jordan, J. B. Tenenbaum, and D. M. Blei, "Hierarchical topic models and the nested chinese restaurant process," in *Advances in neural information processing systems*, pp. 17–24, 2004.

[55] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Sharing clusters among related groups: Hierarchical Dirichlet processes," in *Advances in Neural Information Processing Systems*, pp. 1385–1392, 2005.

[56] S. Grant and J. R. Cordy, "Estimating the optimal number of latent concepts in source code analysis," in *Proceedings of the 10th IEEE International Working Conference on Source Code Analysis and Manipulation (SCAM '10)*, pp. 65–74, IEEE, September 2010.

[57] A. Panichella, B. Dit, R. Oliveto, M. Di Penta, D. Poshynanyk, and A. De Lucia, "How to effectively use topic models for software engineering tasks? An approach based on genetic algorithms," in *Proceedings of the 35th International Conference on Software Engineering (ICSE '13)*, pp. 522–531, IEEE Press, May 2013.

[58] W. Zhao, J. J. Chen, R. Perkins et al., "A heuristic approach to determine an appropriate number of topics in topic modeling," *BMC Bioinformatics*, vol. 16, no. 13, article no. S8, 2015.

[59] A. Karami, A. Gangopadhyay, B. Zhou, and H. Kharrazi, "Fuzzy Approach Topic Discovery in Health and Medical Corpora," *International Journal of Fuzzy Systems*, pp. 1–12, 2017.

[60] R. Mousavi and M. Eftekhari, "A new ensemble learning methodology based on hybridization of classifier ensemble selection approaches," *Applied Soft Computing*, vol. 37, pp. 652–666, 2015.

[61] M. Jordan, *Learning in graphical models*, MIT Press, Cambridge, Mass, USA, 1999.

[62] L. Rokach, "Ensemble-based classifiers," *Artificial Intelligence Review*, vol. 33, no. 1-2, pp. 1–39, 2010.

[63] Z.-H. Zhou, *Ensemble Methods: Foundations and Algorithms*, Chapman and Hall, New York, NY, USA, 2012.

[64] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 4, no. 2, pp. 123–140, 1996.

[65] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 832–844, 1998.

[66] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, no. 2, pp. 241–259, 1992.

[67] Q. Sun and B. Pfahringer, "Bagging ensemble selection," in *Proceedings of the 24th Australasian Joint Conference on Artificial Intelligence*, pp. 251–260, Australia, 2011.

[68] S. Cheng, B. Liu, T. O. Ting, Q. Qin, Y. Shi, and K. Huang, "Survey on data science with population-based algorithms," *Big Data Analytics*, vol. 1, no. 1, article 3, 2016.

[69] J. H. Holland, *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*, MIT press, 1992.

[70] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of the IEEE International Conference on Neural Networks*, pp. 1942–1948, Perth, Australia, December 1995.

[71] X.-S. Yang, "A new metaheuristic bat-inspired algorithm," in *Nature Inspired Cooperative Strategies for Optimization (NICSO 2010)*, pp. 65–74, 2010.

[72] X.-S. Yang and S. Deb, "Engineering optimisation by Cuckoo search," *International Journal of Mathematical Modelling and Numerical Optimisation*, vol. 1, no. 4, pp. 330–343, 2010.

[73] X. S. Yang, *Nature-inspired metaheuristic algorithms*, Luniver press, 2010.

[74] E. Rendón, I. M. Abundez, C. Gutierrez et al., "A comparison of internal and external cluster validation indexes," in *Proceedings of the 2011 American Conference*, vol. 29, San Francisco, Calf, USA, 2011.

[75] D. J. Poirier, *Intermediate statistics and econometrics: a comparative approach*, MIT Press, 1995.

[76] L. I. Kuncheva and C. J. Whitaker, "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy," *Machine Learning*, vol. 51, no. 2, pp. 181–207, 2003.

[77] R. G. Rossi, R. M. Marcacini, and S. O. Rezende, *Benchmarking text collections for classification and clustering tasks*, Institute of Mathematics and Computer Sciences, University of Sao Paulo, 2013.

[78] H. Narasimhan, W. Pan, P. Kar, P. Protopapas, and H. G. Ramaswamy, "Optimizing the Multiclass F-Measure via Biconcave Programming," in *Proceedings of the IEEE 16th International Conference on Data Mining (ICDM)*, pp. 1101–1106, IEEE, 2016.

[79] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[80] X. Min, L. Liu, Y. He et al., *Benchmarking swarm intelligence clustering algorithms with case study of medical data*, 2016.

[81] P. Das, D. K. Das, and S. Dey, "A New Class Topper Optimization Algorithm with an Application to Data Clustering," *IEEE Transactions on Emerging Topics in Computing*, 2018.

Hindawi
Computational and Mathematical Methods in Medicine
Volume 2018, Article ID 7207151, 8 pages
https://doi.org/10.1155/2018/7207151

Research Article

Analysis and Study of Diabetes Follow-Up Data Using a Data-Mining-Based Approach in New Urban Area of Urumqi, Xinjiang, China, 2016-2017

Yukai Li,[1] Huling Li,[1] and Hua Yao [2]

[1]*College of Public Health, Xinjiang Medical University, Urumqi 830011, China*
[2]*Center of Health Management, The First Affiliated Hospital of Xinjiang Medical University, Urumqi, Xinjiang 830054, China*

Correspondence should be addressed to Hua Yao; yaohua01@sina.com

Received 12 February 2018; Revised 29 April 2018; Accepted 17 May 2018; Published 10 July 2018

Academic Editor: Federico Divina

The focus of this study is the use of machine learning methods that combine feature selection and imbalanced process (SMOTE algorithm) to classify and predict diabetes follow-up control satisfaction data. After the feature selection and unbalanced process, diabetes follow-up data of the New Urban Area of Urumqi, Xinjiang, was used as input variables of support vector machine (SVM), decision tree, and integrated learning model (Adaboost and Bagging) for modeling and prediction. The experimental results show that Adaboost algorithm produces better classification results. For the test set, the G-mean was 94.65%, the area under the ROC curve (AUC) was 0.9817, and the important variables in the classification process, fasting blood glucose, age, and BMI were given. The performance of the decision tree model in the test set is relatively lower than that of the support vector machine and the ensemble learning model. The prediction results of these classification models are sufficient. Compared with a single classifier, ensemble learning algorithms show different degrees of increase in classification accuracy. The Adaboost algorithm can be used for the prediction of diabetes follow-up and control satisfaction data.

1. Introduction

Currently, China has the highest number of chronic disease patients in the world, of which those suffering from diabetes and its associated complications are among the most critical. Diabetes is a chronic disease characterized by a long treatment cycle, numerous complications (e.g., kidney and eye diseases), and recurrent illness. With advances in the informatization of medicine, medical industries with large amounts of complicated patient data are keen to extract information from this data to assist the development of these industries. Simultaneously, they also seek to be capable of alleviating the challenges faced by medical personnel, through the forthcoming development of smart medicine. The use of machine learning and other artificial intelligence methods for the analysis of medical data in order to assist diagnosis and treatment is one of the manifestations of smart medicine with the most practical significance.

With the improvement of the living standards of our people and the westernization of our diet, the incidence, mortality, and morbidity of diabetes have significantly increased and have a serious impact on our health. In 2006, Shang [1] made use of the survey data of Xinjiang chronic disease integrated prevention and control demonstration site in the New Urban District of Urumqi in 2004 and surveyed 2031 people over the age of 18 in three communities in the district. The results showed the relationship between diabetes and age and gender: the prevalence of male and female rose with age, because the decrease of glucose tolerance with age and the improvement of living standard are the reasons for the increased incidence. Overweight and obesity are one of the risk factors of diabetes mellitus. The survey found that the prevalence of diabetes in people with BMI>24 was 10. 58%, the prevalence of diabetes in people with BMI≦24 was 4.31%, two groups prevalence by chi-square test was $P < 0.01$, and there was a significant difference between the two groups, indicating that overweight and obese individuals are more susceptible to diabetes. In 2009, Su [2] analyzed the related factors of diabetes in the New Urban District of Urumqi in Xinjiang. The results showed that age, gender,

height, weight, and BMI associated with diabetes were not statistically significant. However, the waist circumference, systolic blood pressure, and triglyceride are factors that are positively correlated with diabetes. In 2017, Mohemaiti [3] used questionnaire to survey the prevalence of 200 elderly patients type 2 diabetes with coronary heart disease from January to December in 2016 in Hangzhou Road community of the New Urban Area of Urumqi; the results showed that smoking, BMI $\geq 24\,kg/m^2$, complications associated with diabetes, hypertension, and dyslipidemia are risk factors for coronary heart disease in elderly patients with diabetes mellitus. It is the key according to the relevant risk factors and the timely development of interventions to reduce the prevalence of coronary heart disease in elderly patients with diabetes mellitus.

Data mining is a significant tool in medical databases, which enhances the sensitivity and/or specificity of disease detection and diagnosis by opening a window of relatively better resources [4]. Applying machine learning and data mining methods in diabetes research is a pivotal way to utilizing plentiful available diabetes-related data for extracting knowledge. The severe social impact of the specific disease makes DM one of the main priorities in medical science research, which inevitably produces large amounts of data. Therefore, there is no doubt that machine learning and data mining approaches in DM are of great concern on diagnosis, management, and other related clinical administration aspects [5]. In order to achieve the best classification accuracy, abundant algorithms and diverse approaches have been applied, such as traditional machine learning algorithms, ensemble learning approaches, and association rule learning. Most noted among the aforementioned ones are the following: Calisir and Dogantekin proposed LDA-MWSVM, a system for diabetes diagnosis [6]. The system performs feature extraction and reduction using the Linear Discriminant Analysis (LDA) method, followed by classification using the Morlet Wavelet Support Vector Machine (MWSVM) classifier. Gangji and Abadeh [7] presented an Ant Colony-based classification system to extract a set of fuzzy rules, named FCSANTMINER, for diabetes diagnosis. In [8], authors regard glucose prediction as a multivariate regression problem utilizing Support Vector Regression (SVR). Agarwal [9] utilized semi-automatically marked training sets to create phenotype models via machine learning methods. Ensemble approaches, which utilize multiple learning algorithms, have been confirmed to be an effective way of enhancing classification accuracy.

This study follows the support vector machine (SVM), Adaboost, Bagging data mining ensemble techniques, and decision tree as our research model. More specifically, the dataset used for decision-making in this study is obtained from the diabetes follow-up data of the New Urban Area of Urumqi, Xinjiang. The purpose of this study is to evaluate the performance of aforementioned techniques of data mining and adopt machine learning methods that combine feature selection and class unbalanced processing to evaluate the health management control satisfaction of diabetic patients. We used health management measure indicators of diabetes patients as the input variables of our models to accurately classify two levels of control satisfaction in follow-up data, namely, (i) satisfied with the control and (ii) unsatisfied with the control. Finally, a classification model with further higher classification accuracy was constructed.

2. Materials and Methods

2.1. Dataset. The dataset used in this study is gathered from the diabetic patient health management follow-up data of the New Urban Area of Urumqi, Xinjiang. The dataset contains 3406 records for a period ranging from December 1, 2016, to February 28, 2017. Each record includes 25 characteristic variables, which are likely to affect the degree of satisfaction with diabetes control. An abstract detail of those relevant factors selected in this study is provided in Table 1 that includes age, sex, race, body mass index (BMI), diabetes complications, systolic blood pressure, diastolic blood pressure, and fasting blood glucose of the patients. The chi-square test was used to compare and analyze the satisfaction of different classification variables and the respondents. By using chi-square test to select a small number of the most relevant features (or by eliminating many irrelevant features), one is able to reduce the risk of overfitting the training data and often produce a better overall model. The difference was statistically significant at $P<0.05$. Categorical variables are statistically significant by chi-square test and continuous variables, which are used as input variables for machine learning.

In our research, the dataset encounters the class imbalance problem. Out of 3406 patients, 2832 patients were satisfied with control of diabetes, which constitutes about 83.21% of the total patients and 574 patients are unsatisfied. The imbalanced ratio equals 5:1 between majority and minority. In other words, a dataset is class-imbalanced if one class includes significantly more sample numbers than the other. In order to resolve the problem, we can pick the random undersampling (RUS), random oversampling (ROS), and SMOTE, which are among the most used resampling methods to counterpoise imbalanced datasets. Here, we only choose SMOTE algorithms, which is used to create one more dataset, where the minority samples were oversampled by 400% and the majority class was undersampled at 123% to approximately make the ratio 1:1. The descriptions of the datasets are given in Table 2. Eventually, the balanced dataset was used to construct the model.

2.2. Algorithms. We selected 4 algorithms to test decision tree, support vector machine (SVM), Bagging, and Adaboost which are common algorithms in machine learning. Decision tree [10] is a category of tree classifier. Generally, decision tree uses information entropy, information gain, or Gini coefficients to assess which characteristic to use as the classification characteristic corresponding to a non-leaf-node [11]. Ordinarily, decision trees can intuitively display the classification process, clearly showing rules that can be understood by humans. SVMs are supervised learning models associated with data analysis and model recognition and are widely used in classification and regression analysis, which use a hypothesis space of polynomial linear functions over a high

Table 1: Analysis of control satisfaction of diabetes patients in New Urban Area of Urumqi (n=3406).

Characteristic	Satisfied (N1=574)	Unsatisfied (N2=2832)	χ^2	*P* values
Age, Median (IQR), Years	57(49-65)	54(46-62)	-	-
Sex				
male	276	1400	0.35	0.555
female	298	1432		
Ethnicity				
Han nationality	479	2544		
Hui	57	183	28.05	*<0.0001*
others	3	28		
Uighur	35	77		
Degree of education				
junior high school	193	866		
College specialties and above	55	392		
High School / Technical School	96	559	12.62	*0.013*
Illiteracy and semi-literacy	56	245		
primary school	174	770		
Marital status				
Divorced / widowed	59	362		
unmarried	3	13	2.79	0.248
married	512	2457		
Diagnosis methods				
clinical	228	1673		
outpatient clinic	333	1099	73.96	*<0.0001*
others	13	60		
Diabetes complications				
Coronary heart disease				
no	525	2462	9.07	*0.003*
yes	49	370		
Hypertension				
no	311	1317	11.27	*0.001*
yes	263	1515		
High cholesterol				
no	483	2579	25.17	*<0.0001*
yes	91	253		
Smoking				
no	270	1546	10.94	*0.001*
yes	304	1286		
Drinking				
no	278	1622	15.13	*<0.0001*
yes	296	1210		
Diet control				
no	187	666	20.88	*<0.0001*
yes	387	2166		
physical activities				
no	158	621	8.48	*0.004*
yes	416	2211		

Table 1: Continued.

Characteristic	Satisfied (N1=574)	Unsatisfied (N2=2832)	χ^2	*P* values
Hypoglycemic agents				
no	175	802	1.10	0.295
yes	399	2030		
Insulin				
no	337	1722	0.88	0.349
yes	237	1110		
Quit smoking				
no	356	1903	5.72	*0.017*
yes	218	929		
Limit wine				
no	333	1863	12.58	*<0.0001*
yes	241	969		
Follow-up method				
phone	50	218	9.75	*0.008*
home	26	234		
clinic	498	2380		
Psychological adjustment				
poor	8	13	78.86	*<0.0001*
good	327	2123		
fair	239	696		
Follow medical practice				
poor	98	103	191.40	*<0.0001*
good	254	1863		
fair	222	866		
Compliance medication				
no medication	80	421	41.89	*<0.0001*
regular	455	2356		
intermittent	39	55		
Systolic blood pressure, Median (IQR), mm Hg	130 (120-140)	130 (120-140)	-	-
Diastolic blood pressure, Median (IQR), mm Hg	78 (70-80)	80 (70-84)	-	-
BMI, Median (IQR), kg/m^2	25.36 (23.53-27.53)	26.27 (24.14-28.43)	-	-
Fasting blood glucose level, Median (IQR), mmol/L	6.4 (6.0-6.8)	8.7 (7.5-11.03)	-	-

Table 2: Dataset description.

Dataset	Samples distribution	Ratio	Description
Original data	2832/574	5:1	Original data with full instances
SMOTE-data	2824/2870	1:1	Dataset is balanced utilizing SMOTE oversampling

dimensional feature space. While SVMs are a "black box" algorithm, they typically outperform other ML algorithms for classification tasks [12, 13]. In 1996, Breiman proposed the popular bootstrap aggregation (Bagging) method [14]. It primarily involves bootstrap sampling techniques in which samples are selected repeatedly with a certain probability and with replacement, which generates numerous different sample subsets. Next, these different sample subsets are used individually to perform training on base classifiers and obtain an integrated classifier with certain diversity. The diversity strategy of Bagging is straightforward and effective, and numerous derivative methods based on this strategy yield adequate classification results [15]. Boosting, also known as reinforcement learning, is a critical ensemble

Table 3: Confusion matrix.

		Predicted classification	
		1	0
Actual classification	1	TP	FP
	0	FN	TN

learning technique that can reinforce a weak classifier, whose prediction accuracy is marginally higher than that of a random guess, into a strong classifier with high prediction accuracy. Adaboost is the most successful representative of this algorithm and has been rated as one of the ten most effective algorithms for data mining [16]. This algorithm is an iterative method that was proposed by Schapire and Freund in 1995 [17–19].

Because each of these algorithms has their own characteristics and advantages, each method will produce different results to classify the degree of satisfaction of diabetes follow-up and control, and for more comprehensive evaluation of predictors in the imbalanced context, G-mean [20] and AUC [21] are frequently used to measure how well the predictor can balance the performance between two classes, so we choose G-mean and area under the ROC curve (AUC) as an index to evaluate the performance of the classification models. By using confusion matrix (see Table 3), we can calculate the accuracy, sensitivity, and specificity.

G-mean is the geometric mean of the sensitivity and specificity; that is,

$$\text{G-mean} = \sqrt{\text{Sensitivity} \times \text{Specificity}} \quad (1)$$

The ROC curve describes the relationship between TP/(TP + FN) and FP/(FP + TN) of the classifier. Since the ROC curve cannot quantitatively evaluate the classifiers, AUC is usually adopted as the evaluation index. AUC (area under ROC curve) value refers to the area under the ROC curve. An ideal classification model has an AUC value of 1, with a value between 0.5 and 1.0, and the larger AUC represents that the classification model has better performance.

The experimentation is performed using open source R software version 3.4.1 (https://www.r-project.org/). The main packages included the following:

(1) The adabag (https://cran.r-project.org/web/packages/adabag/) software package focuses on the Bagging and Adaboost algorithms.

(2) The kernlab (https://cran.r-project.org/web/packages/kernlab/) package was used for the support vector machine algorithm.

(3) The rpart (https://cran.r-project.org/web/packages/rpart/) was used for decision tree classification.

3. Results

Our research dataset is divided into two parts; two-thirds of the data is used as a training set, and one-third of the dataset is defined as a testing set to evaluate the performance of several classifiers. All classifiers were fitted to the same training and testing data. The specific process is shown in Figure 1.

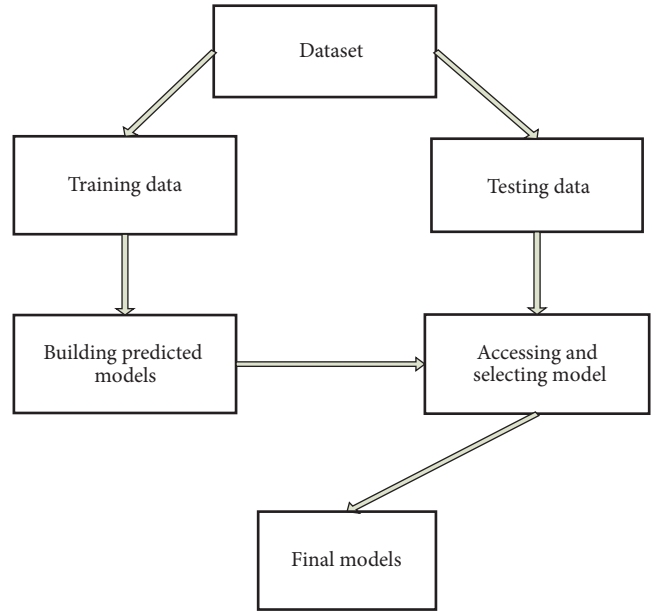


Figure 1: General flowchart of modeling.

As can be seen from Table 4, in this study, the performance of the four final predictive models was evaluated using G-mean, AUC. For the testing dataset, the final comparative analysis results demonstrated that the Adaboost algorithm showed the best with accuracy of 94.84%, and the sensitivity and specificity were 95.76% and 93.56%, respectively. The SVM algorithm came out to be the second best with a classification accuracy of 92.62%, and the sensitivity and specificity gave 94.08% and 91.28%, respectively, followed by the Bagging model (91.15%) and decision tree (91.15%), which exhibited identical results, with the sensitivity and specificity being equal to 90.50% and 91.81%, respectively. In the results, the area under the receiver operating characteristic (ROC) curve (AUC) values of the SVM, Bagging, and decision tree algorithms were 0.9688, 0.9164, and 0.9115, respectively. The area under ROC for Adaboost ensemble method is 98.17% and G-mean of 0.9465, showing a high reliability of discriminative capability among all the methods. Overall, the ML method presented in this paper has obtained the well classification performance of health management control satisfaction of patients with diabetes. Decision tree also yielded better performance. The ROC curves for the four classifiers are shown in Figure 2.

4. Discussion

Health management of diabetic patients is an important part of the national basic public health service project. Diabetics are one of the six key groups defined by the national basic public health service project, and satisfaction is one of the important indicators of the effectiveness of the test project [22]. Patients are satisfied with the services provided; they will take the initiative to participate in the project to form a virtuous circle, further enhance the effectiveness of project health

Table 4: Comparison of prediction performance of the four models.

Algorithms	Accuracy	Sensitivity	Specificity	G-mean	AUC
Decision Trees	0.9115	0.9050	0.9181	0.9115	0.9115
SVM	0.9262	0.9408	0.9128	0.9267	0.9688
Adaboost	0.9484	0.9576	0.9356	0.9465	0.9817
Bagging	0.9115	0.9050	0.9181	0.9115	0.9164

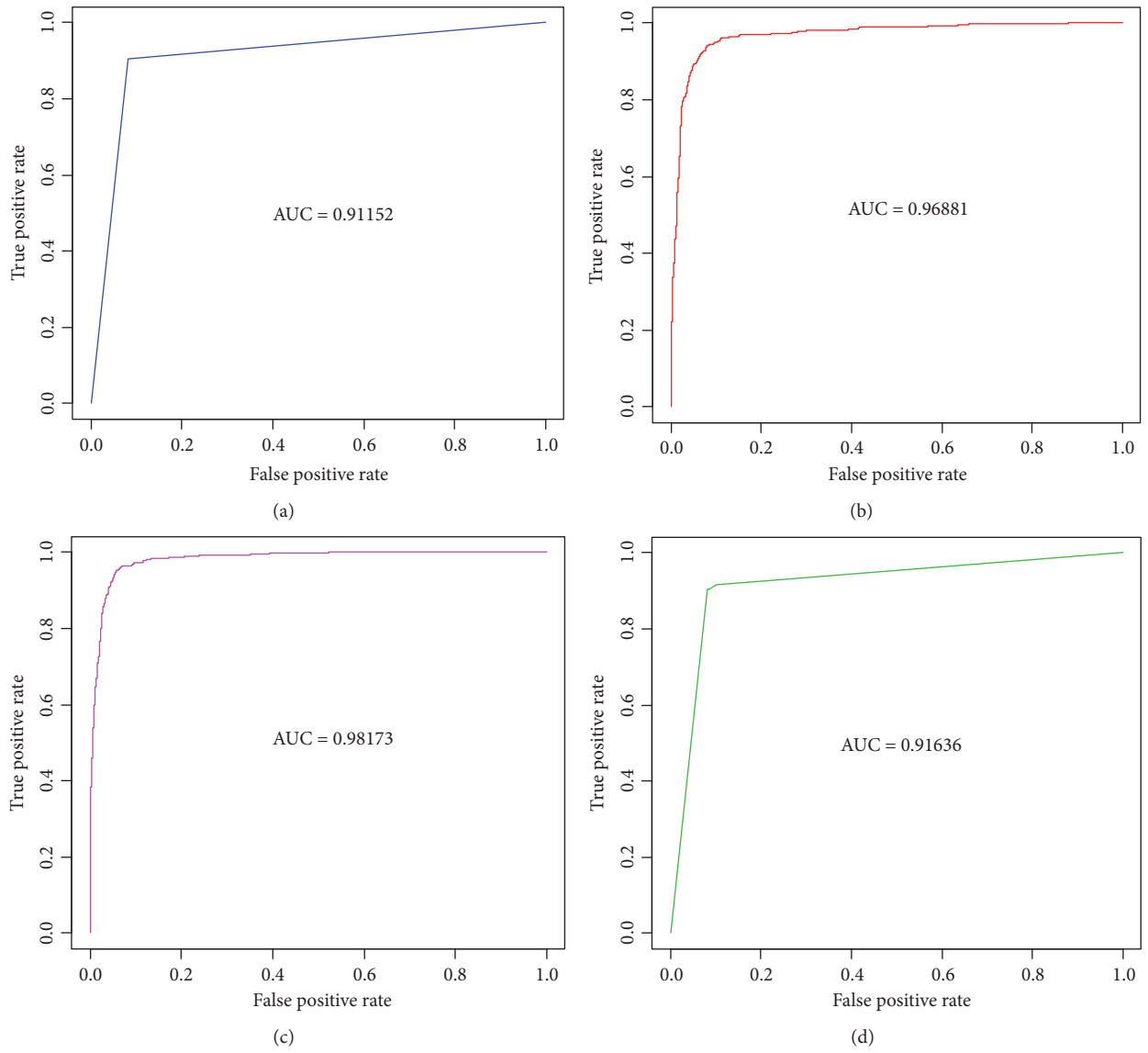


Figure 2: ROC curves for (a) decision tree model, (b) SVM model, (c) Adaboost model, and (d) Bagging model.

management, and then promote the smooth implementation of the project. At the same time, patient satisfaction with health services directly affects the development of health services. Therefore, we must attach great importance to the satisfaction of patients and improve patient satisfaction by continuously improving service capabilities and service quality [23]. Machine learning methods provide a new way to diabetes analytics which is suitable for contemporary Big Data demands. Those approaches could get over many constraints intrinsic in many traditional statistical modeling approaches [24]. Therefore, when focusing on a certain disease, several appropriate classification algorithms should be selected based on the characteristics of the dataset. By comparing the classification accuracy of these classification algorithms on the dataset, the most effective classification algorithm is used as the diagnostic model. In general, the performance of machine learning algorithms is evaluated using predictive accuracy. However, this is not appropriate

when the data is imbalanced and/or the costs of different errors vary markedly.

The dataset used in this study is obtained from the diabetic patient health management follow-up data of the New Urban Area of Urumqi, Xinjiang. This study systematically involves four representative data mining techniques for predictive data mining task. That includes decision tree, SVM, ensemble learning method Bagging, and Adaboost. These algorithms are combined for creating knowledge to render it useful for decision-making. Each algorithm will produce different results to classify the degree of satisfaction with diabetes control. Firstly, chi-square test was used to select the features of the diabetes dataset. Secondly, because the dataset has unbalanced problem, we chose a method to deal with unbalanced data, that is, the SMOTE method. Finally, the dataset after feature selecting and unbalanced processing was classified by four classification algorithms. The experimental results proved that, for the testing dataset, Adaboost algorithm performed best in four models with a AUC equal to 0.9817 and an G-mean equal to 0.9465. An important feature of the Adaboost algorithm is the calculation of the importance of each variable (feature). We can output the importance score of each input variable in the classification process. Variables with high importance are closely related to the predictions results. For instance, Huang [25] mentioned that adequately controlled blood glucose was defined as fasting blood glucose values <7.0 mmol/L. The effect of post-management blood glucose control has a direct impact on patient satisfaction, with a statistically significant difference ($X2=24.128$, $P<0.05$). Moreover, Baccaro [26] also indicated that a significant statistic correlation was observed between the score of the questionnaires and good diabetes control showed by the levels of HbAc1 and fasting blood glucose, among other parameters, which is consistent with the first important variable (fasting blood glucose) reported by the Adaboost algorithm proposed by us. Our results also showed that the age and BMI were also important variables. One study has pointed out [27] higher age, better physical health, less diabetes-related distress, and higher diabetes treatment satisfaction. Another example, a previous study [28] aims to assess the psychological well-being and treatment satisfaction in patients with type 2 diabetes mellitus in a general hospital in Korea. Their result revealed that treatment satisfaction was significantly associated with age, satisfaction with waiting and treatment times, compliance with recommended diet and exercise, and duration of diabetes. For BMI, there is a certain relationship between the satisfaction rate of blood glucose control and overweight or obesity, which explains the importance of BMI in the classification of control satisfaction [29]. Besides, to determine which patient characteristics and laboratory values were independently associated with treatment satisfaction, Boels [30] used a linear mixed model for analysis, whose conclusion was that a number of factors including diabetes education, perceived and actual hyperglycaemia, and macrovascular complications are associated with treatment satisfaction. The Bagging and Adaboost methods [31] combine a large number of decision trees and can significantly increase their prediction efficiency. Ensemble learning algorithm has better performance than simple classification algorithm (decision tree).

The limitations of research should also be recognized. In this paper, only one method of dealing with unbalanced data is used. Of course, all kinds of methods have been developed to deal with unbalanced data, such as random oversampling, cluster-based oversampling, and algorithmic ensemble techniques. This paper does not compare with the performance of the original dataset in the algorithm. In the future work, we can consider, from a variety of perspectives, adopting diverse imbalanced processing methods and a machine learning method to compare the effects of different types of unbalanced processing techniques.

In addition, it should be referred that despite the claims that these machine learning classification algorithms can generate sufficient and effective decision-making, very few have really permeated the clinical practice [32]. Understandably, clinicians are not only interested in the high accuracy of a predictive model, but also in the degree with which the model could explain the pathogenesis of the disease [24]. Although it has powerful learning capabilities, without being supported by the appropriate approaches for determining how they work, the results of machine learning algorithms prediction may encounter a limited applicability in the clinical practices. We used machine learning approaches for diabetes analytics in real-life clinical settings, which is a severe challenge.

5. Conclusions

In this study, we used the diabetic patient health management follow-up data. We have combined feature selection and imbalanced processing techniques, and few researchers have utilized the health management control satisfaction of patients with diabetes for classification predictions. In this work, we offered proof that Adaboost algorithm can be successfully used for health management control satisfaction of patients with diabetes.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Acknowledgments

This work was supported by Key Research and Development Project of Xinjiang (no. 2016B03048).

References

[1] X. J. Shang, L. X. Liu, and Y. M. Guan, "Analysis of the results of diabetes investigation in the New Urban Area of Urumqi,

Xinjiang," *Bulletin of Disease Control and Prevention*, vol. 21, no. 3, pp. 69-69, pp. 69-69.2006.

[2] L. Q. Su, F. P. Wang, and X. Y. Wang, "Analysis of related factors of diabetes in the New Urban Area of Urumqi, Xinjiang," *Xinjiang Medicine*, vol. 39, no. 4, pp. 12-13, 2009.

[3] P. Mohemaiti, Y. Keyoumu, P. Mohemaiti et al., "Current situation and related risk factors of elderly type 2 diabetes mellitus with coronary heart disease in Hangzhou road community of the New Urban Area of Urumqi," *Chinese Journal of Gerontology*, vol. 37, no. 21, pp. 5422–5424, 2017.

[4] S. Perveen, M. Shahbaz, A. Guergachi et al., "Performance Analysis of Data Mining Classification Techniques to Predict Diabetes," *Procedia Computer Science*, vol. 82, pp. 115–121, 2016.

[5] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine Learning and Data Mining Methods in Diabetes Research," *Computational and Structural Biotechnology Journal*, vol. 15, pp. 104–116, 2017.

[6] D. Çalişir and E. Doğantekin, "An automatic diabetes diagnosis system based on LDA-wavelet support vector machine classifier," *Expert Systems with Applications*, vol. 38, no. 7, pp. 8311–8315, 2011.

[7] M. F. Ganji and M. S. Abadeh, "A fuzzy classification system based on ant colony optimization for diabetes disease diagnosis," *Expert Systems with Applications*, vol. 38, no. 12, pp. 14650–14659, 2011.

[8] E. I. Georga, V. C. Protopappas, D. Ardigò et al., "Multivariate Prediction of Subcutaneous Glucose Concentration in Type 1 Diabetes Patients Based on Support Vector Regression," *IEEE Journal of Biomedical Health Informatics*, vol. 17, no. 1, pp. 71–81, 2013.

[9] V. Agarwal, T. Podchiyska, J. M. Banda et al., "Learning statistical models of phenotypes using noisy labeled training data," *Journal of the American Medical Informatics Association* , vol. 23, no. 6, Article ID ocw028, pp. 1166–1173, 2016.

[10] L. Rokach and Z. O. Maimon, *Data mining with decision trees: theory and applications*, vol. 13, World Scientific Publishing Co., Toh Tuck, Singapore, 2008.

[11] L. Rokach and O. Maimon, "Top-down induction of decision trees classifiers - a survey," *IEEE Transactions on Systems Man & Cybernetics Part C*, vol. 35, no. 4, pp. 476–487, 2005.

[12] D. Cossock and T. Zhang, "Statistical analysis of Bayes optimal subset ranking," *Institute of Electrical and Electronics Engineers Transactions on Information Theory*, vol. 54, no. 11, pp. 5140–5154, 2008.

[13] S. Mani, Y. Chen, T. Elasy et al., "Type 2 diabetes risk forecasting from EMR data using machine learning," *AMIA Annual Symposium Proceedings*, vol. 2012, pp. 606–615, 2012.

[14] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[15] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[16] Z. H. Zhou, Y. Yang, X. D. Wu, and V. Kumar, *The Top Ten Algorithms in Data Mining*, CRC Press, New York, NY, USA, 2009.

[17] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, part 2, pp. 119–139, 1997.

[18] Y. Freund and E. R. Schapire, "Experiments with a new boosting algorithm," in *13th International Conference on International Conference on Machine Learning*, vol. 13, pp. 148–156, Morgan Kaufmann Publishers Inc., Bari, Italy, 1996.

[19] R. E. Schapire and Y. Singer, "Improved boosting algorithms using confidence-rated predictions," *Machine Learning*, vol. 37, no. 3, pp. 297–336, 1999.

[20] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: One-sided selection," in *Proceedings of the 14th International Conference on Machine Learning*, pp. 179–186, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997.

[21] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

[22] F. L. Meng and S. H. JIN, "Investigation and Analysis of Satisfaction of Hypertension Patients in Community Health Service in Hangzhou," *Health Research*, vol. 32, no. 2, pp. 132–134, 2012.

[23] Y. F. Zhao, X. Yao, F. Deng et al., "Survey and Analysis of the Satisfaction of the Residents of the Community Health Service Institutions in Karamay City," *Chinese Journal of Social Medicine*, no. 4, pp. 306–308, 2015.

[24] Y. Sebastian, "Advances in Diabetes Analytics from Clinical and Machine Learning Perspectives," *International Journal of Design, Analysis and Tools for Integrated Circuits and Systems*, vol. 6, no. 1, pp. 32–37, 2017.

[25] L. Huang, Y. Liu, X. J. Guan et al., "Satisfaction of Diabetic Patients under Community Health Management in Wuhou District of Chengdu, 2014-2016," *Journal of Preventive Medicine Information*, vol. 33, no. 8, pp. 728–731, 2017.

[26] F. Baccaro, P. P. Novelli, J. Arduin et al., "Diabetes Treatment Satisfaction Questionnaire (DTSQ) of in non-ambulatory type 2 diabetic patients," *Boletín De La Asociación Médica De Puerto Rico*, vol. 108, no. 1, pp. 55–60, 2016.

[27] P. R. Wermeling, J. Janssen, K. J. Gorter, J. W. J. Beulens, and G. E. H. M. Rutten, "Satisfaction of well-controlled type 2 diabetes patients with three-monthly and six-monthly monitoring," *BMC Family Practice*, vol. 14, article no. 107, 2013.

[28] H. Park, S. N. Lee, M. Y. Baek et al., "The Well-Being and Treatment Satisfaction of Diabetic Patients in an Outpatient Setting at a General Hospital in Korea," *The Journal of Korean Diabetes*, vol. 17, no. 2, p. 123, 2016.

[29] H. J. Xie, "Investigation and Analysis of Blood Glucose Control," *The Chinese community physicians (medicine)*, vol. 15, no. 8, pp. 354-355, 2013.

[30] A. M. Boels, R. C. Vos, T. G. Hermans et al., "What determines treatment satisfaction of patients with type 2 diabetes on insulin therapy? An observational study in eight European countries," *BMJ Open*, vol. 7, no. 7, p. e016180, 2017.

[31] S. Mani, Y. Chen, and T. Elasy, "Type 2 Diabetes Risk Forecasting from EMR Data using Machine Learning," *AMIA Annual Symposium Proceedings*, pp. 606–615, 2012.

[32] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Computational and Structural Biotechnology Journal*, vol. 13, pp. 8–17, 2015.

Hindawi
Computational and Mathematical Methods in Medicine
Volume 2018, Article ID 3018356, 11 pages
https://doi.org/10.1155/2018/3018356

Research Article

Exploration of Neural Activity under Cognitive Reappraisal Using Simultaneous EEG-fMRI Data and Kernel Canonical Correlation Analysis

Biao Yang,[1,2] Jinmeng Cao,[1,2] Tiantong Zhou,[1,2] Li Dong,[3] Ling Zou,[1,2] and Jianbo Xiang[4]

[1]*School of Information Science and Engineering, Changzhou University, Changzhou, Jiangsu 213164, China*
[2]*Changzhou Key Laboratory of Biomedical Information Technology, Changzhou, Jiangsu 213164, China*
[3]*School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu, Sichuan 610054, China*
[4]*Changzhou No. 2 People's Hospital Affiliated with Nanjing Medical University, Changzhou, Jiangsu 213164, China*

Correspondence should be addressed to Ling Zou; zouling@cczu.edu.cn and Jianbo Xiang; hx_bob@163.com

Received 5 March 2018; Accepted 21 May 2018; Published 2 July 2018

Academic Editor: Miguel García-Torres

Background. Neural activity under cognitive reappraisal can be more accurately investigated using simultaneous EEG- (electroencephalography) fMRI (functional magnetic resonance imaging) than using EEG or fMRI only. Complementary spatiotemporal information can be found from simultaneous EEG-fMRI data to study brain function. *Method*. An effective EEG-fMRI fusion framework is proposed in this work. EEG-fMRI data is simultaneously sampled on fifteen visually stimulated healthy adult participants. Net-station toolbox and empirical mode decomposition are employed for EEG denoising. Sparse spectral clustering is used to construct fMRI masks that are used to constrain fMRI activated regions. A kernel-based canonical correlation analysis is utilized to fuse nonlinear EEG-fMRI data. *Results*. The experimental results show a distinct late positive potential (LPP, latency 200-700ms) from the correlated EEG components that are reconstructed from nonlinear EEG-fMRI data. Peak value of LPP under reappraisal state is smaller than that under negative state, however, larger than that under neutral state. For correlated fMRI components, obvious activation can be observed in cerebral regions, e.g., the amygdala, temporal lobe, cingulate gyrus, hippocampus, and frontal lobe. Meanwhile, in these regions, activated intensity under reappraisal state is obviously smaller than that under negative state and larger than that under neutral state. *Conclusions*. The proposed EEG-fMRI fusion approach provides an effective way to study the neural activities of cognitive reappraisal with high spatiotemporal resolution. It is also suitable for other neuroimaging technologies using simultaneous EEG-fMRI data.

1. Introduction

Emotional regulation is known as a unique ability of human beings to control experience and expression of their emotions. It has been the focus of many fields (e.g., cognitive neuroscience, clinical medicine, and sociology) due to its importance to human mental health [1, 2]. Two well-established emotional regulation strategies are widely applied to control emotional experiences, including expressive suppression and cognitive reappraisal. The former is a way of response modulation whereby individual voluntarily inhibits emotional expressive behavior [3]. However, according to the catharsis model, emotions are supposed to "pile up" if not expressed [4]. Hence, expressive suppression may enhance emotional experience which harms mental health. Cognitive reappraisal, on the other hand, is an approach to change the way people think about a potentially emotion eliciting condition to decrease the emotional influence [5]. For instance, one's representative reaction to a scene of a person shooting at another one may be decreased by imaging the scene as a film scene. On the contrary, the reaction may be enhanced by imaging the person is shot by his/her

close relative. By utilizing cognitive reappraisal, discomfort to events (e.g., sick, horror, and self-abasement) can be alleviated at an early stage. Despite recent studies show that cognitive reappraisal is correlated to facial frown muscle activities [6], heart rate, and skin conductance [3], studying the essence of cognitive reappraisal is still urgent.

Recently, several neuroimaging technologies (e.g., EEG (electroencephalography) and fMRI (functional magnetic resonance imaging)) are utilized to explore the essence of cognitive reappraisal. Submillisecond temporal resolution of EEG makes it suitable to explore the subtle temporal dynamics of neural activity, which is expressed by electric potential fluctuations spread to the scalp. Event Related Potential (ERP) is widely used to study the characteristics of EEG signals under different emotional states due to its high temporal resolution. An essential component of ERP, Late Positive Potential (LPP), is found to indicate the ability of cognitive reappraisal using emotional regulation. The facilitated processing of emotional stimuli is indicated by the LPP as a central-parietal slow positive deflection in the ERP. The amplitude of LPP turns out to be increased for emotionally eliciting compared with neutral stimuli, beginning with approximately 200ms after stimulus onset and continuing several seconds [7]. Meanwhile, it is susceptible to spontaneous emotional regulation. Hence, a decrease of LPP amplitude can be found when participants are asked to distract attention from the pictures which may arouse unpleasant emotion via cognitive reappraisal [8]. Moreover, LPP reduction can also be found from positive emotional regulation by cognitive reappraisal [9, 10]. However, emotional eliciting sources are hard to locate due to the poor spatial resolution of EEG. FMRI is another widely used technology to study the brain function. It can localize both superficial and deep sources of activity with mm-scale spatial resolution via detecting the variations of blood oxygenation level-dependent (BOLD). Cerebral regions which participate in emotional regulation can be found via fMRI due to its high spatial resolution. Recent fMRI researches show that voluntary reappraisal can influence modulated neural activities in the amygdala [11, 12]. It also indicates that the employment of cognitive reappraisal influences the neural activities in the dorsal parts of the anterior cingulate cortex, the ventromedial prefrontal cortex, and the dorsolateral prefrontal cortex [13]. However, the low resolution temporal variations of these regions are not suitable for studying the neural activity under cognitive reappraisal.

To resolve the abovementioned insufficient of monomodality neuroimaging technology, simultaneous EEG-fMRI fusion is utilized to study the neural activity under cognitive reappraisal due to its high spatiotemporal resolution. In general, there are mainly three approaches for simultaneous EEG-fMRI fusion, including fMRI aided EEG analysis, EEG aided fMRI analysis, and symmetric EEG-fMRI analysis. For fMRI aided EEG analysis, fMRI information with high spatial resolution is used to support the inverse issue of EEG source reconstruction. Kyathanahally *et al.* proposed a framework to invest decision-making in the brain using simultaneous EEG-fMRI data [14]. Thinh *et al.* developed a novel multimodal EEG-fMRI fusion approach by employing the most probable fMRI spatial subsets to guide EEG source localization in a time-variant fashion [15]. For EEG informed fMRI analysis, EEG features (e.g., ERP amplitude, the power spectrum, and epileptic) are used to forecast the BOLD changes in fMRI. Liu *et al.* proposed a general linear model (GLM) model for EEG-fMRI fusion. The fusion results indicate that the intraparietal sulcus and frontal executive areas are the primary sources of biasing influences on task-related visual cortex, whereas task-unrelated default mode network and sensorimotor cortex are suppressive during visual attention [16]. Ahmad *et al.* developed a framework to recognize different visual brain activity patterns using simultaneous EEG-fMRI data. A GLM model was utilized for EEG-fMRI fusion and the results were further classified into different patterns by multilayer perceptron [17]. For symmetric EEG-fMRI analysis, both data are jointly processed by a generative model or changed into a common feature/data space. Yu *et al.* developed a framework to construct multimodal brain graphs using EEG-fMRI data which were simultaneously sampled during eyes open and eyes closed resting states [18]. FMRI data were decomposed into independent components with associated time courses by group independent component analysis (ICA) and EEG time series were segmented into spectral power time courses by superposed average of five frequency bands (alpha, theta, beta, delta, and low gamma). However, ICA assumes that all sources are independent. This strong assumption restricts the power of ICA fusion approach in exploring the underlying sources. Canonical correlation analysis (CCA) was employed by Correa *et al.* to fuse simultaneous EEG-fMRI data with weak assumption [19]. Dong *et al.* also proposed a CCA based EEG-fMRI fusion approach to study familial cortical myoclonic tremor and epilepsy [20]. The proposed local multimodal serial analysis was specifically designed to handle the change of hemodynamic response functions (HRFs).

Despite the widely developed approaches to analyze EEG-fMRI data, there is still no method that focuses on two challenging issues of simultaneous EEG-fMRI fusion; one is to handle the mutual interference between EEG and fMRI, and the other is to handle the nonlinearity of EEG-fMRI data. Aiming to resolve these challenges, we propose an effective fusion framework based on CCA. Empirical mode decomposition (EMD) is used to increase SNR of EEG data that is polluted by MR scanning. FMRI masks are constructed and are used to eliminate unwanted fMRI components that are correlated with wanted EEG components. RBF kernel is embedded into the CCA framework to handle the nonlinearity of EEG-fMRI data. Participants are shown with visual stimuli paradigm. Based on previous researches that study EEG and fMRI, respectively [7], we expect (1) the correlated ERP components and fMRI activated regions related to cognitive reappraisal can be simultaneously be extracted from the EEG-fMRI data, and (2) LPPs under three emotional states can be observed from the correlated ERP components and amplitudes of different LPPs coincided with the existed studies, and (3) the correlated fMRI activated regions coincided with the previous found regions (e.g., amygdala, dorsomedial PFC, dorsolateral prefrontal cortex (PFC), anterior cingulate cortex, and orbitofrontal cortex). There are mainly two contributions of our work: (1) providing

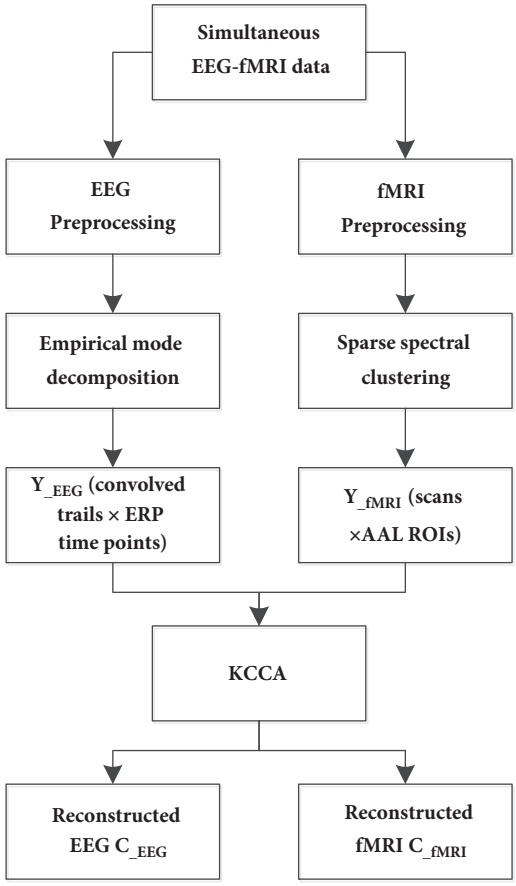


Figure 1: Pipeline of the proposed EEG-fMRI fusion approach.

an effective framework for simultaneous EEG-fMRI fusion and (2) exploring the neural activity under cognitive reappraisal in high spatiotemporal resolution.

2. Materials and Methods

2.1. The EEG-fMRI Fusion Framework. The framework of the fusion approach is demonstrated in Figure 1. Simultaneous EEG-fMRI data is preprocessed, respectively. EMD is further used to eliminate noise of EEG data. Sparse spectral clustering (SSC) is employed to construct fMRI masks that indicate the emotion-related cerebral regions. EEG feature to be fused is defined as $Y_{\_EEG}$ (convolved trails × ERP time points), which are obtained by convolving the ERP values at different time points with a standard HRF. On the other hand, fMRI feature to be fused is defined as $Y_{\_fMRI}$ (scans ×AAL ROIs), which are obtained by calculating mean values in anatomical automatic labeling (AAL) cerebral regions under the constraints of fMRI masks. Then, $Y_{\_EEG}$ and $Y_{\_fMRI}$ are fused using a kernel-based CCA (KCCA) framework. EEG and fMRI components ($C_{\_EEG}$ and $C_{\_fMRI}$) are finally reconstructed based on the selected correlated components.

2.2. Subjects. A total of 15 healthy adults, 5 females and 10 males, aged from 19 to 24 years (M (mean value) =23, SD (standard deviation) =1.48), are recruited from Changzhou University to implement the experiments.

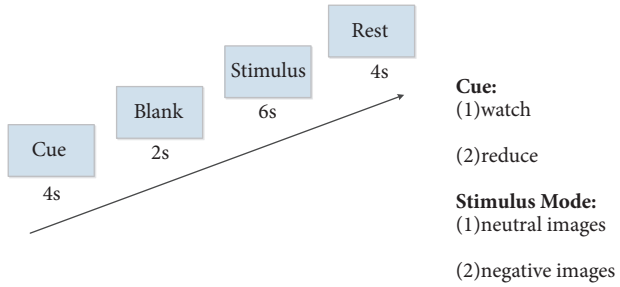


Figure 2: Illustration of the visual stimuli paradigm.

Participants have regular or corrected regular vision without history of neurological, medical, or psychiatric disorders. They have been tested for psychological profile to discard some comorbid issues as depression or psychiatric symptoms that can affect emotional evaluation. All participants provide written informed consent to be part of the experiment, which is approved by the local ethics committee (Changzhou University, Changzhou, China). Each subject receives 42-minute fMRI scan (structure: 5 min, resting state: 5 min, and task state: 32 min).

2.3. Paradigm. The visual stimuli paradigm [21] is implemented in a block fMRI design as shown in Figure 2. The entire experiment for one participant contains 120 trials, including 4 circulations in which 30 trials are implemented. Three conditions, including watching neutral images (e.g., buildings, neutral faces, and food), watching negative images (e.g., sadness, disasters, and violence), and watching negative images with cognitive reappraisal, are randomly implemented in 40 trials, respectively. All the images used are chosen from the international affective picture gallery. The arousal for neutral images is M (mean) = 2.91 and SD (standard deviation) = 1.93; meanwhile, for negative images it is M = 5.71 and SD = 2.61. Procedure of a single trial can last at most 16 seconds as proposed in [22]. Initially, cue word "reduce" or "watch" is shown on the screen for 4 seconds in the cue period. After a 2-second blank period, the stimulus period will last for 6 seconds. At this period, neutral and negative images will randomly appear with cue word "watch", while only negative images will appear with cue word "reduce". Notably, cognitive reappraisal will be used if the cue word "reduce" appears. Finally, the rest period will last for 4 seconds.

2.4. Simultaneous EEG-fMRI Acquisition. EEG acquisition system of EGI company (Eugene, the USA) is used in the experiment. EEG is sampled continuously at 1000Hz. An amplifier is placed inside the MR scanner room. Subjects are fitted with an electrode cap containing 64 electrodes with Cz as online reference. Later, the data are referenced to zero by reference electrode standardization technique [23]. It is recently confirmed being close to the idea of zero reference [24, 25]. Impedances are kept low below 50kΩ. The helium pump is turned off during experiments to avoid related artifacts.

Functional imaging data are sampled with 3-Tesla superconducting type nuclear magnetic resonance imaging system

Figure 3: The foam pads used to prevent head movement.

of Philips Company. Single excitation gradient echoes planar sequence is utilized to acquire functional images. After a whole paradigm finished, 960 BOLD sensitive echo planar images (EPI) are gathered during four sessions. EPI volumes are aligned with the anterior-posterior commissural line. It contains 24 axial slices with 4mm thickness including flip angle: 90 degree; TR (repetition time): 2s; TE (echo time): 35ms; FOV (field of view): 230mm∗182mm; matrix: 96×74. Subjects are mandated to lie in the MRI scanning room, staying awake, and blinking as little as possible. The foam pads (Figure 3) are used to prevent head movement.

2.5. EEG Data Processing. Processing of EEG data contains two parts, one is denoising and the other is extracting EEG feature. In consideration of the influence caused by MR scanning, denoising is achieved through two steps: traditional denoising using net-station toolbox and further increasing SNR using EMD.

For traditional denosing, noises such as gradient artifact, ECG, and power interference are eliminated as follows: (1) Gradient artifact is removed by template elimination method. The gradient artifact template is constructed in a weighted average mean by labeling the timing that fMRI triggered EEG. Then, an average artifact subtraction method is utilized to eliminate the gradient artifact. (2) Band-pass filtering is employed with the band 0.01-40Hz. (3) Optimal basis set approach is used to eliminate ballistocardiogram artifacts caused by the heartbeat. (4) The EEG data are segmented into different fragments based on the stimulus time point. Each fragment ranges from 200ms before stimulus and 1500ms after it. (5) Artifacts such as head movements and blinking are detected in all fragments of all electrodes. The electrode with artifacts is labeled as bad electrode. (6) The bad electrode is replaced by the average of its 3 surrounding electrodes. (7) The first 200ms of each fragment is used for baseline correction.

After traditional denoising, EMD is employed to further increase SNR of EEG data that is affected by MR scanning [26]. EMD tries to find functions which form a complete and nearly orthogonal basis of the original signal. These functions are termed as Intrinsic Mode Functions (IMFs). Then, increasing SNR can be achieved through removing IMFs that are taken as disturbance. Details of increasing SNR through EMD can be found in our former work [27].

After denoising, emotion-related ERP extracted from EEG is used to study the neural activity under cognitive reappraisal [7]. Amplitudes of ERP (extracted from Poz channel) at different time points are termed as EEG feature. At each time point, the trial-to-trial dynamics are convolved with a standard HRF to coincide with fMRI (5 volumes in each trial) due to the BOLD delay. We restrict the analysis to 900ms (225 uniform and consecutive time points) after stimulus onset because the most emotion-related components in the EEG are considered to appear during the first 200-700ms after stimulus onset. Finally, the dimension of the extracted EEG feature is 600 (convolved trails) × 225 (ERP time points).

2.6. FMRI Data Processing. FMRI data are processed using reference electrode standardization technique and statistical parametric mapping (SPM) to correct slice time and exclude head motion. Then the data are normalized and further registered to the Montreal Neurological Institute (MNI) space. Finally, a Gaussian filter (full-width at half-maximum of 8 mm) is used for smoothing filtering and only five fMRI activation regions (three in stimulus period and two in rest period) after stimulus presentation are selected in each trial. Each fMRI activation region is represented by its mean values in different AAL ROIs [28]. Finally, the dimension of the extracted fMRI feature is 600 (scans) × 90 (AAL ROIs).

Notably, some fMRI regions irrelevant to emotion processing are also activated. These undesired activation regions should be removed to guarantee the accuracy of EEG-fMRI fusion. Otherwise, they may correlate with the wanted EEG components. In this work, an fMRI mask is constructed through spatiotemporal clustering of all fMRI activation. SSC is used to cluster the fMRI activation because SSC is insensitive to the number of features and, thus, can avoid dimension disaster [29]. Then, there is no undesired fMRI activation in the fMRI mask because undesired activation mostly sustain for a short period in certain cerebral regions. Finally, for each row of the fMRI feature, an "and" operation with fMRI mask will be performed to restrain the influence of undesired fMRI activation.

2.7. Simultaneous EEG-fMRI Data Fusion Using KCCA. CCA searches for a pair of linear transformations of the variable set in the manner of one for each. It is commonly used for symmetric EEG-fMRI analysis. Given two data X ($Y_{\_EEG}$) and Y ($Y_{\_fMRI}$), their generative models are given by

$$X = A_X C_X$$
$$Y = A_Y C_Y \quad (1)$$

where A_X and A_Y are canonical variate matrices and C_X and C_Y are associated EEG and fMRI components. Let a_{Xk} and a_{Yk} represent the k^{th} column of A_X and A_Y (the k^{th} pair of canonical variate); then their relational degree is defined as

Table 1: Correlated components of EEG-fMRI with high relational degrees (> 0.55) under three emotional states.

correlated components	relational degrees (KCCA approach / CCA approach)		
	neutral state	negative state	reappraisal state
component 1	**0.951** / 0. 944	**0.971** / 0.966	**0.932** / 0.913
component 2	**0.892** / 0.888	**0.952** / 0.947	**0.834** / 0.801
component 3	0.833 / **0.841**	0.863 / **0.877**	**0.805** / 0.731
component 4	**0.765** / 0.741	**0.821** / 0.816	0.704 / **0.716**
component 5	0.643 / **0.681**	0.753 / **0.765**	**0.613** / 0.606
component 6	**0.586** / N/A	**0.712** / 0.660	**0.551** / 0.537
component 7	N/A / N/A	**0.605** / 0.584	N/A / N/A

$$\rho_k = \frac{a_{Xk}{}^T S_{XY} a_{Yk}}{\sqrt{a_{Xk}{}^T S_{XX} a_{Xk}} \times \sqrt{a_{Yk}{}^T S_{YY} a_{Yk}}} \tag{2}$$

$$S = S(X, Y) = \begin{bmatrix} S_{XX} & S_{XY} \\ S_{YX} & S_{YY} \end{bmatrix} \tag{3}$$

where ρ_k indicates the relational degree of the k^{th} pair of associated components. The total covariance matrix S is represented as a block matrix. The within-sets covariance matrices are S_{XX} and S_{YY}. The between-sets covariance matrices are $S_{XY} = S_{YX}{}^T$. Then, those associated components whose relational degrees are larger than a given threshold (0.55) are used to reconstruct the wanted EEG component $\widehat{C}_X$ and fMRI component $\widehat{C}_Y$, which are defined as follows:

$$\begin{aligned} \widehat{C}_X &= \left(\widehat{A}_X^T \widehat{A}_X\right)^{-1} \widehat{A}_X^T X \\ \widehat{C}_Y &= \left(\widehat{A}_Y^T \widehat{A}_Y\right)^{-1} \widehat{A}_Y^T Y \end{aligned} \tag{4}$$

where $\widehat{A}_X$ and $\widehat{A}_Y$ only contain the selected pairs of canonical variate. Details of solving a CCA problem can be referred to [19].

However, CCA cannot process nonlinear data. Thus, kernel is used to resolve such problem through mapping data into a high dimensional feature space. A kernel κ for all X, Y $\in$ R is defined as follows:

$$\kappa(X, Y) = \langle \varphi(X), \varphi(Y) \rangle \tag{5}$$

where φ is a mapping from the original data space R to a new feature space F (φ: R->F). Great flexibility can be achieved by applying different kernels such as linear kernel, Gaussian kernel, and RBF. Based on kernel, the directions a_{Xk} and a_{Yk} can be represented as follows:

$$\begin{aligned} a_{Xk} &= X\alpha \\ a_{Yk} &= Y\beta \end{aligned} \tag{6}$$

where α and β indicate the transformations from original data to their canonical variate. Then, (2) can be represented as follows:

$$\rho_k = \frac{\alpha' X' X Y' Y \beta'}{\sqrt{\alpha' X' X X' X \alpha \bullet \beta' Y' Y Y' Y \beta}} \tag{7}$$

Notable, linear transformations X'X and Y'Y cannot process nonlinear data very well. Hence, RBF kernel is used to replace the linear transformations due to its superiority in processing nonlinear data. Then, (2) can be rewritten as follows:

$$\rho_k = \frac{\alpha' K_X K_Y \beta}{\sqrt{\alpha' K_X{}^2 \alpha \bullet \beta' K_Y{}^2 \beta}} \tag{8}$$

where K_X and K_Y represent the RBF kernel matrices. Relational degrees calculated using (8) is more suitable to nonlinear EEG-fMRI data than that calculated using (2).

3. Experimental Results

3.1. Comparisons between KCCA and CCA. This work focuses on the highly correlated components between EEG temporal evolution and fMRI spatial activation. Ninety correlated components are obtained using KCCA fusion. Table 1 demonstrates the correlated components whose relational degrees are larger than 0.55. As shown in the table, there are six pairs of correlated components under neutral and reappraisal states, and seven pairs of correlated components under negative state. Our former work using CCA fusion is used as comparison [27]. It is obvious that relational degrees obtained using KCCA is larger than that obtained using CCA.

Figures 4–6 illustrate the fifteen subjects' superposed average results of the correlated EEG-fMRI components. Correlated components whose relational degrees are larger than 0.55 are used for superposed average. For each figure, subfigure (a) indicates the superposed average result of correlated EEG component extracted from Poz electrode. Furthermore, x-axis represents time (ms) and y-axis represents normalized amplitude (dimensionless). Subfigure (b) illustrates the correlated fMRI activation under the same state, while the color-bar indicates the normalized activated intensity. Then, neural activities caused by the same stimuli can be observed in both high temporal (correlated EEG component) and spatial resolutions (correlated fMRI activation).

Aside from the differences in relational degrees, differences in EEG components are also evaluated between CCA [27] and KCCA. Figure 7(a) illustrates the fifteen subjects' superposed average results of reconstructed EEG

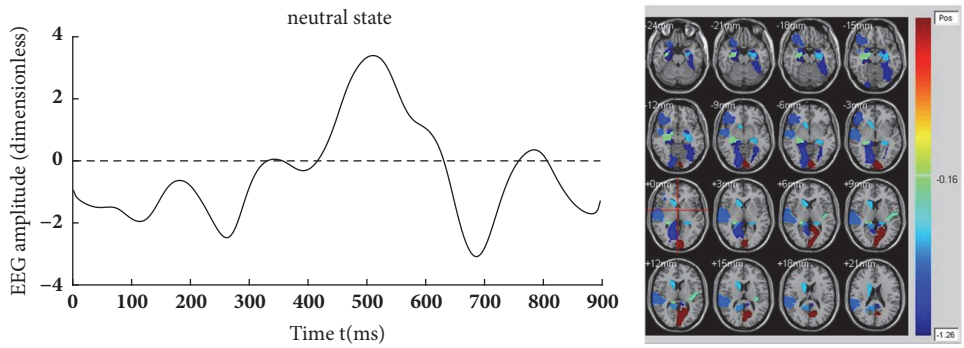


Figure 4: Fifteen subjects' superposed average result of correlated EEG-fMRI under neutral state. (a) Correlated EEG component and (b) correlated fMRI activation.

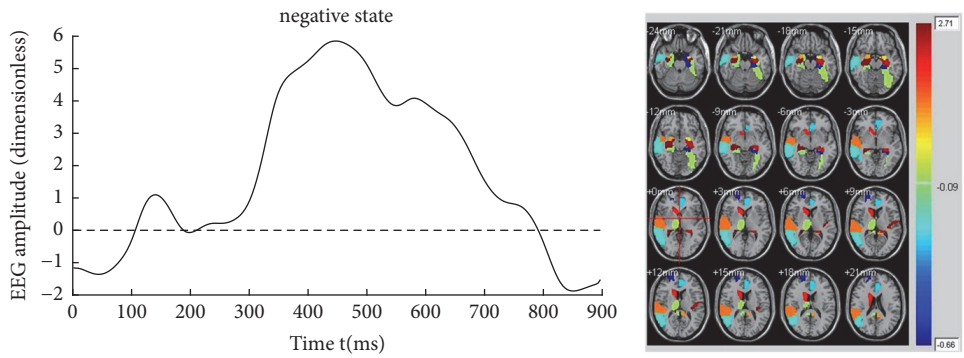


Figure 5: Fifteen subjects' superposed average result of correlated EEG-fMRI under negative state. (a) Correlated EEG component and (b) correlated fMRI activation.

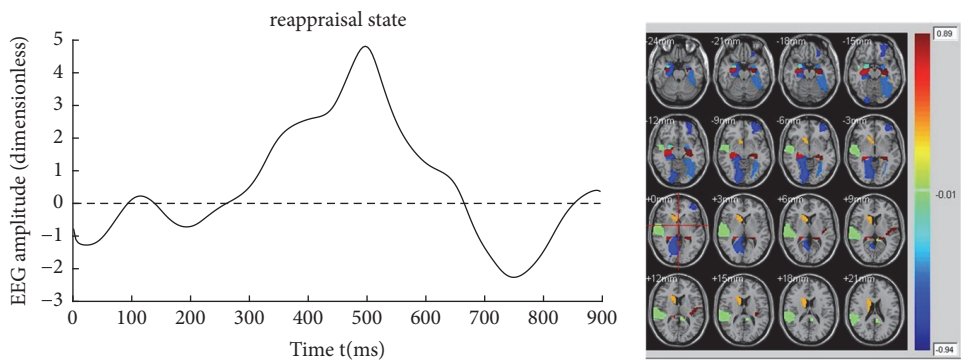


Figure 6: Fifteen subjects' superposed average result of correlated EEG-fMRI under reappraisal state. (a) Correlated EEG component and (b) correlated fMRI activation.

components that are calculated by KCCA fusion. All the EEG components are extracted from Poz electrode. Obvious differences can be observed among their LPP components. The amplitude of LPP component under reappraisal state is smaller than that under negative state and is obviously larger than that under neutral state. Figure 7(b) illustrates the fifteen subjects' superposed average results of reconstructed EEG components that are calculated by CCA fusion. Similar results can be observed. However, amplitudes of LPP component under negative state and that under reappraisal state are more or less intersecting at the reported emotion arousing period (200-700ms by [7]) as illustrated in Figure 7(b). The intersection may be caused by nonlinearity of simultaneous EEG-fMRI data.

Table 2: Fifteen subjects' superposed average results of correlated fMRI activations under neutral state, negative state, and reappraisal state (using KCCA).

under neutral state		under negative state		under reappraisal state	
AAL ROIs (No)	Z-score	AAL ROIs (No)	Z-score	AAL ROIs (No)	Z-score
Calcarine_L (43)	0.355	Hippocampus_R (38)	2.711	Heschl_L (79)	0.887
Hippocampus_R (38)	0.208	Heschl_L (79)	2.485	Hippocampus_L (37)	0.870
Heschl_L (79)	0.178	Hippocampus_L (37)	2.317	Hippocampus_R (38)	0.776
Caudate_R (72)	0.068	Caudate_R (72)	1.990	Caudate_R (72)	0.550
N / A	N / A	Temporal_Sup_R (82)	1.383	Amygdala_ R (42)	0.043
N / A	N / A	Cingulum_Post_L (35)	1.249	Temporal_Sup_R (82)	0.038
N / A	N / A	Amygdala_R (42)	1.233	Amygdala_L (41)	0.031
N / A	N / A	Cingulum_Mid_R (34)	0.672	Cingulum_Post_L (35)	0.006
N / A	N / A	Cingulum_Mid_L (33)	0.627	N / A	N / A
N / A	N / A	Amygdala_L (41)	0.567	N / A	N / A
N / A	N / A	Fusiform_L (55)	0.223	N / A	N / A
N / A	N / A	Thalamus_R (78)	0.220	N / A	N / A
N / A	N / A	Cingulum_Post_R (36)	0.121	N / A)	N / A
N / A	N / A	ParaHippocampal_R (40)	0.090	N / A	N / A

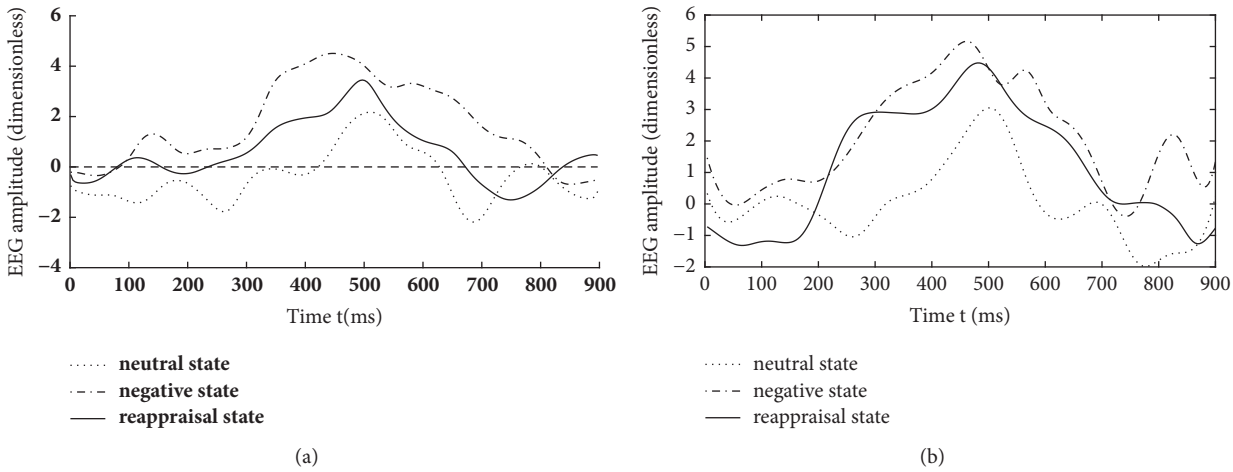


Figure 7: Fifteen subjects' superposed average results of EEG correlation components under three emotional states using (a) KCCA fusion and (b) CCA fusion.

A quantitative comparison is performed on the average correlated EEG components of fifteen subjects under three emotional states. 225 samples are uniformly sampled from 700ms EEG component and their amplitudes are used as input. F-test is used for evaluation and different emotional states are used as the factors of ANOVA. The result shows distinct differences in EEG components of different emotional states. The mean of the differences (MOD) between conditions under negative and neutral states is 23, with $F(1, 224) = 262.65 (P < 0.01)$. The MOD between conditions under reappraisal and negative states is 11, with $F(1, 224) = 70.49 (P < 0.01)$. The MOD between conditions under reappraisal and neutral states is 13, when $F(1, 224) = 83.04 (P < 0.01)$. Obviously, the quantitative result is confirmed to the result of Figure 7.

3.2. Comparisons between KCCA and GLM. Comparisons between KCCA and GLM are performed to verify the superiority of symmetric EEG-fMRI analysis in studying the neural activities of cognitive reappraisal. Table 2 demonstrates fifteen subjects' superposed average results of correlated fMRI activation under three emotional states using KCCA. Intensities of fMRI activation are measured by the Z-score values in different AAL ROIs. A big Z-score value indicates a strong fMRI activation. Notably, only AAL ROIs whose

Table 3: Fifteen subjects' superposed average results of correlated fMRI activations under neutral state, negative state, and reappraisal state (using GLM).

under neutral state		under negative state		under reappraisal state	
AAL ROIs (No)	Z-score	AAL ROIs (No)	Z-score	AAL ROIs (No)	Z-score
Parietal_Sup_L (59)	0.516	Heschl_L (79)	1.587	Parietal_Inf_R (62)	0.887
Paracentral_Lobule_R (70)	0.366	Parietal_Sup_L (59)	1.466	Parietal_Sup_L (59)	0.870
Parietal_Sup_R (60)	0.159	Parietal_Sup_R (60)	1.039	Occipital_Mid_R (52)	0.776
Occipital_Mid_L (51)	0.020	Precuneus_L (67)	0.922	ParaHippocampal_L (39)	0.350
N / A	N / A	Paracentral_Lobule_L (69)	0.790	Angular_R (66)	0.006
N / A	N / A	Paracentral_Lobule_R (70)	0.725	N / A	N / A
N / A	N / A	Occipital_Mid_R (52)	0.569	N / A	N / A
N / A	N / A	Occipital_Mid_L (51)	0.507	N / A	N / A
N / A	N / A	Occipital_Sup_R (50)	0.478	N / A	N / A
N / A	N / A	Parietal_Inf_L (61)	0.292	N / A	N / A
N / A	N / A	Temporal_Pole_Mid_L (87)	0.159	N / A	N / A
N / A	N / A	SupraMarginal_L (63)	0.126	N / A	N / A
N / A	N / A	Precuneus_R (68)	0.116	N / A	N / A
N / A	N / A	SupraMarginal_R (64)	0.114	N / A	N / A
N / A	N / A	Heschl_R (80)	0.108	N / A	N / A
N / A	N / A	Cingulum_Mid_L (33)	0.084	N / A	N / A
N / A	N / A	ParaHippocampal_L (39)	0.080	N / A	N / A
N / A	N / A	Cingulum_Ant_L (31)	0.022	N / A	N / A

Z-score values are larger than 0 (a negative Z-score value in certain AAL ROI indicates that this ROI is irrelevant to emotion processing) are preserved in this table. Meanwhile, no EEG component is evaluated because GLM is mainly used for analyzing fMRI activation. Table 3 demonstrates fifteen subjects' superposed average results of correlated fMRI activation under three emotional states using GLM. Differences between KCCA and GLM exist in both activated regions and intensities. Discussions of their differences will be given in Section 4 in detail.

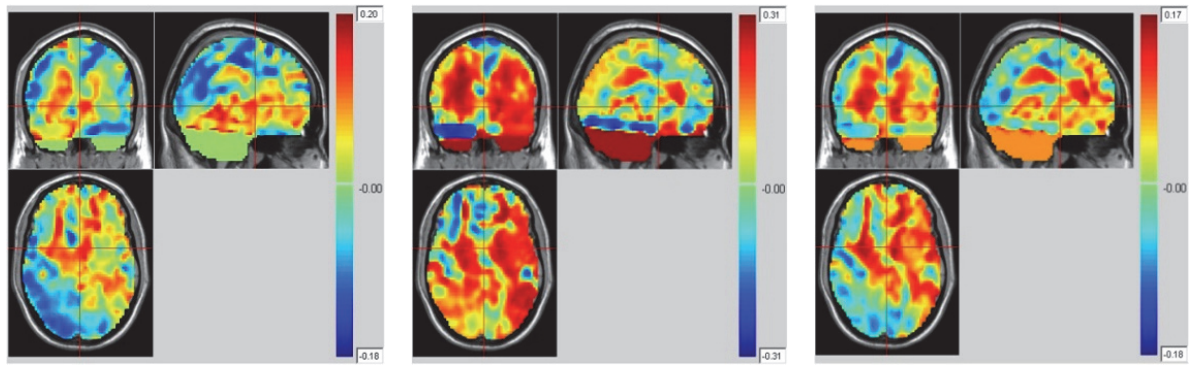
3.3. Evaluation of the fMRI Masks. FMRI masks are used to restrain the activated fMRI regions due to their ability to eliminate the regions uncorrelated to emotion processing. The clustering results of all subjects under three emotional states are illustrated in Figure 8. As shown in the figure, activated regions under neutral state are the smallest while activated regions under negative state are the biggest.

KCCA fusion without fMRI masks is performed to evaluate the effectiveness of fMRI masks. For fMRI, fifteen subjects' superposed average results of correlated fMRI activation obtained through KCCA but without fMRI masks are illustrated in Figure 9. Correlated fMRI activation varies a lot due to whether fMRI masks are used, especially under negative and reappraisal states. For example, there is obvious activation in cerebral regions such as hippocampus, amygdala, and temporal lobe that are directly related to emotion processing in Figure 5(b). However, no activation can be found in these cerebral regions in Figure 9(a). Same phenomena can be observed in Figures 6(b) and 9(b). There is no activation in emotion-related cerebral regions such as hippocampus and temporal lobe in Figure 9(b), while obvious activation can be observed in these regions in Figure 6(b). There is no obvious difference in activated fMRI regions between Figures 4(b) and 9(c) because fMRI masks do not focus on cerebral regions unrelated to emotion processing.

For EEG, fifteen subjects' superposed average results of EEG correlated components under three emotional states but without clustering mask are shown in Figure 10. Compared with the results in Figure 7, no obvious decrease can be observed in ERP amplitude from negative state to reappraisal state. Meanwhile, EEG evolutions under different emotional states are hard to separate.

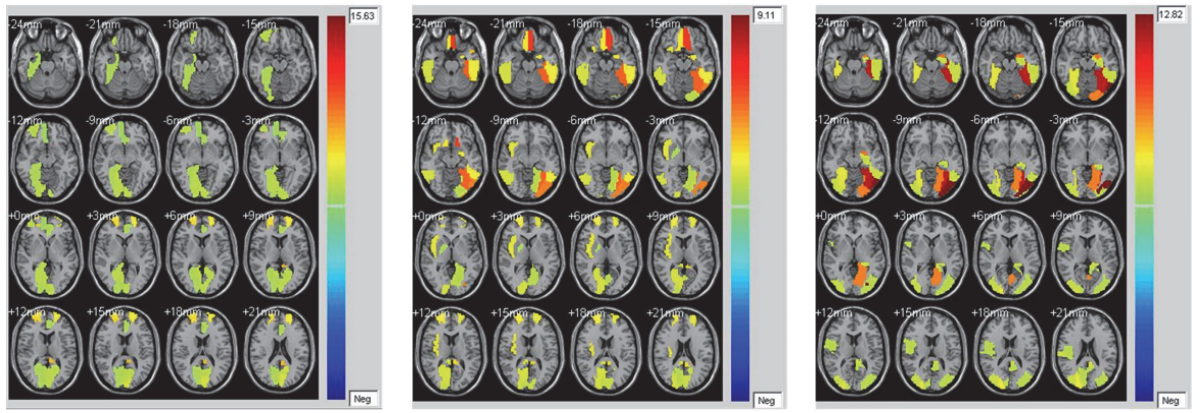
4. Discussion and Conclusion

The aim of cognitive reappraisal is to regulate human experience under negative emotion such as depression, fear, and disappointment. Simultaneous EEG-fMRI analysis is used to study the neural activity under cognitive reappraisal due to its complementarity in both spatial and temporal domains. In this work, these neural activities are studied using a KCCA fusion framework. Meanwhile, EMD is used to further increase SNR of EEG data that is sampled under MR scanning. FMRI masks are calculated using SSC and are used to eliminate the activation unrelated to emotion processing. With all these processing, both EEG and fMRI components can be reconstructed based on the selected correlated components (Figures 4, 5, and 6). Results of these figures are very important to study the mechanism of cognitive reappraisal that is useful for human to regulate



(a) (b) (c)

Figure 8: fMRI clustering results of all subjects (a) under neutral state; (b) under negative state; and (c) under reappraisal state (color-bar indicates the activated intensity).



(a) (b) (c)

Figure 9: Fifteen subjects' superposed average results of correlated fMRI activation using the proposed method without fMRI masks (a) under negative state, (b) under reappraisal state, and (c) under neutral state.

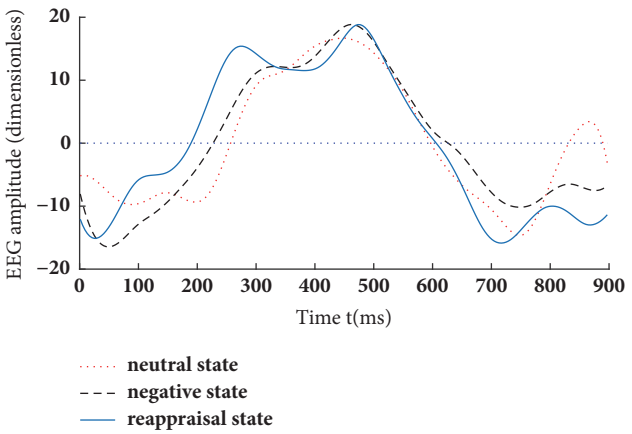


Figure 10: Fifteen subjects' superposed average results of EEG correlated components under three emotional states using the proposed method but without clustering mask.

his/her emotion. For spatial analysis, activation in emotion-related cerebral regions (e.g., amygdala, hippocampus, and temporal lobe) under reappraisal state is obviously weaker than that under negative state through introducing the cognitive reappraisal strategy. It reveals that negative emotion can be effectively restricted in emotion-related cerebral regions after applying cognitive reappraisal strategy. For temporal analysis, obvious differences can be observed among different LPP components which are considered to be highly correlated to emotion processing. Peak value of LPP component under reappraisal state is smaller than that under negative state, and obviously larger than that under neutral state. Both the shrunken fMRI activated regions and decreased peak value of LPP component verify the assumptions that negative emotions, e.g., sorrow, fear, and disappointment, can be restrained by using cognitive reappraisal.

Effectiveness of kernel strategy can be observed through the comparisons between KCCA and CCA. CCA fusion is widely used for symmetric EEG-fMRI analysis. However, nonlinearity of the EEG-fMRI data may decrease the fusion accuracy. Thus, we improve the CCA fusion with a kernel strategy. It is not very novel but is effective. KCCA fusion is specially designed to process nonlinear EEG-fMRI data. As

shown in Table 1, relational degrees of correlated components derived using KCCA fusion are mainly larger than that derived by CCA fusion. Notably, a larger relational degree indicates a stronger relationship between two components. Thus, the results in Table 1 may indicate the superiority of KCCA fusion to traditional CCA fusion.

The superiority of KCCA fusion to CCA fusion can be also observed from the reconstructed EEG and fMRI components. For fMRI that concentrates on spatial activation, no obvious activation can be observed in hippocampus which is emotion-related under negative or reappraisal states using CCA fusion. It may be caused by the fact that CCA cannot process nonlinear EEG-fMRI data. However, obvious activation can be observed in these regions under the same emotional states using KCCA fusion. It reveals the ability of KCCA in mining effective fMRI activation from nonlinear EEG-fMRI data. For EEG that concentrates on temporal evolutions, amplitude of LPP component under reappraisal state is obviously weaker than that under negative state at the same period using KCCA fusion. The decrease in amplitude indicates the ability of cognitive reappraisal to restrain sorrowful emotion, as pointed out by [4]. However, no obvious decrease can be observed in amplitude of LPP component from negative state to reappraisal state using CCA fusion. Thus, the larger relational degrees, the more fMRI activation, and the obvious decrease in amplitude of LPP component between negative and reappraisal states reveal the superiority of KCCA fusion to CCA fusion. Such superiority is obtained due to the effect of kernel strategy in processing nonlinear EEG-fMRI data.

The superiority of symmetric EEG-fMRI analysis to EEG informed fMRI analysis can be observed through the comparisons between KCCA and GLM (Tables 2 and 3). Only fMRI activation is compared because GLM cannot be used to study the EEG evolutions. Then, for KCCA (Table 2), obvious activation can be observed under negative state in cerebral regions such as the temporal lobe, the hippocampus, the amygdala, and the cingulate gyrus. Meanwhile, activation can be observed under reappraisal state in cerebral regions such as the amygdala, the temporal lobe, the cingulate gyrus, the hippocampus, and the frontal lobe. These activation regions indicate the important role of these cerebral regions in emotional regulation. In the perspective of activated intensity (Z-score), activation in cerebral regions under reappraisal state, especially the regions (e.g., the amygdala, the hippocampus, and the temporal lobe) directly related to emotion processing, is obviously weaker than activation in those regions under negative state through using cognitive reappraisal. Activation in these cerebral regions under neutral state is much weaker than activation in the same regions under the other two states. All these results are basically consistent with the conclusions proposed by [28]. Compared with the fusion results obtained using GLM (Table 3), two results can be concluded: (1) by utilizing KCCA approach, more regions are found to be activated under negative and reappraisal states, and (2) activated intensities of these regions calculated using KCCA fusion are larger than those calculated using GLM fusion. Both results indicate the superiority of KCCA fusion (symmetric EEG-fMRI analysis) in studying the neural activity of cognitive reappraisal.

As a special preprocessing, fMRI masks are useful due to the assumption that strong fMRI activation uncorrelated with emotion processing may be correlated with EEG components, thus leading to omitting the fMRI activation which we are truly interested in. As shown in Figures 4(b), 5(b), and 6(b), obvious fMRI activation can be observed in emotion-related cerebral regions such as the hippocampus and the temporal lobe under negative and reappraisal states. However, no activation can be observed in these regions if fMRI masks are not used as preprocessing. Meanwhile, obvious decrease can be observed in EEG amplitude from negative state to reappraisal state using our fusion approach (Figure 7(a)), while little decrease can be observed under the same condition without fMRI masks (Figure 10).

Based on the above discussions, our fusion approach may provide a fine solution for analyzing simultaneous EEG-fMRI data in high resolution spatiotemporal domains. It can synchronously tell when and where the neural activities related to certain tasks such as cognitive reappraisal occur. It may also provide a useful technological means for fusion-based cerebral area positioning, ERP-induction time determination, and brain imaging feature extraction in the area of brain-human interface. Our fusion approach can be also used in paradigms which can cause LPP with further study on cognitive researches and clinical trials.

There are still some limitations in the proposed fusion approach: (1) for the data to be fused, activation in AAL ROIs is employed instead of original fMRI voxels, aiming at reducing computational complexity. As a result, one cannot study the activation of reconstructed fMRI components at voxel level. (2) Prior knowledge is necessary for our KCCA fusion approach. It is hard to choose suitable parameters that are significant for satisfactory fusion results. There is also no certain criterion in determining the threshold of relational degrees. (3) The number of enrolled subjects is far from enough; thus, the evaluations may lack persuasion. Our future work focuses on implementing EEG-fMRI fusion at voxel level instead of AAL ROIs. Thus, the spatial resolution of reconstructed fMRI components can be greatly boosted.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Biao Yang and Tiantong Zhou contributed equally to this paper.

Acknowledgments

This work has been partially supported by the National Natural Science Foundation of China (61201096 and 61501060), the Natural Science Foundation of Jiangsu Province (BK20150271), and Qing Lan Project of Jiangsu Province.

References

[1] G. A. Bonanno, A. Papa, K. Lalande, M. Westphal, and K. Coifman, "The importance of being flexible: The ability to both enhance and suppress emotional expression predicts long-term adjustment," *Psychological Science*, vol. 15, no. 7, pp. 482–487, 2004.

[2] R. J. Davidson, K. M. Putnam, and C. L. Larson, "Dysfunction in the neural circuitry of emotion regulation—a possible prelude to violence," *Science*, vol. 289, no. 5479, pp. 591–594, 2000.

[3] D. Driscoll, D. Tranel, and S. W. Anderson, "The effects of voluntary regulation of positive and negative emotion on psychophysiological responsiveness," *International Journal of Psychophysiology*, vol. 72, no. 1, pp. 61–66, 2009.

[4] S. Paul, D. Simon, R. Kniesche, N. Kathmann, and T. Endrass, "Timing effects of antecedent- and response-focused emotion regulation strategies," *Biological Psychology*, vol. 94, no. 1, pp. 136–142, 2013.

[5] J. J. Gross, "Antecedent- and response-focused emotion regulation: divergent consequences for experience, expression, and physiology," *Journal of Personality and Social Psychology*, vol. 74, no. 1, pp. 224–237, 1998.

[6] S. H. Kim and S. Hamann, "The effect of cognitive reappraisal on physiological reactivity and emotional memory," *International Journal of Psychophysiology*, vol. 83, no. 3, pp. 348–356, 2012.

[7] G. Hajcak, A. Macnamara, and D. M. Olvet, "Event-related potentials, emotion, and emotion regulation: an integrative review," *Developmental Neuropsychology*, vol. 35, no. 2, pp. 129–155, 2010.

[8] A. MacNamara, K. N. Ochsner, and G. Hajcak, "Previously reappraised: The lasting effect of description type on picture-elicited electrocortical activity," *Social Cognitive and Affective Neuroscience*, vol. 6, no. 3, pp. 348–358, 2011.

[9] G. Hajcak, J. S. Moser, and R. F. Simons, "Attending to affect: Appraisal strategies modulate the electrocortical response to arousing pictures," *Emotion*, vol. 6, no. 3, pp. 517–522, 2006.

[10] J. W. Krompinger, J. S. Moser, and R. F. Simons, "Modulations of the Electrophysiological Response to Pleasant Stimuli by Cognitive Reappraisal," *Emotion*, vol. 8, no. 1, pp. 132–137, 2008.

[11] S. H. Kim and S. Hamann, "Neutral correlates of positive and negative emotional regulation," *Journal of Cognitive Neuroscience*, vol. 19, pp. 776–798, 2007.

[12] K. N. Ochsner, K. Knierim, D. H. Ludlow et al., "Reflecting upon feelings: An fMRI study of neural systems supporting the attribution of emotion to self and other," *Cognitive Neuroscience*, vol. 16, no. 10, pp. 1746–1772, 2004.

[13] H. L. Urry, "Using Reappraisal To Regulate Unpleasant Emotional Episodes: Goals and Timing Matter," *Emotion*, vol. 9, no. 6, pp. 782–797, 2009.

[14] S. P. Kyathanahally, A. M. Francowatkins, X. Zhang et al., "A realistic framework for investigating decision-making in the brain with high spatio-temporal resolution using simultaneous EEG/fMRI and joint ICA," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–12, 2016.

[15] T. Nguyen, T. Potter, T. Nguyen, C. Karmonik, R. Grossman, and Y. Zhang, "EEG Source Imaging Guided by Spatiotemporal Specific fMRI: Toward an Understanding of Dynamic Cognitive Processes," *Neural Plasticity*, vol. 2016, Article ID 4182483, 10 pages, 2016.

[16] Y. L. Liu, J. Bengson, H. Q. Huang et al., "Top-down modulation of neural activity in anticipatory visual attention: control mechanisms revealed by simultaneous EEG-fMRI," *Neuroscience*, vol. 35, no. 20, pp. 7938–7949, 2015.

[17] R. F. Ahmad, A. S. Malik, N. Kamel, F. Reza, H. U. Amin, and M. Hussain, "Visual brain activity patterns classification with simultaneous EEG-fMRI: A multimodal approach," *Technology and Health Care*, vol. 25, no. 3, pp. 471–485, 2017.

[18] Q. Yu, L. Wu, D. A. Bridwell et al., "Building an EEG-fMRI Multi-Modal Brain Graph: A Concurrent EEG-fMRI Study," *Frontiers in Human Neuroscience*, vol. 10, 2016.

[19] N. M. Correa, T. Eichele, T. Adali, Y.-O. Li, and V. D. Calhoun, "Multi-set canonical correlation analysis for the fusion of concurrent single trial ERP and functional MRI," *NeuroImage*, vol. 50, no. 4, pp. 1438–1445, 2010.

[20] L. Dong, P. Wang, Y. Bin et al., "Local Multimodal Serial Analysis for Fusing EEG-fMRI: A New Method to Study Familial Cortical Myoclonic Tremor and Epilepsy," *IEEE Transactions on Autonomous Mental Development*, vol. 7, no. 4, pp. 311–319, 2015.

[21] L. Yuan, R. Zhou, and S. Hu, "Cognitive reappraisal of facial expressions: Electrophysiological evidence of social anxiety," *Neuroscience Letters*, vol. 577, pp. 45–50, 2014.

[22] J. M. Carlson, D. Foti, L. R. Mujica-Parodi, E. Harmon-Jones, and G. Hajcak, "Ventral striatal and medial prefrontal BOLD activation is correlated with reward-related electrocortical activity: a combined ERP and fMRI study," *NeuroImage*, vol. 57, no. 4, pp. 1608–1616, 2011.

[23] D. Yao, "A method to standardize a reference of scalp EEG recordings to a point at infinity," *Physiological Measurement*, vol. 22, no. 4, pp. 693–711, 2001.

[24] Y. Qin, P. Xu, and D. Yao, "A comparative study of different references for EEG default mode network: The use of the infinity reference," *Clinical Neurophysiology*, vol. 121, no. 12, pp. 1981–1991, 2010.

[25] Y. Tian and D. Yao, "Why do we need to use a zero reference? Reference influences on the ERPs of audiovisual effects," *Psychophysiology*, vol. 50, no. 12, pp. 1282–1290, 2013.

[26] H. J. Dong, "Exotic collections asset pricing: the Lagrangian Optimization," *British Journal of Mathematics & Computer Science*, vol. 5, no. 1, pp. 82–91, 2015.

[27] L. Zou, Y. Yan, and B. Yang, "Feature fusion analysis of simultaneously recorded EEG-fMRI in emotion cognitive reappraisal," *Acat Automatica Sinica*, vol. 42, no. 5, pp. 771–781, 2016.

[28] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou et al., "Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain," *NeuroImage*, vol. 15, no. 1, pp. 273–289, 2002.

[29] L. Zou, Y. Xu, Z. Y. Jiang et al., "Functional connectivity analysis of cognitive reappraisal using sparse spectral clustering method," in *Proceedings of the Fifth International Conference on Cognitive Neurodynamics*, pp. 291–297, 2016.

Hindawi
Computational and Mathematical Methods in Medicine
Volume 2018, Article ID 9674108, 11 pages
https://doi.org/10.1155/2018/9674108

Research Article

Structure Optimization for Large Gene Networks Based on Greedy Strategy

Francisco Gómez-Vela,[1] **Domingo S. Rodriguez-Baena**,[1] **and José Luis Vázquez-Noguera**[2]

[1]*Division of Computer Science, Pablo de Olavide University, 41013 Seville, Spain*
[2]*Carrera de Ingeniería Informática, Universidad Americana, Asunción, Paraguay*

Correspondence should be addressed to Francisco Gómez-Vela; fgomez@upo.es

Received 14 March 2018; Revised 23 April 2018; Accepted 11 May 2018; Published 14 June 2018

Academic Editor: Ting Hu

In the last few years, gene networks have become one of most important tools to model biological processes. Among other utilities, these networks visually show biological relationships between genes. However, due to the large amount of the currently generated genetic data, their size has grown to the point of being unmanageable. To solve this problem, it is possible to use computational approaches, such as heuristics-based methods, to analyze and optimize gene network's structure by pruning irrelevant relationships. In this paper we present a new method, called GeSOp, to optimize large gene network structures. The method is able to perform a considerably prune of the irrelevant relationships comprising the input network. To do so, the method is based on a greedy heuristic to obtain the most relevant subnetwork. The performance of our method was tested by means of two experiments on gene networks obtained from different organisms. The first experiment shows how GeSOp is able not only to carry out a significant reduction in the size of the network, but also to maintain the biological information ratio. In the second experiment, the ability to improve the biological indicators of the network is checked. Hence, the results presented show that GeSOp is a reliable method to optimize and improve the structure of large gene networks.

1. Background

One of the most important challenges in systems biology is to understand how individual biological components behave and interact in the context of large and complex systems [1]. This knowledge provides the opportunity of controlling and/or optimizing different parts of biological processes to generate a specific effect in the whole system. Therefore, this system-wide view may lead to new applications in areas such as biotechnology and medicine [2]. In particular, the high amount of data generated in the last years allows the inference of relationships between DNA, RNA, proteins, and other cellular components. The sum of these interactions leads to various types of interaction networks (including protein-protein interaction, metabolic, signalling, and transcription-regulatory networks) called gene networks for the sake of simplicity.

Gene networks are usually inferred from gene expression data and have been widely used to model gene relationships in a biological process [3]. In the last decade, many computational approaches have been proposed for the reverse engineering of gene networks [4]. However, the continuous advances in high-throughput technologies enable carrying out large-scale analyses on the DNA and RNA levels the same as on the protein and metabolite level. As a result, the sources of data from which the gene networks are inferred have increased in size, complexity, and diversity [2]. Due to this, new computational challenges have arisen. For example, some methods have been redesigned to improve their performance during large-scale dataset processing [5]. Other research works have focused their efforts on integrating different sources of data for a more accurate gene network reconstruction, such as the work of [6], in which time data sets from different perturbation experiments are simultaneously considered, or that in [7], where the proposed model integrates big data of diverse types to increase both the power and accuracy of networks inference. Different inference algorithms are combined for reconstructing genome-scale

and high-quality gene network from massive-scale RNA-seq samples in [8]. Even other works, like [9], adapt known gene network construction methods to highly parallel execution using distributed high-throughput computing resources.

As a result of these new researches, inferred gene networks are more complex and larger. This fact makes it difficult to visually detect interesting connections between nodes, even though analysis tools have been created recently to apply both advanced statistics and innovative visualization strategies to support efficient knowledge extraction from gene networks [10]. Regarding the gene network structure, some pieces of evidence, like those from the analysis of metabolism and genetic regulatory networks, have proven most biological networks to be sparse, following a scale-free topology. That is, the nodal degree distribution of the network is a power law distribution [11]. Scale-free networks are highly nonuniform; that is, most of the nodes have only a few links while a few nodes have a very large number of links, which are called Hubs. Hubs in a network play a crucial role in how the information is processed in the network since they connect different highly interconnected group of nodes (modules) that could represent different biological functions [12]. Nowadays, the generation of gene networks with a scale-free topology is harder due to the great size and complexity of the networks obtained from the high quantity of data available, so the optimization of gene network structures is currently an important challenge.

In this paper, a new method for automatic optimization of the topology of a large gene network is presented. The method, called Gene Network Structure Optimization (GeSOp), is a backward elimination procedure based on a greedy heuristic method to perform a prune of the irrelevant relationships of the input network. Through this novel method, large genetic networks can improve their topological characteristics without losing their biological information.

1.1. Related Works. Explicit structure optimization methods examine networks models and apply a scoring function to assess the degree to which the resulting structure explains the data, while penalizing the complexity of the model. For this aim, interactions are added and/or removed until the best score is reached. Therefore, heuristic search algorithms are one of the most used techniques since exploring all possible combinations of interactions is an NP-hard problem, specially with very big and complex networks [2, 13]. Several optimization techniques have been developed. However, they are usually limited by the high dimensionality of the problem, as well as computational power required for large networks [14].

Some research works use evolutionary techniques. To reduce the large search spaces, elitist selection method is often used in genetic algorithms, ensuring that the algorithm does not waste time in the rediscovery of previously discarded partial solutions. For example, in [15], a random Boolean network is evolved to look for an accurate model based only on experimental data, without taking into account prior biological knowledge. Other research works use other methods to improve the algorithm's performance, like [16] that proposes a multiagent genetic algorithm to reconstruct large-scale gene regulatory networks. This algorithm is based on fuzzy cognitive maps and includes efficient search operators to reduce the search space.

The optimization algorithms that are based on one objective function, for example, error minimization, can lead to over-fitting and many false positive connections in large networks inference. For example, in [17], the inference problem of N genes is decomposed into $N \times (N-1)$ different regression problems, in which the expression level of a target gene is predicted from the expression level of a potential regulation gene by using the sum of squared residuals and the Pearson correlation coefficient. To reduce the over-fitting phenomena, some works use multiple objective functions and/or add prior biological knowledge to infer an accurate network model. For example, authors in [18] import some a priori regulatory information about extracted gene networks from existing publications or biological web sites with the aim of enhancing veracity of the network. The proposal presented in [19] was the first one to incorporate functional association databases. They create undirected, confidence-weighted likelihood matrix by means of pairwise confidence scores from those databases and use it to infer gene networks, improving their accuracy.

Other works focus their efforts on looking for scale-free properties. For example, in [20], a new proposal is presented which takes the scale-free topology into account as prior information to prune the search space during the inference process. This way, the search space traversed by the method integrates the exploration of all predictors sets combinations, like when having a small number of combinations, when performing a floating search, or when the number of combinations becomes excessive.

This process is guided by scale-free prior information. In [21], informative prior based on scale-free property is also used to improve inference accuracy. In particular, during a Bayesian-based inference process, prior knowledge about scale-free properties is used to evaluate the relative importance of nodes from the linkage characteristics of the entire network.

As can be observed, most research works in literature integrate different network structure optimization strategies within the inference process. Therefore, these optimization efforts depend on concrete input data and the network generation tasks. In this sense, to the best of our knowledge, the new method proposed in this paper is the first one that is independent of the network inference process. As a result, this method is able to optimize any input gene network.

2. Materials and Methods

In this section, the methods and the different materials used in this paper are presented. Firstly, the GeSOp method to optimize large gene network structures is exhaustively described. Secondly, the gene network generation method applied in the experimentation will be presented, along with the input datasets and biological databases used.

2.1. Gene Network Structure Optimization. GeSOp is a novel method for large gene networks topology optimization.

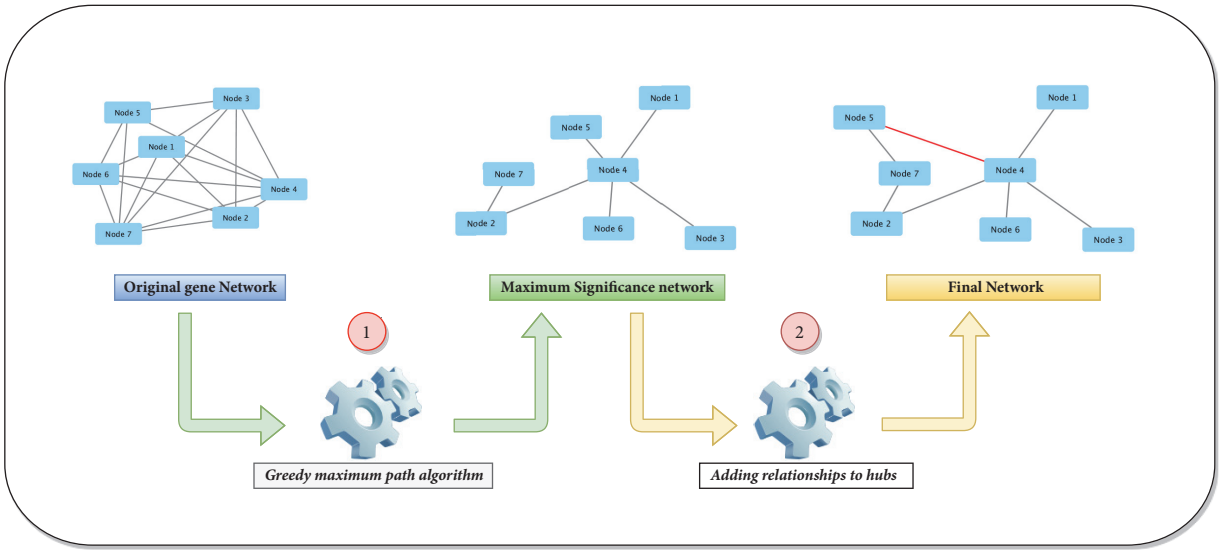


Figure 1: GeSOp method is composed of two different steps: 1. application of a greedy algorithm to prune the original network and 2. detection of Hubs in the resulting network and their enrichment by adding new interactions.

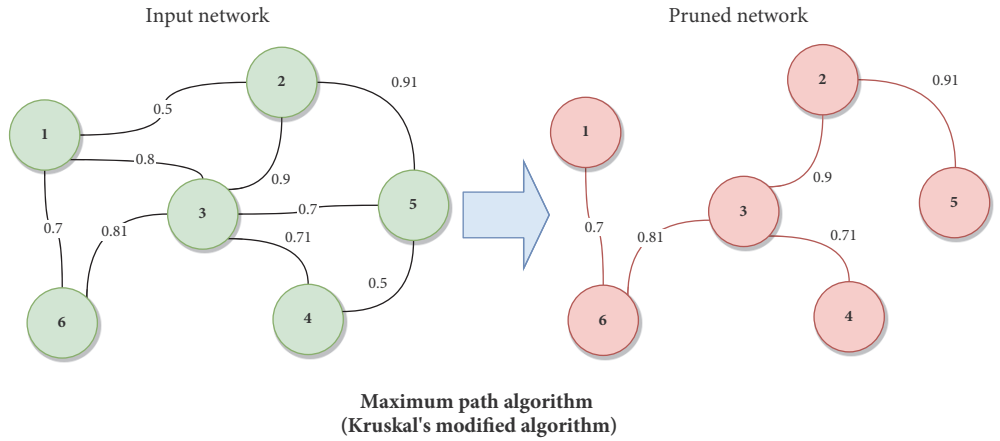


Figure 2: Representation of step 1, in which an input network is pruned using the maximum path algorithm.

The method uses undirected influence networks since they represent the highest level of abstraction in the gene networks as was discussed in [3]. Due to this, our method can be applied for a larger number of networks since almost any gene network can be transformed into a nondirected influence network.

The main goal of the GeSOp is to transform the input gene network into a simpler and more efficient network in terms of information transfer, keeping the biological meaningfulness [2]. For this aim, a new backward removal procedure composed of two different steps has been developed. Initially, GeSOp uses a greedy-based heuristic strategy to prune the original network and select the most biologically relevant interactions. Then, the method looks for the most connected nodes (Hubs) in the resultant network and proceeds by adding relevant interactions which were pruned on the previous step. A description of the general schema of the method, along with a toy example, is shown in Figure 1. A complete description of the two steps and a pseudocode of the method are detailed below.

Step1: Greedy Maximum Relevance Path. The first step of GeSOp uses a greedy-based heuristic algorithm to perform a prune of the input network, taking into account most relevant interactions from a biological point of view (see Figure 2). To do so, a modification of Kruskal's algorithm for the shortest path problem in graphs has been developed [22].

In particular, our method does not select the shortest path between nodes. On the contrary, it selects the longest path according to the weight of edges. Therefore, the relationships with the highest level of significance are selected with respect to the weight of the edges for later network reconstruction.

As a result, the pruned network generated contains the same number of genes (nodes) as the original network but it keeps only most relevant relationships. Hence, it implies a large reduction in terms of the number of edges, while

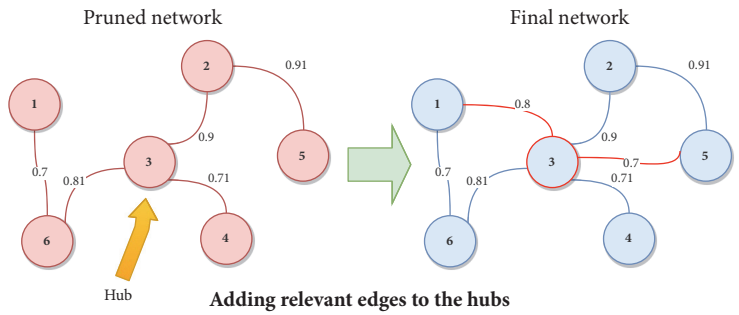


FIGURE 3: An example of the second step of our method, in which the Hubs of the pruned network are identified and relevant edges are added to them. Note that the relationships are added if their weight exceeds the Th_β; in this example, $Th_\beta \geq 0.7$.

still depending on the degree of connectivity of the original network, as is shown in Figure 2.

Step2: Addition of Missing Relationships. As is mentioned in Section 1, Hubs have been reported to have special properties regarding their neighbouring nodes in a gene network. Due to this, in this second step, a topological analysis of the pruned network is performed in order to identify network's Hubs. For this aim, Hubs are selected as those nodes whose connection degree exceeds average network connectivity [12]. A toy example is depicted in Figure 3, where the node "3" is identified as a Hub on the left network.

After the Hubs identification, a threshold (Th_β) is set to determine which relationships of those removed in step 1 should be added to the Hubs. The threshold Th_β is an input parameter of GeSOp algorithm (see Algorithm 1) and it is determined by the user. In this sense, the user may select the threshold which better fits the problem studied. Thus, a new relationship is added to the final network if exceeding Th_β. The process is represented in Figure 3, where two pruned relationships are added to the Hub node in the network on the right.

The final network is generated after each Hub of the pruned networks is processed.

A general pseudocode of the complete method described in this paper is presented in Algorithm 1.

Finally, the complexity of GeSOp combines the complexity of the Step1 ($\Theta(E\log(V))$) and the Step2 ($\Theta(V(E^2))$) resulting in and average case complexity of

$$\Theta\left(E\log\left(V\right)\right) + \Theta\left(V\left(E^2\right)\right), \quad (1)$$

where V and E represent the number of genes and relationships of the input network, respectively.

2.2. Input Datasets. In this section, experimental datasets used for the generation of input gene network used to test GeSOp implementation are shown. In particular, we have selected two different datasets from two different organisms with different features.

Saccharomyces cerevisiae Cell Cycle Dataset. The first dataset used was the one presented by Spellman et al. [23], in relation to the well-known Yeast Cell Cycle. This microarray describes the expression level of 5521 genes in samples from yeast cultures, which were synchronized by three independent methods: α factor arrest, elutriation, and arrest of a cdc15 temperature-sensitive mutant. Particularly, we focus on data generated by cdc15 experiments.

input: Input Network, $G := \langle V, E\rangle$
V: genes, E: relationships
input: Relevant Threshold, Th_β
output: Final network, $G_\beta := \langle V, E_\epsilon\rangle$
, where $E_\epsilon \in E$
/*Step1: maximum path graph*/
$G_\beta \leftarrow$ ***maximumPathAlgorithm***(G);
/*Step2: adding missing edges to Hubs nodes*/
$i \leftarrow 0$;
for $v_i \in V$ **do**
  if $isHub(v_i)$ **then**
    $j \leftarrow 0$;
    for $e_j \in E$ **do**
      if $contains(e_j, v_i) \wedge e_j.weight \geq Th_\beta$ **then**
        $G_\beta \leftarrow addEdge(e_i)$;
      end
      $j \leftarrow j + 1$
    end
  end
  $i \leftarrow i + 1$
end
Return G_β;

ALGORITHM 1: A general pseudocode of the proposed method. The algorithm is divided into two different steps.

Homo sapiens Single Nucleotide Polymorphism (SNP) Dataset. In order to prove the usefulness of our proposed method, the *Homo sapiens* SNP, presented in the work of Hodo et al. [24], has been also selected. This dataset was obtained to study associations of interleukin 28B with carcinoma recurrence in patients with chronic hepatitis C, and it contains information about 54616 genes of *Homo sapiens*.

2.3. Gene Networks Generation Methods. In the following, the methods used to extract gene networks from the two datasets

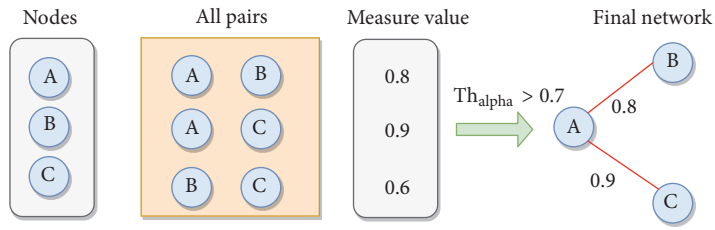


Figure 4: An example of the generation of the input networks. Note that the relationships are added if their weight exceeds the Th_α.

presented above are described. In total, three networks were generated for each dataset. Gene networks based on information theory are one of the most widely used types in literature [2] since they are able to identify coexpression relationships among genes. In this sense, we have selected this kind of networks since they are computationally simple and allow the fitting of large datasets. In particular, three standard measures from information theory to generate coexpression gene networks have been used: **Spearman's** correlation algorithm, **Kendall's** Rank correlation algorithm [1, 25], and **Symmetric Uncertainty** measure (SU) [26, 27].

Gene networks were constructed by calculation of the presented measures (Kendall, Spearman, and SU) from the expression levels in each pair of genes from the input datasets. If the result of the measure exceeds a determinate threshold (here after Th_α) selected by the user, a new edge is added to the network between the nodes as is represented by Figure 4.

For our study, we have selected a low threshold, $Th_\alpha = 0.5$, in order to obtain over-connected networks as was discussed in [3].

2.4. Biological Databases. The aim of this section is to present the biological databases used as reference in the experiment section.

In particular, we have selected three different databases: (a) the GeneMANIA database for evaluating yeast and human networks, (b) YeastNet database for yeast, and (c) HumanNet for human.

GeneMANIA [28] contains information presented in the form of web application for generating hypotheses about gene functions. A prediction server uses a large set of functional association data, including protein and genetic interactions, pathways, coexpression, colocalization, and protein domain similarities. The information stored in GeneMANIA is freely available online. This information is stored in a structure categorized by organisms, where genes (nodes) are related (gene-gene relationship) if at least one piece of evidence of this relation exists in the literature.

YeastNet, which was presented in [29], is a probabilistic functional gene network obtained from 5794 protein-coding genes of the yeast extracted from *Saccharomyces cerevisiae* Genome Database [30]. This network combines protein-protein interactions, protein-DNA interactions, coexpression, phylogenetic conservation, and also literature information, in total covering 102803 linkages among 5483 yeast proteins.

Finally **HumanNet**, which was presented in [31], is a probabilistic functional gene network of 18714 validated protein-coding genes of *Homo sapiens*. It is constructed by modified Bayesian integration of 21 types of "omics" data from multiple organisms. Each data type is weighted according to how well it associates known genes to a biological function in *Homo sapiens*. Each interaction in HumanNet has an associated log-likelihood score that rates the probability of a relationship representing a true functional linkage between two genes.

3. Results and Discussion

The performance of the proposed method was tested by means of two different experiments. The aim of the first experiment is proving that the networks processed by our method do not lose rate of biological information. To this end, we have used different networks, generated using standard methods of literature, and different databases (see Sections 2.3 and 2.4). In the second experiment, a topological analysis of different networks is carried out to check how biological structure indicators are improved.

3.1. Biological Information Analysis. The aim of this experiment is to show how the networks processed by our method reduce the size of the network, keeping their biological information ratio. To do so, for each dataset used, we present a comparison, in terms of size and performance, between the original inferred network and those optimized by GeSOp.

3.1.1. Performance Evaluation. The quality of the optimized networks was assessed by a direct comparison with a gold standard, that is, the biological databases presented in Section 2.2. To compute the quality measures, the following indices were defined as they were presented in [32]:

(i) **True positives (TP):** both networks contain the gene-gene relationship evaluated.

(ii) **False positives (FP):** the input network contains a relationship which is not present in the biological database.

(iii) **True negatives (TN):** the relationships are not present neither in the input network nor in the biological database.

(iv) **False negatives (FN):** the relationship exists in the biological database but it does not in the input network.

Table 1: Results of yeast cell cycle networks processed with GeSOp. As it is shown, networks are significantly reduced in size.

	Yeast								
	Kendall			Spearman			SU		
	Input	GeSOp	diff. %	Input	GeSOp	diff. %	Input	GeSOp	diff. %
Nodes	5466	5466	-	5521	5521	-	4802	4802	-
Edges	619552	10801	-98.25 %	2555009	446704	-82.51%	145329	26421	-81.81%

Once these indices are obtained, other measures used in the literature have been selected to rate the quality of gene networks [2, 3], *Precision* and *Recall* [2, 33], which are defined below.

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

3.1.2. Yeast Experiment. As was stated before, in this subsection, the results obtained by the networks generated by the Yeast Cell Cycle dataset are presented. The input networks were generated using a $Th_\alpha = 0.5$ as cut-off to generate over-connected networks as was introduced in [3]. On the other hand, GeSOp uses a threshold $Th_\beta = 0.7$ for adding relationships. We have selected this threshold as relevant correlation value as was also discussed in [3].

The first analysis is presented in Table 1, in which the number of nodes and edges of the original networks and the optimized ones are exposed.

The table presents the different results obtained by the networks generated by the following methods: Kendall, Spearman, and SU. The first column of each method represents the original input network (network obtained by method on the dataset with $T_h = 0.5$) and the second one ("GeSOp") the final network obtained by our method. On the other hand, the rows of the table represent the number of nodes presented in the network ("Nodes") and the number of relationships comprising the network ("edges"), respectively. Finally, the column "diff. %" represents the difference between the number of edges of the input and final network.

Firstly, it is worth mentioning that the network generation methods present different results for the same dataset. Spearman's method is the one that obtains larger networks since the method is able to find less strictness coexpression levels. On the other hand, SU's method is the most restrictive, as this technique is based on detecting not only the lineal dependencies, but also the nonlinear ones. Finally, Kendall's method is more restrictive than the Spearman method but more relaxed than the SU's.

Regarding the size of the networks, results show that the networks optimized by GeSOp have reduced their size from 81, 81% to 98.25%, in terms of number of edges. Note that GeSOp preserves the nodes, as was described previously. These results represent a significant size reduction, which implies that the final networks are simpler and more user-friendly for researchers in terms of size and visualization.

Once it has been shown that GeSOp is capable of carrying out a reduction in the size of gene networks, it is also important to check if these optimized networks keep the ratio of biological information that they originally contained. For this aim, Tables 2 and 3 are presented. In them, for each method of generation (i.e., Kendall, Spearman, and SU), three columns are displayed. The columns "Input" represent the results for the input network, columns "GeSOp" represent the optimized networks generated by GeSOp. In addition, the results obtained by the networks computed only in step 1 of our method are presented in the "Pruned" columns. The rows "Precision" and "Recall" indicates the ratio of biological information of the networks according to the biological databases used.

Results show that the networks do not suffer any loss of information. On the contrary, the value of the Precision measure for these networks is increased. For example, in the case of the Kendall's network compared to YeastNet, Precision value goes from 0.01 to 0.09, which is a significant improvement. This behaviour is also presented in the Spearman's and SU's networks, where Precision's values increase from 0.01 to 0.02.

Regarding the Recall, it has been reduced in all the networks optimized by our method. This fact makes sense, since Recall value is inversely proportional to the number of FN, which are the relationships that are present in the biological databases. Therefore, our method for reducing the size of the network is inherently increasing the number of FN. Thus, the greater the database used to rate the network, the lower the value of its Recall because there will be more FN.

3.1.3. Homo sapiens Experiment. In this subsection, the experiments carried out by means of the human SNP dataset are described. The obtained networks were generated using the same parameters as in the previous section ($Th_\alpha = 0.5$ and $Th_\beta = 0.7$).

The analysis carried out on the size of the different human networks is shown in Table 4. The results follow the same pattern as of the yeast networks. Spearman is the method which presents the larger network while SU presents the smaller.

GeSOp is able to reduce considerably the size of the networks (e.g., −85.68% for Kendall's network and −89.46% for Spearman's), but the case of SU's network is remarkable. In this case, the reduction is about −40.08%, which is significantly lower than the rest of the cases. This result is consistent with the fact that the SU's network is significantly smaller than the rest of the studied networks, so it is difficult to reduce the size of this network without losing biologically

Table 2: Yeast's network results against YeastNet.

	Kendall			Spearman			SU		
	Input	Pruned	GeSOp	Input	Pruned	GeSOp	Input	Pruned	GeSOp
TP	8331	94	909	19706	64	6589	1744	94	436
FP	444362	4035	9449	1864316	4374	328473	102890	3496	20850
Precision	0.01	0.02	0.094	0.01	0.01	0.02	0.01	0.026	0.02
Recall	0.08	$9.18 \cdot 10^{-4}$	0.009	0.2	$6.25 \cdot 10^{-4}$	0.006	0.01	$9.18 \cdot 10^{-4}$	0.004

Table 3: Yeast's network results against GeneMANIA.

	Kendall			Spearman			SU		
	Input	Pruned	GeSOp	Input	Pruned	GeSOp	Input	Pruned	GeSOp
TP	194918	1942	7863	692753	1909	147360	43991	1722	10281
FP	400383	3273	8423	1770378	3326	293279	95244	2824	18206
Precision	0.32	0.37	0.48	0.28	0.36	0.33	0.31	0.37	0.36
Recall	0.04	$4.01 \cdot 10^{-4}$	0.016	0.08	$3.94 \cdot 10^{-4}$	0.003	0.009	$3.56 \cdot 10^{-4}$	0.002

Table 4: Results of human SNP networks processed with GeSOp. The size of the networks is also significantly reduced.

	Human								
	Kendall			Spearman			SU		
	Input	**GeSOp**	**diff. %**	**Input**	**GeSOp**	**diff. %**	**Input**	**GeSOp**	**diff. %**
Nodes	8068	8068	-	31061	31061	-	1431	1431	-
Edges	68329	9783	-85.68%	5387473	567590	-89.46%	1871	1121	-40.08%

Table 5: Human's network results against GeneMANIA.

	Kendall			Spearman			SU		
	Input	Pruned	GeSOp	Input	Pruned	GeSOp	Input	Pruned	GeSOp
TP	17144	1282	2085	351686	1305	52563	525	299	303
FP	26416	2759	3116	2512234	11646	248969	745	545	553
Precision	0.39	0.31	0.4	0.12	0.10	0.18	0.40	0.35	0.36
Recall	0.0024	$1.83 \cdot 10^{-4}$	$2.98 \cdot 10^{-4}$	0.04	$1.86 \cdot 10^{-4}$	0.0075	$0.7 \cdot 10^{-4}$	$0.4 \cdot 10^{-4}$	$0.43 \cdot 10^{-4}$

Table 6: Human's network results against HumanNet.

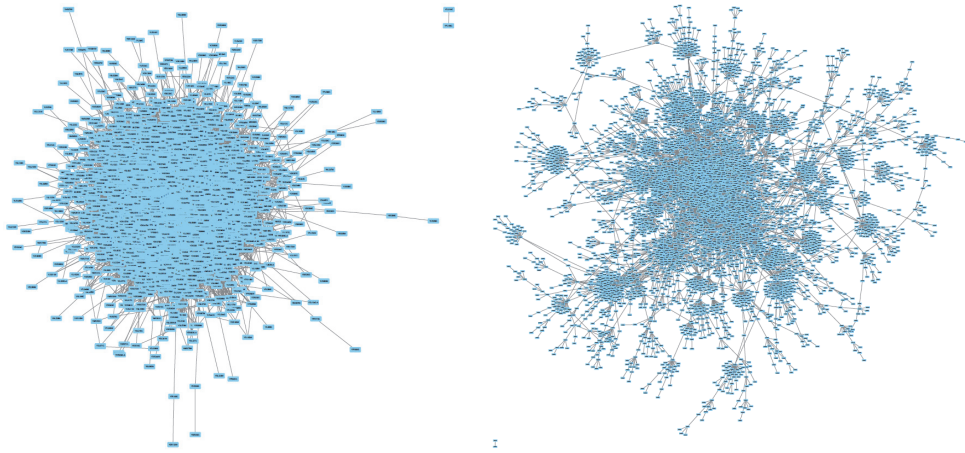
	Kendall			Spearman			SU		
	Input	Pruned	GeSOp	Input	Pruned	GeSOp	Input	Pruned	GeSOp
TP	4216	276	586	46850	141	8202	125	77	77
FP	35931	3291	4084	2465035	10540	258413	1045	699	711
Precision	0.10	0.07	0.12	0.01	0.01	0.03	0.10	0.09	0.09
Recall	0.008	$5.79 \cdot 10^{-4}$	0.001	0.09	$2.95 \cdot 10^{-4}$	0.017	$2.4 \cdot 10^{-4}$	$1.61 \cdot 10^{-4}$	$1.66 \cdot 10^{-4}$

relevant relationships. Due to this result, it is possible to argue that GeSOp performs better with larger gene networks which contain spurious relationships.

The biological validation of the different networks using GeneMANIA and HumanNet databases (see Section 2.2 for more details) is presented in Tables 5 and 6, respectively.

The validation results follow the same pattern as for the yeast networks. The accuracy value increases for all cases except for SU's networks. As was discussed above, it is difficult to prune small networks without losing relevant relationships. Even so, the loss of Precision value is very small (0.04 with GeneMANIA and 0.01 on HumanNet).

In conclusion, the results obtained by both experiments show how GeSOp is able to perform a pruning process on large networks, by reducing their size while keeping their ratio of biological information. The relevance of our method became more evident since, as was discussed in literature [14], the optimization usually implies loss of information in the majority of the cases. However, for almost all analyzed cases, Precision of the network is improved by GeSOp.



(a) $Input_y$ (b) $GeSOp_y$

Figure 5: Visual comparison of yeast network. The original Kendall's network is shown on (a). On (b), the final network obtained with GeSOp is depicted. As can be observed, the optimized network presents a scale-free topology.

3.2. Topological Analysis. In this section, the ability of GeSOp to improve the topology of gene networks is analyzed.

As was stated in Section 1, biological networks usually follow topological patterns, in particular the scale-free topology. The topology of a network is crucial to understand the biological network's architecture and performance [34]. Therefore, gene networks inferred by computational methods should present this type of topology [3]. Based on this assumption, we present a topological analysis of some of the networks optimized by GeSOp in the previous section. The objective is to identify if their topology indicators have been improved in terms of scale-free topology.

Scale-free networks have a structure containing only a few Hubs, among some other features. The most important and commonly used topological features of scale-free networks are presented [35, 36] as follows:

(i) **Characteristic path length (CPL):** The CPL of a network indicates the shortest path length between two nodes, averaged over all pairs of nodes comprising the network. A high path length indicates that the network is in a linear chain. A lower value means that is more compact. Scale-free networks usually have a great CPL.

(ii) **Diameter**: The diameter of a network indicates the maximal distance between two nodes. As in the case of CPL, a greater diameter of the network indicates that it follows a biological pattern.

(iii) **Clustering coefficient**: For one node, this coefficient can be calculated as the number of links among the nodes within its neighbourhood divided by the number of links that are possible among them. A high clustering coefficient for a network is another indicator of the existence of biological relationships.

(iv) **Graph density:** The density of a network defines the ratio of the number of edges to the number of possible edges. Gene networks are generally sparsely connected. Therefore, a low density should indicate biological meaning in the network.

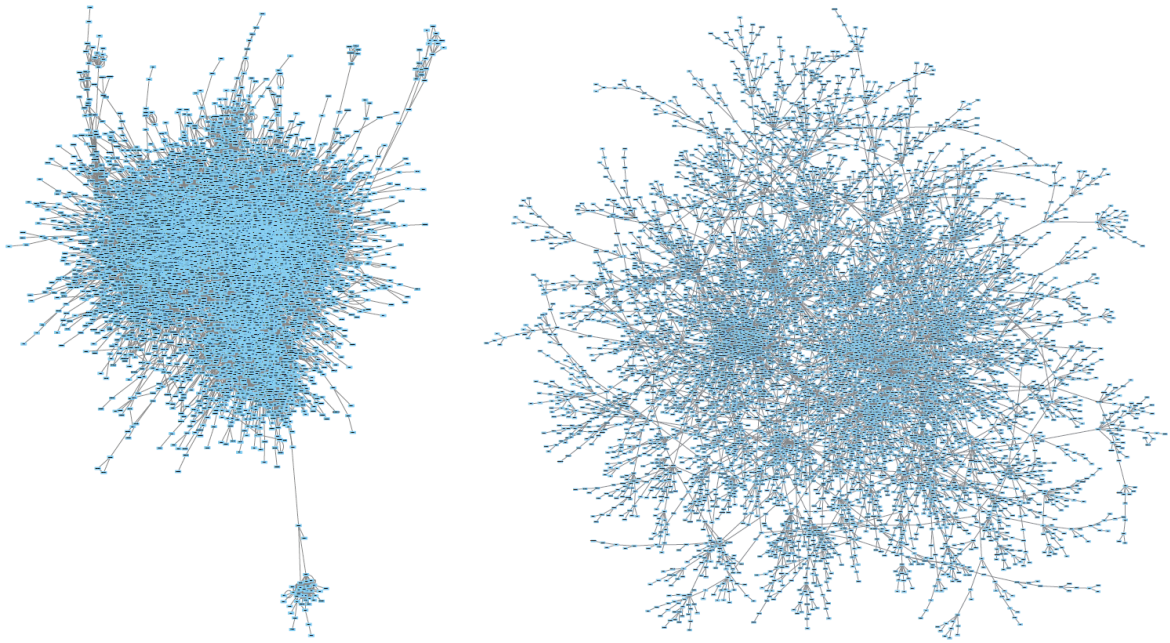
(v) **The node degree distribution**: It indicates the ratio of nodes in the network with degree k. Scale-free networks usually follow a power law: $P(k) \sim k^{\gamma}$, where γ is a constant (≥ 0). A high γ is an indicator of a scale-free topology.

For this experiment, the networks obtained by Kendall's method on Yeast and Human datasets have been used as reference, for the sake of simplicity. Thus, we present a topological study for four networks, the originals (named "$Input_{organism}$") and the processed ones (hereafter "$GeSOp_{organism}$"). Visual representation of the networks is depicted in Figures 5 and 6, where it is possible to check the topological differences of the networks.

As can be seen in the figures, the optimized networks ("$GeSOp_x$") present a more linear and less compact topology than the input ones, so they fit better with the scale-free topology. In addition, an exhaustive topological analysis of the four networks has been carried out based on the indicators presented above. The topological analysis of the network has been performed using the tool Network Analyzer [37] and the results obtained are depicted in Table 7.

The results presented in Table 7 show that the networks improve their topological indicators once they are processed by GeSOp. Moreover, it is possible to argue that these networks follow a biological pattern according to [36]. That is, after the optimization process, networks show, on the one hand, a lower mean clustering coefficient and density. On the other hand, they present higher characteristic path length, diameter, and γ constant. These results mean that networks have improved in terms of the biological relevance of their relationships.

Moreover, the optimized networks present characteristics closer to a scale-free topology as their node degree distribution follows a power law with $\gamma \geq 0$[34] (see Figure 7).



(a) $Input_h$ (b) $GeSOp_h$

Figure 6: Visual comparison of human networks used in this experiment. The original Kendall's network is shown on (a). On (b), the optimized network obtained with GeSOp is depicted.

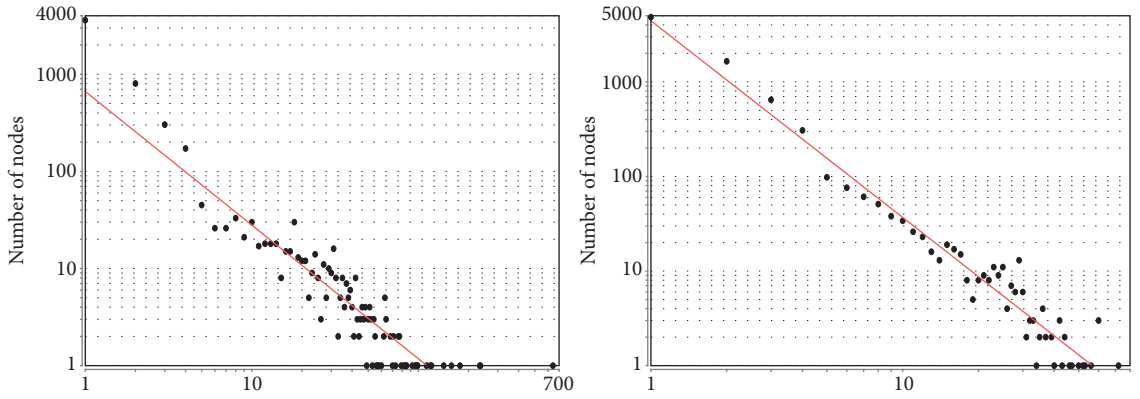


Figure 7: Node degree for the optimized networks obtained with GeSOp. The fitted power law indicates that the networks follow a scale-free topology.

Table 7: Topological indicator of four selected networks. The results presented show how the optimized networks obtained by GeSOp improve their indicators.

	Network	Clust. coef.	CPL	Diameter	Density	Gamma (γ)
Yeast	$Input_y$	0.411	2.697	9	0.041	0.845
	$GeSOp_y$	0.085	6.156	20	0.001	1.375
Human	$Input_h$	0.21	4.954	19	0.003	1.394
	$GeSOp_h$	0.024	10.84	33	~ 0.000	2.079

This fact can be verified by the results presented in column "Gamma" of Table 7, in which the values of γ (from power law) are improved in the optimized networks.

The results generated by this second experiment probes that GeSOp is a reliable method to improve the topological features of the gene networks, in terms of biological structure.

4. Conclusions

In this work, a new backward elimination method for optimization of large gene networks structure, namely, GeSOp, has been presented. The method, which is based on a greedy strategy, is able to perform a drastic reduction of size of the input network in terms of the number of gene-gene relationships. The prune of the less biologically significant relationships produces simpler and more user-friendly networks for researchers in terms of size and visualization.

On one hand, the results presented show that the method is able not only to perform a prune of the input network, but also to keep the ratio of the biological information presented in the original network. Furthermore, for almost all studied cases, this ratio is improved. On the other hand, topological analyses carried out in the experiments show how networks optimized by GeSOp improve their biological indicators by acquiring a scale-free topology. Finally, regarding the generated results, it is possible to argue that the relevance of our method becomes evident for the processing and optimization of large gene networks.

As future works, we will work on the inclusion of previous biological knowledge, in form of gene networks as gold standard, in the second step of the methodology. Thus, the method will take into account not only the existing Hubs in the input network, but also the genes that have a great relevance in the networks used as gold standard. Another future work is based on the implementation; we are working in paralleling implementation of the algorithm to improve its performance.

Data Availability

In this section, we provide the links to the datasets and databases presented above. In particular, the links for the datasets are as follows:

(1) **Yeast dataset**: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE23

(2) **Human dataset**: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL570

and those for the databases are as follows:

(1) **GeneMANIA**: http://genemania.org/data/

(2) **YeastNet**: https://www.inetbio.org/yeastnet/

(3) **HumanNet**: http://www.functionalnet.org/humannet/

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] Y. X. R. Wang and H. Huang, "Review on statistical methods for gene network reconstruction using expression data," *Journal of Theoretical Biology*, vol. 362, pp. 53–61, 2014.

[2] M. Hecker, S. Lambeck, S. Toepfer, E. van Someren, and R. Guthke, "Gene regulatory network inference: data integration in dynamic models—a review," *BioSystems*, vol. 96, no. 1, pp. 86–103, 2009.

[3] F. Gómez-Vela, C. D. Barranco, and N. Díaz-Díaz, "Incorporating biological knowledge for construction of fuzzy networks of gene associations," *Applied Soft Computing*, vol. 42, pp. 144–155, 2016.

[4] D. Marbach, R. J. Prill, T. Schaffter, C. Mattiussi, D. Floreano, and G. Stolovitzky, "Revealing strengths and weaknesses of methods for gene network inference," *Proceedings of the National Acadamy of Sciences of the United States of America*, vol. 107, no. 14, pp. 6286–6291, 2010.

[5] A. Lachmann, F. M. Giorgi, G. Lopez, and A. Califano, "ARACNe-AP: gene network reverse engineering through adaptive partitioning inference of mutual information," *Bioinformatics*, vol. 32, no. 14, pp. 2233–2235, 2016.

[6] N. Omranian, J. M. O. Eloundou-Mbebi, B. Mueller-Roeber, and Z. Nikoloski, "Gene regulatory network inference using fused LASSO on multiple data sets," *Scientific Reports*, vol. 6, Article ID 20533, 2016.

[7] F. Petralia, P. Wang, J. Yang, and Z. Tu, "Integrative random forest for gene regulatory network inference," *Bioinformatics*, vol. 31, no. 12, pp. i197–i205, 2015.

[8] H. Yu, B. Jiao, L. Lu et al., "NetMiner-an ensemble pipeline for building genome-wide and high-quality gene co-expression network using massive-scale RNA-seq samples," *PLoS ONE*, vol. 13, no. 2, p. e0192613, 2018.

[9] W. L. Poehlman, M. Rynge, D. Balamurugan, N. Mills, and F. A. Feltus, "OSG-KINC: high-throughput gene co-expression network construction using the open science grid," in *Proceedings of the 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1827–1831, Kansas City, MO, November 2017.

[10] J. Xia, E. E. Gill, and R. E. W. Hancock, "NetworkAnalyst for statistical, visual and network-based meta-analysis of gene expression data," *Nature Protocols*, vol. 10, no. 6, pp. 823–844, 2015.

[11] A. Barabási and Z. N. Oltvai, "Network biology: understanding the cell's functional organization," *Nature Reviews Genetics*, vol. 5, no. 2, pp. 101–113, 2004.

[12] R. R. Vallabhajosyula, D. Chakravarti, S. Lutfeali, A. Ray, and A. Raval, "Identifying Hubs in protein interaction networks," *PLoS ONE*, vol. 4, no. 4, Article ID e5344, 2009.

[13] Y. Wang, X. Zhang, and L. Chen, "Optimization meets systems biology," *BMC Systems Biology*, vol. 4, no. Suppl 2, p. S1, 2010.

[14] S. A. Thomas and Y. Jin, "Reconstructing biological gene regulatory networks: where optimization meets big data," *Evolutionary Intelligence*, vol. 7, no. 1, pp. 29–47, 2014.

[15] M. R. Mendoza and A. L. Bazzan, "Evolving random boolean networks with genetic algorithms for regulatory networks reconstruction," in *Proceedings of the the 13th annual conference*, p. 291, Dublin, Ireland, July 2011.

[16] J. Liu, Y. Chi, and C. Zhu, "A dynamic multiagent genetic algorithm for gene regulatory network reconstruction based on fuzzy cognitive maps," *IEEE Transactions on Fuzzy Systems*, vol. 24, no. 2, pp. 419–431, 2016.

[17] J. Xiong and T. Zhou, "Gene regulatory network inference from multifactorial perturbation data using both regression and correlation analyses," *PLoS ONE*, vol. 7, no. 9, Article ID e43819, 2012.

[18] J. Li and X.-S. Zhang, "An optimization model for gene regulatory network reconstruction with known biological information," in *Proceedings of the First International Symposium on Optimization and Systems Biology*, pp. 35–44, 2007.

[19] M. E. Studham, A. Tjärnberg, T. E. M. Nordling, S. Nelander, and E. L. L. Sonnhammer, "Functional association networks as priors for gene regulatory network inference," *Bioinformatics*, vol. 30, no. 12, pp. I130–I138, 2014.

[20] F. M. Lopes, D. C. Martins Jr., J. Barrera, and R. M. Cesar Jr., "A feature selection technique for inference of graphs from their known topological properties: revealing scale-free gene regulatory networks," *Information Sciences*, vol. 272, pp. 1–15, 2014.

[21] B. Yang, J. Xu, B. Liu, and Z. Wu, "Inferring gene regulatory networks with a scale-free property based informative prior," in *Proceedings of the 8th International Conference on BioMedical Engineering and Informatics (BMEI '15)*, pp. 542–547, October 2015.

[22] D. B. West, *Introduction to Graph Theory*, Prentice-Hall of India Private Limited, New Delhi, India, 2000.

[23] P. T. Spellman, G. Sherlock, M. Q. Zhang et al., "Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization," *Molecular Biology of the Cell (MBoC)*, vol. 9, no. 12, pp. 3273–3297, 1998.

[24] Y. Hodo, M. Honda, A. Tanaka et al., "Association of interleukin-28B genotype and hepatocellular carcinoma recurrence in patients with chronic hepatitis C," *Clinical Cancer Research*, vol. 19, no. 7, pp. 1827–1837, 2013.

[25] P. A. Jaskowiak, R. J. G. B. Campello, and I. G. Costa, "On the selection of appropriate distances for gene expression data clustering," *BMC Bioinformatics*, vol. 15, article no. S2, 2014.

[26] L. Song, P. Langfelder, and S. Horvath, "Comparison of co-expression measures: mutual information, correlation, and model based indices," *BMC Bioinformatics*, vol. 13, no. 1, article no. 328, 2012.

[27] H. Liu, L. Liu, and H. Zhang, "Ensemble gene selection for cancer classification," *Pattern Recognition*, vol. 43, no. 8, pp. 2763–2772, 2010.

[28] D. W. Farley, S. L. Donaldson, O. Comes et al., "The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function," *Nucleic Acids Research*, vol. 38, no. 2, pp. W214–W220, 2010.

[29] H. Kim, J. Shin, E. Kim et al., "YeastNet v3: a public database of data-specific and integrated functional gene networks for Saccharomyces cerevisiae," *Nucleic Acids Research*, vol. 42, no. 1, pp. D731–D736, 2014.

[30] J. M. Cherry, E. L. Hong, and C. Amundsen, "Saccharomyces genome database: the genomics resource of budding yeast," *Nucleic Acids Research*, pp. D700–D705, 2012.

[31] I. Lee, U. M. Blom, P. I. Wang, J. E. Shim, and E. M. Marcotte, "Prioritizing candidate disease genes by network-based boosting of genome-wide association data," *Genome Research*, vol. 21, no. 7, pp. 1109–1121, 2011.

[32] E. R. Dougherty, "Validation of inference procedures for gene regulatory networks," *Current Genomics*, vol. 8, no. 6, pp. 351–359, 2007.

[33] D. M. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," *International Journal of Machine Learning Technology*, vol. 2, no. 1, pp. 37–63, 2011.

[34] N. T. Doncheva, Y. Assenov, F. S. Domingues, and M. Albrecht, "Topological analysis and interactive visualization of biological networks and protein structures," *Nature Protocols*, vol. 7, no. 4, pp. 670–685, 2012.

[35] G. A. Pavlopoulos, M. Secrier, C. N. Moschopoulos et al., "Using graph theory to analyze biological networks," *BioData Mining*, vol. 4, no. 1, article 10, 2011.

[36] W. Winterbach, P. V. Mieghem, M. Reinders, H. Wang, and D. D. Ridder, "Topology of molecular interaction networks," *BMC Systems Biology*, vol. 7, article no. 90, 2013.

[37] Y. Assenov, F. Ramírez, S.-E. Schelhorn, T. Lengauer, and M. Albrecht, "Computing topological parameters of biological networks," *Bioinformatics*, vol. 24, no. 2, pp. 282–284, 2008.