INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
Online ISSN 1841-9844, ISSN-L 1841-9836, Volume: 15, Issue: 6, Month: December, Year: 2020
Article Number: 3894, https://doi.org/10.15837/ijccc.2020.6.3894

A Rough Set and Cellular Genetic Fusion Algorithm for Acute Critical Disease Prediction

L. Jia, H. Wang, H. Zhuang, X. Li, Y. Zhao, S. Pan, K. Wu, J. Li, T. Li

Lijing Jia, Heng Zhuang, Yuzhuo Zhao, Tanshi Li
Department of Emergency,
Chinese PLA General Hospital, Beijing, China

Hongxin Wang
Department of Emergency,
Armed Police Characteristic Medical Center, Tianjin, China

Xueyan Li
Management School
Beijing Union University, Beijing, China

Kainan Wu
Department of School of Economics and Management
University of Chinese Academy of Sciences, Beijing, China

Shuxiao Pan, Jing Li*
Department of School of Economics and Management
Beijing Jiaotong University, Beijing, China
*Corresponding author:jingli@bjtu.edu.cn

Abstract

This study is to solve the problems of an overly-broad scale of medical indicators, lack of retrospective research samples, insufficient depth of data mining, and low disease prediction accuracy. In this paper, we propose an intelligent screening algorithm that combines a genetic algorithm, cellular automata, and rough set theory. This algorithm can achieve high accuracy in predicting patient outcomes with a small number of indicators. And we compare it with the traditional genetic algorithm. We built the prediction model with 64 indicators based on the logistic regression (AUC 0.8628), support vector machine (AUC 0.5319), Naïve Bayes (AUC 0.7102), and AdaBoost algorithms (AUC 0.9095). Using the cellular genetic algorithm for attribute screening not only effectively reduces the number of indicators but also achieve almost the same accuracy of prediction with 8 indicators based on the logistic regression (AUC 0.8782), support vector machine (AUC 0.8525), Naïve Bayes (AUC 0.8408), and AdaBoost algorithms (AUC 0.8770). Compared with the traditional scoring system, the predictive model established in this paper can more accurately predict rebleeding accidents based on physiological test indicators and continuous patient indicators.

Keywords: cellular genetic algorithm, key indicator, disease prediction, machine learning.

1 Introduction

In the modern world, big data technology is widely used in medical research. Driven by medical informatization, hundreds of millions of pieces of medical data are generated every day. Big data technology and the first aid database now provide technical and statistical support for studying the key indicators associated with patients with acute critical disease. The ability to extract key indicators from massive data in light of unknown causes and quickly provide pre-warning about the risk of rebleeding accidents is crucial to improving the survival rate of patients with acute critical disease. Furthermore, the pre-warning model can provide enough time for medical staff to take measures and allocate treatment resources.

Due to financial, time, and human limitations, the sample size of a traditional retrospective medical analysis is small. The indicators obtained from tracked patients are limited, and researchers often face the problem of missing data. Once the number of patients and the index dimension increase greatly, the workload of traditional retrospective medical analysis will increase exponentially. It is difficult to provide doctors with all-around and multi-angle decision support [46]. Currently, it is necessary to consider advanced theories and methods to balance the number of indicators and the accuracy of disease prediction. In order to solve this problem from the perspective of evolutionary computation, this study proposes a screening method combining rough set, genetic algorithm, and cellular automata that can effectively balance the number of indicators and the accuracy of disease prediction, and compares it with existing methods.

The primary outcome of this study is that cellular genetic algorithm was used to construct more accurate and effective prediction models based on the logistic regression (AUC 0.8782), support vector machine (AUC 0.8525), Naïve Bayes (AUC 0.8408), and AdaBoost algorithms (AUC 0.8770).

2 Literature review

In the field of emergency critical illness, most data sources are prospective analyses or retrospective analyses [49]. The amount of data is limited. Although most of the analyses use statistical tests to process the data [20, 30], the sample size is small and the mining depth is insufficient. The conclusions are mostly from the perspective of treatment and evaluation, which has limited value for prevention and prediction [19, 23].

With the development of medical big data, there is a new way to resolve the huge medical index system. Hu first used multivariate logistic regression analysis to determine independent predictors when predicting the need for paediatric emergency patient return visits [17]. Wei et al. used logistic regression, AdaBoost, and XGBoost to form predictions based on 728 patients in the emergency database of PLA General Hospital, and finally selected the XGBoost method to screen key indicators [44].

It can be seen from the above research that compared with traditional methods, machine learning has achieved better results in outcome prediction and risk warning systems for critical patients. In the face of a huge medical index system, it is better able to identify the index that has high correlation with the prediction under the condition of ensuring prediction accuracy.

Key indicator selection, also known as feature selection or attribute selection, is the process of selecting some effective features from the original features to reduce the dataset dimensions, and is the key pre-processing step in machine learning and data mining [50]. Scholars worldwide have done much research on the design of screening for key indicators. There are three common methods: filter, embedding, and wrapper [24]. The attribute reduction method based on rough set is a typical filtering method [51]. A filtering method divides the key index screening and the subsequent prediction model training into two independent parts. In the face of a large amount of data and multi-dimensional indexes, this method is simple and easy to use, and has a good interpretability for the data without consuming a lot of calculation time.

The rough set (RS) theory was proposed by Pawlak in the early 1980s [28]. This theory provided a new mathematical tool for dealing with fuzzy and uncertain problems [25]. The major advantages of rough set theory are that it can efficiently discover implied knowledge by analysing imperfect

(uncertain and vague) [27], inconsistent [29], and incomplete data [11], without the use of priori information or assumptions [31]. The main theoretical difference between rough set theory and the most commonly used statistical method, regression, is that rough set theory as a classification tool is able to predict the "belongingness" of items to discrete classes, which in turn can be used to generate decision rules from the results, whereas regression tends to predict a value from continuous variables to obtain a mathematical formula for the result. In addition, rough set theory does not require any check on assumptions; instead, it focuses more on the accuracy of the model's predictions. The medical field often involves many subdivided medical indicators, and rough set theory is very suitable for solving large medical index analysis and prediction problems. Ali et al. (2015) proposed a new hybrid rough set model to analyse 17 indicators of 50 diabetic patients in order to distinguish the type of diabetes and predict disease progression [2]. Gil-Herrera et al. (2011) established a dataset consisting of 9105 cases and 15 variables, and used rough set theory to predict the late lifespan of terminally-ill patients in order to improve the referral process for hospice care [10]. Wang et al. (2006) introduced particle swarm optimization into rough set theory to predict the degree of malignancy of gliomas. Additionally, 14 conditional attributes and one decision attribute were extracted from 280 cases, revealing the relationship between the features found on the MRI and the degree of malignancy [43].

However, in traditional rough set methods, the decision attributes have mostly consisted of single sequences. Decision-making methods that combine multiple attributions are often used. It will affect the speed of attribute reduction. Because of its sensitivity to noise, the decision rules for extraction are very unstable, and the accuracy needs to be improved. The genetic algorithm was created in 1975 by Holland and his students from the University of Michigan [16]. It has a natural implicit parallelism and powerful global search ability, and obtains the global optimal solution for solution space by simulating the genetic evolution principle of bio-property survival.

However, the traditional genetic algorithm has disadvantages including low search efficiency, poor local search ability, and ease of falling into local optima when solving problems. In 1948, Von Neumann proposed the idea of cellular automata [40], which is based on the characteristics of complex systems, to simulate and describe complexity. In 1993, Whitley first proposed the cellular genetic algorithm (CGA) [45], which combines genetic algorithms and cellular automata to find global optimal solutions for complex systems. The core idea is realizing the wide spread of excellent individual information among the population through the information interaction among multiple subjects. The CGA demonstrates excellent performance in local extremum. Using cellular automata (CA) [39], the algorithm assigns the individuals comprising a group to a two-dimensional grid (cell space). The genetic operation of the algorithm is different from that of a simple GA. The selection operation of a simple GA is randomly paired in the whole population, while the CGA avoids too much information exchange, confines the individual's genetic selection operation to a limited neighborhood, completes the selection and recombination in the neighborhood, performs mutations, and leaves good individuals as offspring and fathers. The limited neighborhood slows down the diffusion of individual genetic information throughout the population, which is conducive to maintaining the diversity of the population during evolution, and provides conditions for avoiding local convergence and responding to changes in a timely manner [1][8]. Therefore, this study uses a CGA to reduce rough set attributes and screen key indictors that have high correlation with acute critical disease. The combination of rough sets and CGA is still relatively rare.

In summary, the existing reduction algorithm mainly uses a heuristic search method to construct the minimum reduction of the conditional attributes, that is, the minimum reduction, from the kernel of the rough set. However, this algorithm becomes increasingly complicated as the scale of the problem increases, and it is difficult to find the global optimum. The genetic algorithm is very suitable for solving the rough set attribute reduction problem because it has the advantages of global optimization and implicit parallelism. The neighbour learning model of the CGA also better maintains the diversity of the population, which offers a good balance between global search and local optimization.

According to the characteristics of rough set, genetic algorithms and cellular automata, this paper proposes a new intelligent screening algorithm for the prediction of key indicators in patients with acute critical disease. The algorithm introduces a combination of a genetic algorithm and cellular au-

tomata into rough set theory and is, in essence, an evolutionary algorithm based on genetic algorithms. Evolutionary computation is a subdomain of computational intelligence that involves combinatorial optimization problems. The algorithm is influenced by the natural selection mechanism of the "survival of the fittest" and the transmission of genetic information in the process of biological evolution. Through the iterative simulation of this process, the problem to be solved is regarded as the environment, and in some possible solutions, the optimal solution is sought through natural evolution [21]. One of its important research branches is the genetic algorithm [15]. Compared with traditional optimization methods, such as statistics-based methods and exhaustive methods, evolutionary computation is a mature global optimization method with high robustness and wide applicability. It has the characteristics of self-organization, self-adaptation, and self-learning [9], and can effectively address complex problems (such as non-deterministic polynomial -hard optimization problems) that are difficult to solve with traditional optimization algorithms without being limited by the nature of the problem.

In the context of current medical big data, combined with the idea of evolutionary computing, the proposed intelligent screening algorithm can effectively compensate for the shortcomings of existing statistical testing methods, such as insufficient computational power and difficulty in finding optimal solutions. It is possible to extract the research objects that satisfy the conditions from the existing datasets as well as efficiently and stably calculate the key indicators that affect rebleeding in patients with acute critical disease. The algorithm can provide more timely, efficient, and scientific information for clinical treatment decisions, and solve clinical practical problems using clinical "real world" data.

3 Methods

In this study, the screening of key patient indicators is regarded as a single objective attribute reduction problem: taking patients' physiological indicators as condition attributes and patients' conditions as decision attributes. The optimization problem is solved by the combination of rough set, genetic algorithm, and cellular automata, and the fitness function is constructed by the grey correlation degree. The optimal key index combination is generated by the iterative updating of the population.

3.1 Preliminaries

The disaster knowledge system can be represented as a four-tuple:

The knowledge system of disease can be expressed as a four-tuple: $S = \{U, A, R, D\}$, where F is the attribute value of the object.

$U = \{x_1, x_2, \cdots, x_n\}$ is a set of patients;

$A = (a_1, a_2, \cdots, a_m)$ represents the physiological index of patients;

$D = (d_1, d_2, \cdots, d_d)$ represents the option of patients' conditions;

The parameter p of $D(p)$ indicates the degree of association between the decision attribute (column) and the condition attribute (column), which can be calculated by the grey correlation degree $p_i = \frac{1}{m}\sum_{i=1}^{m}\gamma_i$.

D is the decision attribute of the disease knowledge system. For example, when the observed data for patient k are $\mathrm{D}(k)$, $k = 1, \cdots, n$; then, $D(k) = \{d(1), d(2), \cdots, d(n)\}$ is the decision attribute sequence of the disease knowledge system.

R is the conditional attribute of the disease knowledge system. When the physiological parameter observation data for patient k are $R_i(k)$, $i = 1, \cdots, m$, $k = 1, \cdots, n$; then, $R_i(k) = \{a_i(1), a_i(2), \cdots, a_i(n)\}$ is the conditional attribute sequence of the disease knowledge system.

Dependency reflects the relationship between attributes. If an attribute is regarded as a type of knowledge reflecting the object, then the attribute dependence can be regarded as the ability to derive knowledge from other knowledge, and is a measure of knowledge dependence. Knowledge Q is derivable from knowledge P when all concepts of Q can be defined by some of the concepts in P. When Q is derivable from knowledge P, it is stated that Q depends on P, denoted by $P \Rightarrow Q$. The formal definition of dependency is as follows:

Let $K = (U, R)$ be a knowledge base, and let $P,\ Q \subseteq R$:

(1) When IND $(P) \subseteq$ IND(Q), knowledge Q depends on knowledge P;

(2) When $P \Rightarrow Q$ and $Q \Rightarrow P$, knowledge P and Q are equivalent, denoted by $P \equiv Q$;

(3) When there is no $P \Rightarrow Q$ and there is no $Q \Rightarrow P$, P and Q are independent.

Obviously, if and only if IND $(P) \subseteq$ IND(Q), then $P \Rightarrow Q$. Let $K = (U,\ R)$ be a knowledge base, and let $P,\ Q \subseteq R$;

when $K = r\,(P,\ Q) = r_p\,(Q) = |\mathrm{POS}_p(Q)|\,/|\mathrm{card}(U)|$, knowledge Q depends on P $(0 \leq k \leq 1)$, denoted by $P \Rightarrow_k Q$, where card represents the number of sets. $\mathrm{POS}_p\,(Q) = \bigcup_{x \in U/Q} P(x)$ is the P positive domain of the domain U in Q. Therefore, $K = r\,(P,\ Q) = \sum_{x \in U/Q} \frac{|P(x)|}{U}$.

Let $X_i(k)$ be the attribute sequence, $X_i\,(k) = \{D\,(k)\,, R_i(k)\}$. E_1, E_2 are the sequence operators.

$X_iE_1 = \{x_i\,(1)\,e_1, x_i\,(2)\,e_1, \cdots, x_i\,(n)\,e_1\}$, $X_iE_2 = \{x_i\,(1)\,e_2, x_i\,(2)\,e_2, \cdots, x_i\,(n)\,e_2$, where $x_i\,(k)\,e_1 = x_i(k)/x_i(1)$ and X_iE_1 is the initial value image. $x_i\,(k)\,e_2 = x_i(k)/\overline{x_i}$, and X_iE_2 is the mean image.

In the knowledge system of gastrointestinal bleeding, $S = \{U, A, F, D\}$, $R \subseteq A$, $U = \{x_1, x_2, \ldots, x_n\}$, the rough membership of the disaster set relative to the decision attribute i is expressed as:

$$\mu_U\ (D_i\,) = \frac{card(U \cap D_i)}{\mathrm{card}\,(U)}. \tag{1}$$

The concept of 'rough membership' indicates the dependence between the physiological parameter attribute set and the decision attribute set. That is, in a set, the higher the frequency of a certain decision attribute, the greater the importance of this attribute for this set.

$p_i = \frac{1}{m}\sum_{i=1}^{m} \gamma_i$, p_i is the grey correlation degree of the condition attribute and decision attribute [21], indicating the degree of association between gastrointestinal rebleeding and patient physiological index. This paper uses Deng's grey relational degree. $D_i\,(k) = \{\mathrm{di}\,(1)\,, \mathrm{di}\,(2)\,, \cdots, \mathrm{di}(n)\}$ is the decision attribute sequence of the disease knowledge system. $R_i\,(k) = \{a_i\,(1)\,, a_i\,(2)\,, \cdots, a_i(n)\}$ is the conditional attribute sequence of the disease knowledge system. γ_i is the correlation coefficient between $d_i\,(k)$ and $a_i\,(k)$. The formula is as follows:

$$\gamma_{\mathrm{ij}}(k) = \frac{\min\limits_{j}\min\limits_{k} |D_j\,(k) - R_i\,(k)| + \rho \max\limits_{j}\max\limits_{k} |D_j\,(k) - R_i\,(k)\,|}{|D_j\,(k) - R_i\,(k)| + \rho \max\limits_{j}\max\limits_{k} |D_j\,(k) - R_i\,(k)\,|}, \tag{2}$$

where $|D_j\,(k) - R_i\,(k)\,|$ is the absolute difference between the points $D_j\,(k)$ and $R_i\,(k)$. ρ is the identification coefficient, generally 0.5; the degree of correlation between the decision attribute $D_j\,(k) = \{\mathrm{dj}\,(1)\,, \mathrm{dj}\,(2)\,, \cdots, \mathrm{dj}(n)\}$ and the condition attribute $R_i\,(k) = \{a_i\,(1)\,, a_i\,(2)\,, \cdots, a_i(n)\}$ is

$$\gamma\,(D_j, R_i) = \frac{1}{n}\sum_{k=1}^{n} \gamma_{\mathrm{ij}}(k). \tag{3}$$

Set a real number $\gamma\,(D_j, R_i)$, if

(1) $0 < \gamma\,(D_j, R_i) < 1, \gamma\,(D_j, R_i) = 1 \Longleftrightarrow\ D_j\ = R_i$

(2) The smaller $|D_j\,(k) - R_i\,(k)\,|$ is, the larger $\gamma\,(D_j, R_i)$ is.

Let $\gamma\,(D_j, R_i)$ be the grey relational degree of decision attribute Dj with condition attribute Ri.

The above formula (1) organically combines the grey correlation degree and the traditional definition of dependence between different attribute sets. By obtaining the physiological indicators of the patients during actual clinical diagnosis and treatment, a set of key influencing factor rules are extracted to detect early gastrointestinal rebleeding risk.

3.2 Design of an intelligent screening algorithm

Screening for key physiological indicators in this paper is achieved through attribute reduction. The so-called attribute reduction refers to reducing the amount of redundant knowledge (attributes) in the knowledge base without affecting knowledge expression, thus ensuring that the information system's classification ability remains unchanged. This procedure is implemented to make the expression of the knowledge base more concise, and thus, to finally extract the rules of the knowledge system. The attribute reduction problem belongs to a discrete coding optimization problem. When the number of variables is large, the problem is an NP-hard problem. Studies have shown that some artificial intelligence algorithms, such as genetic algorithms, have achieved good results in solving NP-hard problems that result from incomplete information [15, 21].

In recent years, some studies have shown that compared with traditional genetic algorithms, the CGA can better maintain population diversity and has a strong global search ability for complex optimization problems [18]. When facing multi-dimensional, large-scale medical big data, traditional genetic algorithms are prone to premature convergence due to poor local search ability, while CGAs can effectively retain good individuals and maintain population diversity. Therefore, we selected a CGA to optimize the rough set attribute reduction problem.

It is known from section 3.1 that in the decision table, the importance of a condition attribute to the decision attribute has a one-to-one relation. The principle of the attribute reduction algorithm is observing changes in the decision table after removing an attribute, and then measuring how important that attribute is to the decision attribute. The greater the change in the classification ability upon removing a condition attribute from the decision table relative to the decision attribute, the more important that condition attribute is. Based on this principle, this paper designs a new coding method using a CGA.

(1) Encoding method

Let $A = (a_1, a_2, \cdots, a_m)$ represent the set of condition attributes in the decision table of clinical diagnosis and treatment of diseases. Let $a_i = 0$ indicate that the condition attribute can be reduced; let $a_i = 1$ indicate that the condition attribute cannot be reduced.

(2) Fitness function

The fitness function uses the dependence degree shown in (1). The greater the dependency between the condition attribute and the decision attribute, the more important the condition attribute is, that is, the greater the individual's fitness.

(3) The algorithm steps are as follows (See Figure 1).

Step 1: Generate an initial population.

In the $n \times n$ cell space, n^2 condition attribute combinations are randomly generated, represented by x_{ij}, where $i, j \in [1, n]$. Let K denote the number of condition attributes; then, $x_{\mathrm{ij}} = [a_{\mathrm{ij1}}, \cdots, a_{\mathrm{ijk}}, \cdots, a_{\mathrm{ijK}}]$, and a_{ijk} randomly takes 0 or 1. Zero means that the attribute a_{ijk} is not included in the individual x_{ij}, and 1 means that the attribute a_{ijk} is included in the individual x_{ij}.

Step 2: Calculate the fitness.

Each cell individual calculates its own condition attribute dependency, denoted by y_{ij}. Standardize it for comparison. Let $\mathrm{fit}_{\mathrm{ij}} = y_{\mathrm{ij}} - \min\limits_{i,j\in[1,n]} y_{\mathrm{ij}}, \mathrm{fit}_{\mathrm{ij}}$ be the fitness of the individual i, j.

Step 3: Selection.

The Moore-type neighbour structure ($[i-1, i+1]$,$[j-1, j+1]$) is used, which is represented by Ω. Each individual x_{ij} searches for the optimal individual $\left\{ x_{\mathrm{ef}} \middle| y_{\mathrm{ef}} = \max\limits_{i,j\in\Omega} y_{\mathrm{ij}} \right\}$ in the "neighbour" as the learning object.

Step 4: Recombination.

Set the crossover probability to p_c. For each attribute a_{ijk} in x_{ij}, the probability of being interchanged with the attribute a_{efk} in x_{ef} is p_c.

Step 5: Mutation.

Set the mutation probability to p_m. For each attribute a_{ijk} in x_{ij}, the probability of generating a variation (1 to 0, or 0 to 1) is p_m.

Step 6: Return to Step 2 until the fitness is no longer increasing.

The pseudo code of the above process is given as follows:

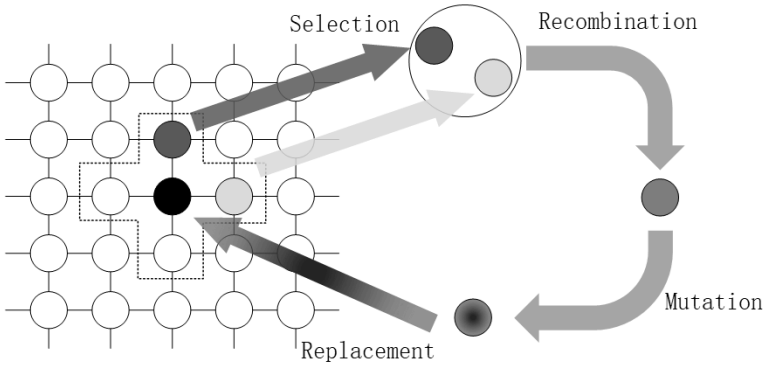

Figure 1: Cellular genetic algorithm

Algorithm	Cellular Genetic Algorithm-based attribute reduction
1	While not Termination Condition() do
2	For each cell i from 1 to population size
3	The attribute combination of cell i is coded by binary code
4	End
5	For each cell i from 1 to population size
6	Calculate the attribute dependability of cell i's attribute combination
7	Establish the fitness function based on dependability
8	Select the cell with the best fitness from the neighbourhood i^*
9	Let pc and pm be the crossover and mutation probability respectively
10	If rand $<$ pc
11	Implement the crossover operation between i and i^*
12	End
13	If rand $<$ pm
14	Implement the mutation operation in cell i
15	End
16	End
17	End

4 Results and validation

4.1 Indexes, data pre-processing, and algorithm parameters

In this paper, 647 cases diagnosed as gastrointestinal bleeding in the first aid database of PLA General Hospital were used to build the decision-making table for gastrointestinal bleeding. Among them, there were 313 cases of nosocomial rebleeding in the experimental group, and 334 cases of non-nosocomial rebleeding in the control group. The 64 vital signs and laboratory indexes (as shown in Appendix A1) that are most likely to be related to nosocomial rebleeding of the gastrointestinal tract were selected as condition attributes, and each patient was marked with the outcome state (whether there was a rebleeding condition) as decision attributes, and the original decision table is shown in Appendix A2. For the blank values in the original data, this paper uses the method of filling column by column. To meet the rough set data requirements, only discrete data can be processed by rough set. According to clinical experience, the clinician gave the discrete standard of each index (See Appendix A3). After data dispersion and cleaning, the final decision table for gastrointestinal bleeding was obtained (See Appendix A4)

If there was no special description, the calculation environment and algorithm parameters were set according to Table 1 and Table 2.

Table 1: Computing environment

Name	Configuration
CPU	E5-1650 v2
Graphics card	GTX1060 6G ASUS
RAM	Kingston 8G Recc
Motherboard	ASUS Z9PA
Hard disk	Samsung 850 EVO 250G SSD

Table 2: Algorithm parameter setting

Algorithm parameter	Value
Number of populations	50
Probability of intersection	0.7
Probability of variation	0.01

4.2 Key index screening algorithm and performance analysis

In order to verify the performance of the intelligent screening algorithm built in this paper in an all-around manner, the condition attributes with fewer times of elimination were retained after 50 independent repeated attribute reduction experiments using the genetic algorithm and the CGA, respectively. The results are shown in Table 3 and Table 4.

Table 3: Attribute reduction result based on the cell genetic algorithm

Condition attribute number	Physiological indicators
25	Thrombin Time
35	Inorganic Phosphate
36	Haemoglobinometry
50	Glu (blood gas analysis)
57	pH
59	Systolic Blood Pressure
60	Diastolic Blood Pressure
64	Heart Rate

Table 4: Attribute reduction result based on the genetic algorithm

Condition attribute number	Physiological indicators
2	Gamma-Glutamyl Transpeptidase
3	Leukocyte Count
5	Monocyte
23	N-terminal Pro-brain Natriuretic Peptide
30	Glucose
33	Eosinophil
36	Haemoglobinometry
60	Diastolic Blood Pressure

4.2.1 Algorithm performance

The performance of an algorithm is usually examined by its ability to find the global optimum and the convergence rate. The ability to find the global optimum refers to the exploration capability and the convergence quality; the convergence speed refers to the number of algorithm iterations required to calculate the optimal solution. The genetic algorithm and the CGA were used to perform 50 independent repeated attribute reduction experiments on the decision table data. The adaptation

process of one of the reduction algorithms is shown in Figure 2, and a box diagram of 50 independent reductions is shown in Figure 3.

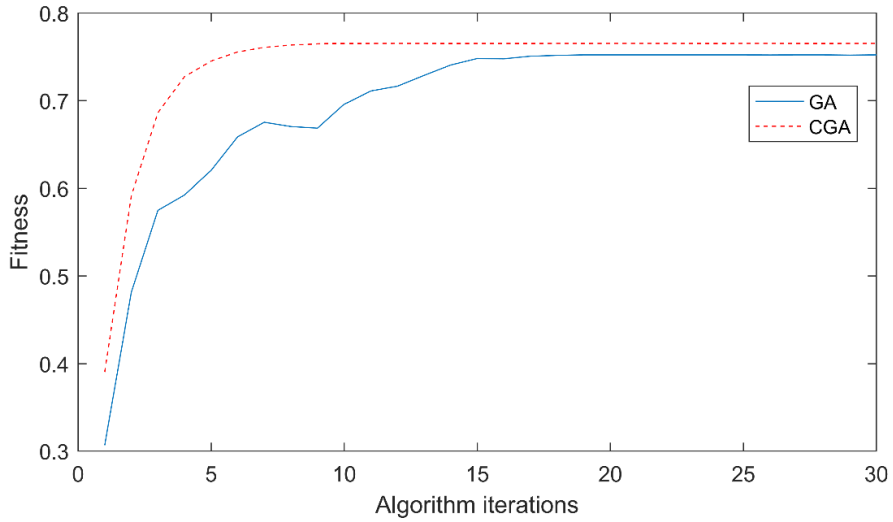

Figure 2: The fitness evolution process of the condition attribute (physiological index) reduction algorithm

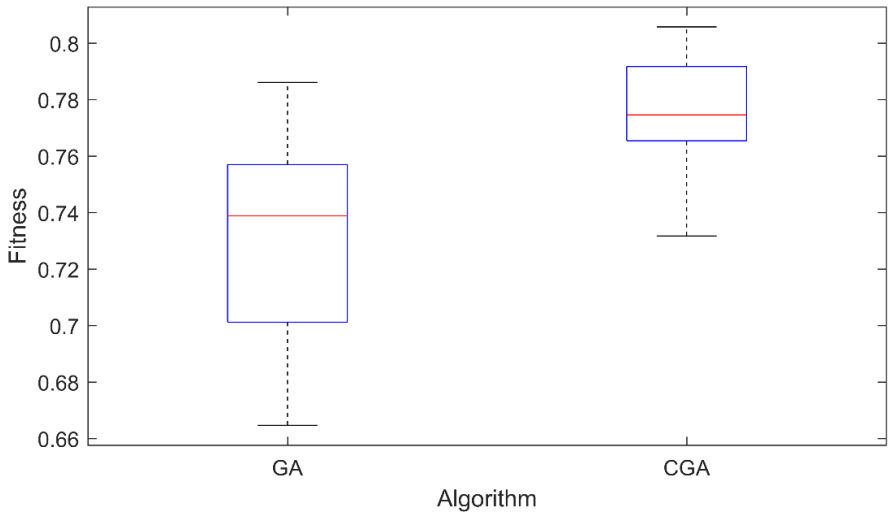

Figure 3: Comparison of fitness box line graphs between the genetic algorithm and the cellular genetic algorithm

First, the degree of fitness can reflect the quality of the algorithm's convergence result when calculating the optimal solution. The greater the fitness, the stronger the ability of the algorithm to find the optimal solution. It can be clearly seen from Figures 2 and 3 and Table 5 that the fitness of the CGA is larger than that of the genetic algorithm, and therefore, the CGA has a stronger ability to find the optimal solution.

Second, the variance of the fitness can reflect the algorithm's stability. The smaller the variance of fitness and the mean square error of the fitness evolution process, the more stable the algorithm. As shown in Figure 3, by comparing the fitness values of 50 independent replicate experiments, the variance of the fitness of the CGA is smaller than that of the genetic algorithm. That is, the fitness of the CGA is more stable, and the overall convergence effect is better. When the condition attribute is no longer being reduced, the last condition attribute iteration calculation is selected for comparison. To eliminate the influence of the initial value, the variance and the average value of the fitness evolution

process of the last 15 algorithm iterations are taken. Table 5 shows that the mean square error of the fitness of the genetic algorithm is much larger than that of the CGA, indicating that the convergence quality of the CGA is better.

In summary, when solving the same problem, the ability of the CGA to find the global optimum is higher than that of the genetic algorithm.

Table 5: Comparison of algorithm performance between the genetic algorithm and the cellular genetic algorithm

Algorithm	Maximum fitness	Mean value of fitness	Mean square error of fitness	Number of algorithm iterations required for convergence
Genetic algorithm	0.7861	0.7322	0.0001560371	20
Cellular genetic algorithm	0.8057	0.7754	0.0000000246	10

Finally, convergence speed is an important indicator of evaluating algorithm performance. It can be seen from Figure 2 and Table 5 that the number of algorithm iterations required for the convergence of the genetic algorithm is greater than that of the CGA, and thus, the CGA is superior to the genetic algorithm in computational efficiency.

By comparing the above genetic algorithm with the CGA, it can be found that the CGA has a higher global optimal ability (higher exploration ability and higher quality of convergence) than the genetic algorithm. The convergence speed of the CGA is also faster (the number of algorithm iterations required for convergence is small), and therefore, the CGA performs better in solving the rough set attribute reduction problem.

4.2.2 The ability of key indicators to predict problems

In this paper, a key indicator prediction model for fatal rebleeding in patients with gastrointestinal bleeding based on rough intensive reduction and machine learning algorithms was established. First, the decision table was reduced by CGA or genetic algorithm, and the conditional attributes of redundancy were removed. There is a certain correlation between the physiological indicators. Attribute reduction can not only eliminate the information overlap between the indicators but also play a role in reducing the dimensions. The speed of the model is improved, while the workload of the collection indicator is reduced, and the diagnosis is more targeted and time-sensitive. Then, based on the key indicators selected, the prediction is made using machine learning algorithms (logical regression, support vector machine, plain Bayesian, and AdaBoost algorithms) [7, 13, 14, 17, 38, 48].

The evaluation indicators of machine learning selected for this study included Accuracy, Precision, and Recal. In the face of medical prediction problems, potential injuries should be identified as far as possible, so recall rates of the evaluation indicators are more important. This study made $\alpha = 1.5$.

The key indicators obtained by the attribute reduction experiment were used as input, and the prediction results were obtained by the cross-validation calculation of ten-folds as shown in Table 6. The corresponding ROC curves are shown in Figures 4, 5, and 6.

It can be seen from Table 6 and Figures 4, 5 and 6 that:

(1) Comparing the CGA-key indicator set and GA-key indicator set, when the intelligent screening algorithm reduces the number of indicators from 64 to eight, the prediction effects of the SVM (Support Vector Machine, SVM) and Naïve Bayes algorithms do not decrease, but increase to be better than the whole indicator set. At the same time, the prediction effects of the AdaBoost algorithm and logistic regression algorithm are not much different from that of the whole index set. When the number of indicators of the genetic algorithm is also reduced from 64 to eight, the prediction accuracy of the model is far less effective than the former two, and there is still a big gap compared with the whole indicator set. This not only shows that the constructed prediction model has a high and stable prediction accuracy, and has a good fitting effect on patient outcomes, but also demonstrates the advantages of the intelligent screening algorithm.

Table 6: Predicted results of rebleeding in patients with gastrointestinal bleeding

Predicts results	Machine learning algorithm	Index number	Validation				
			$F_{1.5}$	AUC	Accuracy	Precision	Recall
All indicators set	AdaBoost	64	0.8273	0.9095	0.8370	0.7937	0.8479
	Logistic regression	64	0.7974	0.8682	0.8109	0.7964	0.8009
	SVM	64	0.2356	0.5319	0.5628	0.0946	0.7457
	Naïve Bayes	64	0.5649	0.7102	0.4890	0.9648	0.4785
CGA-set of key indicators	AdaBoost	8	0.7824	0.8770	0.7976	0.7689	0.7911
	Logistic regression	8	0.7839	0.8782	0.7972	0.7643	0.7950
	SVM	8	0.7683	0.8525	0.7839	0.7541	0.7763
	Naïve Bayes	8	0.5820	0.8408	0.5169	0.9929	0.4930
GA-set of key indicators	AdaBoost	8	0.5658	0.6121	0.5775	0.6110	0.5494
	Logistic regression	8	0.6173	0.6699	0.6421	0.5684	0.6448
	SVM	8	0.5478	0.5701	0.5784	0.5287	0.5578
	Naïve Bayes	8	0.6117	0.6475	0.6091	0.7699	0.5608

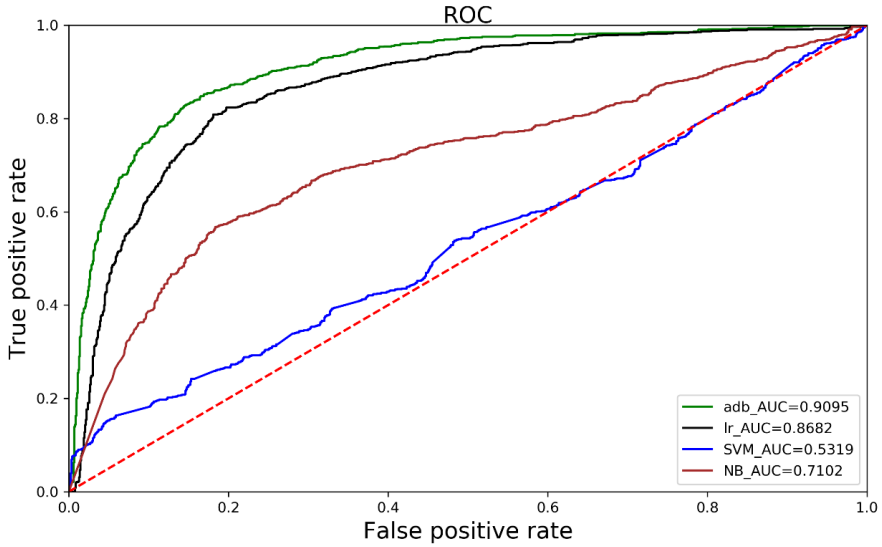

Figure 4: First aid database - full indicator set ROC curve and AUC

(2) Compared with the four prediction models constructed, SVM and Naïve Bayes are more suitable for cases where there are fewer indicators, while AdaBoost and logistic regression have better generalization ability for higher dimensional data.

5 Discussion

The prediction accuracy for traditional disease is affected by interventions in different groups of patients [26, 42], different disease causes [5, 41, 49], and different levels of medical institutions [32]. Strict interventions (such as considering subjects that do not have a large number of comorbidities or complications, i.e., single-disease states), treatment timing (clinical capacity issues that may delay drug treatments), differences in hospital medical levels and conditions (studies are usually performed at large hospitals or developed areas) and the limited number of cases make the application of existing

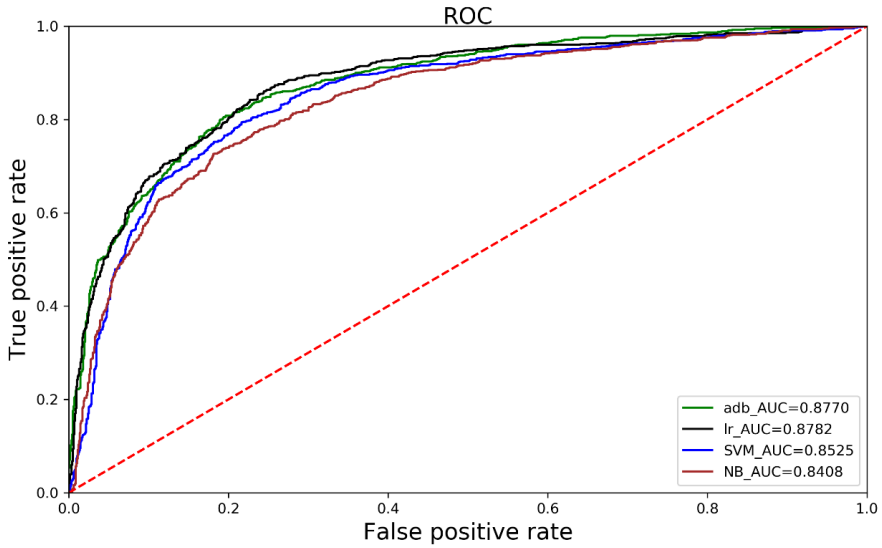

Figure 5: First aid database - intelligent screening algorithm - key indicator set ROC curve and AUC

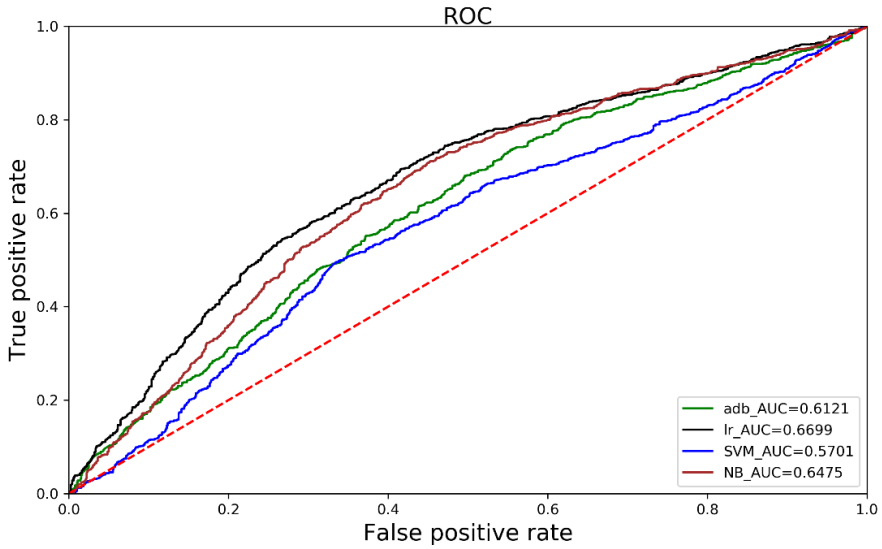

Figure 6: First aid database - genetic algorithm - key indicator set ROC curve and AUC

research results not fully translatable to improving clinical efficacy. Results derived from models including a subset of patients cannot be extended to all patients, which means that the traditional method of key indicator identification still has certain limitations on clinical universality.

Among existing disease prediction methods, two are most common: the traditional statistical method and the artificial intelligence method. Classical statistical methods manually select candidate features based on medical domain knowledge, calculate the importance of each feature based on statistical methods, and construct predictive models. The artificial intelligence method starts with a complete dataset, constructs a machine learning model, automatically extracts features from big data, and emphasizes a large amount of hidden information carried within the medical data itself, thereby achieving greater data mining capacity and outputting values faster.

From the perspective of big data, traditional statistical methods obviously have certain limitations, but their appropriateness and generality are still undeniable. The medical big data analysis method is based on available multi-source medical data, powerful artificial intelligence, and machine learning

algorithms to ensure the accuracy and timeliness of analysis results. However, as some problems remain, such as difficultly controlling confounding factors and imbalanced baseline data, the reliability of the analysis results requires further rigorous clinical trials. Even so, it is undeniable that big data medical care is rooted in real clinical practice and has great research prospects in a wide range of clinical research areas [33].

Based on a large amount of accumulated actual case data, this study introduces an artificial intelligence algorithm to effectively increase the value of data by calculating the key influencing factors of fatal gastrointestinal rebleeding in the hospital. Prediction results of various machine learning algorithms were computed and compared. To quickly realize a more targeted clinical scoring system based on artificial intelligence, technical routes were explored, and the theoretical basis of the system was described. Further clinical trials will be carried out in the future for a wide range of applications.

The application of rough set theory to clinical medical decision support has existed for a long time, and there have been many research results obtained in the fields of clinical decision support, medical image analysis, and others. Tsumoto et al. [35, 36, 37] were the first to apply rough set theory to the clinical expert system. They used rough set theory to generalize rules based on medicinal databases and significantly improved the accuracy of clinical predictions. Xu et al. [47] proposed a new method based on fuzzy rough set theory and information theory to reduce data redundancy, and this method has been used in a cancer recognition classification diagnosis model. Bazan et al. [3] used rough set theory to model classifier networks to identify infant death risk. In 2013, Bazan et al. [4] discussed an application of rough set tools for modelling networks of classifiers and provided clinical decision support for respiratory failure in infants. Son et al. [34] proposed the use of rough set theory and a decision tree to establish a decision model for the early diagnosis of congestive heart failure with an accuracy of 97.5%. Inbarani et al. [12] used new supervised feature selection methods based on a hybridization of particle swarm optimization and rough set theory. These methods were applied to clinical medical diagnosis decision-making to solve the problem of decreased predictive accuracy due to irrelevancy and redundancy of a medical dataset. Chowdhary et al. [6] used an intuitionistic fuzzy rough set technique to extract the features of cancer medical images to assist in cancer diagnosis with an overall accuracy of 98.3%. In summary, it is feasible to apply the rough set theory to clinical medical diagnosis decisions, and the accuracy is much higher than when doing so manually.

Additionally, the method based on rough set theory has been used for disaster rescue. For example, Li et al. [22] and others applied rough set theory for analysing the common points of various disaster medical features upon which many disasters are classified.

6 Conclusion

According to the characteristics of rough set theory, genetic algorithms, and cellular automata, this paper takes the key index prediction of fatal rebleeding in patients with gastrointestinal haemorrhage as an example, and proposes a new intelligent screening algorithm combining genetic algorithms and cellular automata into the rough set theory. The key indicators were selected from a large number of medical indicators and used as input to design prediction models based on logistic regression, SVM, Naïve Bayes, and AdaBoost algorithms.

The results show that: (1) CGA has better performance than the traditional genetic algorithm in solving complex multi-peak optimization problems. (2) In the face of huge medical index analysis problems, attribute reduction can effectively remove redundant data, reduce the dimension of data, and reduce the workload of index extraction and analysis. (3) The combination of intelligent screening algorithm and machine learning algorithm can improve prediction accuracy.

In clinical practice, the population of patients with acute critical disease is large. Due to its complex aetiology, the degree of judgment of medical personnel in most medical institutions depends largely on the individual's subjective experience and intuition, resulting in an inability of institutions to properly identify and offer timely treatment for critically ill patients, prevent potential health hazards, and even prevent disputes. Therefore, objective evaluation criteria are needed in clinical work to provide early warning, so that the medical workers can detect the existence of critical disease or potential critical disease.

In this paper, utilizing big data, a prediction model based on the intelligent screening algorithm and machine learning algorithm was obtained that can effectively identify the high-risk rebleeding population, observe the trend of disease development, and effectively assist doctors' treatment decisions by providing early risk warning.

This paper also found that in terms of early warning for disease, the algorithm and prediction model are superior in improving prediction accuracy. In terms of clinical diagnosis, on the basis of CGA, patients with acute critical disease were further classified according to the causes to help emergency physicians make scientific and accurate triage decisions.

Acknowledgements

This work was partly supported by the National Natural Science Foundation of China (81701961, 71103014), and the National Key Research and Development Plan for Science and Technology Winter Olympics of the Ministry of Science and Technology of China (2019YFF030058).

References

[1] Alba, E.; Dorronsoro, B. (2005). The exploration/exploitation tradeoff in dynamic cellular genetic algorithms. *IEEE Transactions on Evolutionary Computation*, 9(2), 126-142, 2005.

[2] Ali, R.; Hussain J.; Siddiqi, M. H. et al. (2015). H2RM: A Hybrid Rough Set Reasoning Model for Prediction and Management of Diabetes Mellitus. *Sensors*, 15(7), 15921-15951, 2015.

[3] Bazan, J. ; Kruczek, P.; Bazan-Socha, S. et al. (2006). Risk Pattern Identi cation in the Treatment of Infants with Respiratory Failure through Rough Set modeling, *Proceedings of IPMU*, 2–7, 2006.

[4] Bazan, J. ; Kruczek, P.; Bazan-Socha, S. et al. (2006). Automatic Planning of Treatment of Infants with Respiratory Failure through Rough Set modeling, *International Conference on Rough Sets and Current Trends in Computing, Springer, Berlin, Heidelberg*, 418–427, 2006.

[5] Budimir, I.; Gradišer, M.; Nikolić, M. et al. (2016). Glasgow Blatchford, pre-endoscopic Rockall and AIMS65 scores show no difference in predicting rebleeding rate and mortality in variceal bleeding. *Scandinavian Journal of Gastroenterology*, 51(11), 1375-1379, 2016.

[6] Chowdhary, C.L.; Acharjya, D.P. (2016). A hybrid scheme for breast cancer detection using intuitionistic fuzzy rough set technique, *Int. J. Healthc. Inf. Syst. Inform.*, 11 (2), 38–61, 2016.

[7] Direkvand-Moghadam, A.; Khosravi, A.; Sayehmiri, K. (2012). Predictive factors for preeclampsia in pregnant women: a unvariate and multivariate logistic regression analysis, *Acta Biochimica Polonica*, 59(4), 2012.

[8] Dorronsoro, B. (2013). Cellular genetic algorithms without additional parameters. *Journal of Supercomputing*, 63(3), 816-835, 2013.

[9] Fogel, D.B. (1994). An introduction to simulated evolutionary optimization, *IEEE Trans Neural Netw.*, 5(1), 3-14, 1994.

[10] Gil-Herrera, E.; Yalcin, A.; Tsalatsanis, A. et al. (2011). Rough Set Theory based prognostication of life expectancy for terminally ill patients/ Engineering in Medicine and Biology Society, Embc, *2011 International Conference of the IEEE*, 6438-6441, 2011.

[11] Grzymala-Busse, J W. (2008). Three Approaches to Missing Attribute Values: A Rough Set Perspective. In: *Lin T.Y., Xie Y., Wasilewska A., Liau CJ. (eds) Data Mining: Foundations and Practice. Studies in Computational Intelligence, Springer, Berlin, Heidelberg*, 118, 139-152, 2008.

[12] Inbaran, H.H.; Azar, A.T.; Jothi, G. (2014). Supervised hybrid feature selection based on PSO and rough sets for medical diagnosis, *Comput. Methods Progr. Biomed.*, 113(1), 175–185, 2014.

[13] Hadzikadic, M.; Hakenewerth, A.; Bohren, B. et al. (1996). Concept formation vs. logistic regression: predicting death in trauma patients, *Artif. Intell. Med.*, 8(5), 493-504, 1996.

[14] Hamilton, S.L.; Hamilton, J. R. (2012). Predicting in-hospital-death and mortality percentage using logistic regression, *2012 Computing in Cardiology, Krakow*, 489-492, 2012.

[15] Holland, J.H. (1975). Adaptation in natural and artificial systems. *Ann Arbor*, 6(2), 126–137, 1975.

[16] Holland, J.H. (1992). *Adaptation in natural and artificial systems*, MIT Press, 1992.

[17] Hu, Y. H.; Tai, C. T.; Chen, C. C. et al. (2017). Predicting return visits to the emergency department for pediatric patients: Applying supervised learning techniques to the Taiwan National Health Insurance Research Database. *Computer Methods and Programs in Biomedicine*, 144, 105-112, 2017.

[18] Ip, W.C.; Hu, B.Q.; Wong, H. et al. (2009). Applications of grey relational method to river environment quality evaluation in China. *Journal of Hydrology*, 379(3), 284-290, 2009.

[19] Kim, B. J.; Park, M. K.; Kim, S. J. et al. (2009). Comparison of Scoring Systems for the Prediction of Outcomes in Patients with Nonvariceal Upper Gastrointestinal Bleeding: A Prospective Study. *Digestive Diseases & Sciences*, 54(11), 2523-2529, 2009.

[20] Lee, H. H.; Park, J. M.; Lee, S. W. et al. (2015). C-reactive protein as a prognostic indicator for rebleeding in patients with nonvariceal upper gastrointestinal bleeding. *Digestive & Liver Disease*, 47(5), 378-383, 2015.

[21] Kusiak, A.(2000). Evolutionary Computation and Data Mining, *Proceedings of the SPIE Conference on Intelligent Systems and Advanced Manufacturing, B.Gopalakrishnan and A. Gunasekaran (Eds)*, 4192, 1-10, 2000.

[22] Li, T.S.; Li, Z.Y.; Zhao, W. et al. (2020). Analysis of medical rescue strategies based on a rough set and genetic algorithm: A disaster classification perspective. *International Journal of Disaster Risk Reduction*, 42(C), 2020

[23] Lim, K.; Lee, B.M.; Kang, U.; Lee, Y. (2018). An Optimized DBN-based Coronary Heart Disease Risk Prediction, *International Journal of Computers Communications & Control*, 13(4), 492-502, 2018.

[24] Li, J.D.; Cheng, K.W.; Wng, S.H. et al. (2017). Feature selection: a data perspective. *ACM Computing Surveys*, 50(6): 1-45, 2017.

[25] Liu, H.; Dzitac, I.; Guo, S.(2018). Factors Space and its Relationship with Formal Conceptual Analysis: A General View. *International Journal of Computers Communications & Control*, 13(1), 83-98, 2018.

[26] Najarian, K.; Hakimzadeh, R.; Ward, K. et al. (2009). Combining predictive capabilities of transcranial doppler with electrocardiogram to predict hemorrhagic shock, *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE*, 2621-2624, 2009.

[27] Øhrn, A.; Komorowski, J.(1999). Diagnosing Acute Appendicitis with Very Simple Classification Rules, *Principles of Data Mining and Knowledge Discovery. Springer Berlin Heidelberg*, 462-467, 1999.

[28] Pawlak, Z. (1982). Rough set, *International Journal of Computer & Information Sciences*, 11(5), 1982.

[29] Pawlak, Z. (1991). *Rough Sets*, Kluwer Academic Publishers, 1991.

[30] Romagnuolo, J.; Barkun, A.N.; Enns, R. et al. (2007). Simple clinical predictors may obviate urgent endoscopy in selected patients with nonvariceal upper gastrointestinal tract bleeding. *Archives of Internal Medicine*, 167(3), 265-270, 2007.

[31] Sabita Mahapatra, Sreekumar, Mahapatra, S.S. (2010). Attribute selection in marketing: A rough set approach. *IIMB Management Review*, 22(1-2), 16-24, 2010.

[32] Sanders, D. S.; Perry, M. J.; Jones, S. G. et al. (2004). Effectiveness of an upper-gastrointestinal haemorrhage unit: a prospective analysis of 900 consecutive cases using the Rockall score as a method of risk standardisation. *European Journal of Gastroenterology & Hepatology*, 16(5),487, 2004.

[33] Sherman, R E.; Anderson, S A.; Dal Pan, G. J. et al. (2016). Real-World Evidence - What Is It and What Can It Tell Us?. *N Engl J Med*, 375(23), 2293-2297,2016.

[34] Son, C.S.; Kim, Y.N.; Kim, H.S. et al. (2012). Decision-making model for early diagnosis of congestive heart failure using rough set and decision tree approaches, *J. Biomed. Inform.*, 45(5), 999–1008, 2012.

[35] Tsumoto, S.; Tanaka, H. (1994). Induction of Medical Expert System Rules Based on Rough Sets and Resampling methods, *Proceedings of the Annual Symposium on Computer Application in Medical Care, American Medical Informatics Association*, 1066,1994.

[36] Tsumoto, S.; Tanaka, H. (1995). Induction of expert system rules based on rough sets and resampling methods, *Medinfo. MEDINFO 8*, 861–865, 1995.

[37] Tsumoto, S.; Tanaka, H. (1996). Automated Discovery of Medical Expert System Rules from Clinical Databases Based on Rough Sets, *KDD*, 63–69, 1996.

[38] Tao, Z.; Huiling, L.; Wenwen, W. et al. (2019). GA-SVM based feature selection and parameter optimization in hospitalization expense modeling. *Applied Soft Computing*, 75, 323-332, 2019.

[39] Von Neumann, J.; Burks, A.W. (1965). Theory of self-reproducing automata. *IEEE Transactions on Neural Networks*, 5(1): 3-14.

[40] Von Neumann, J. (1948). The general and logical theory of automata. *Papers of John Von Neumann on Computing & Computer Theory*, 1-41, 1948.

[41] Vreeburg, E M.; Terwee, C B.; Snel, P. et al. (1999). Validation of the Rockall risk scoring system in upper gastrointestinal bleeding. *Gut*, 44(3), 331-335, 1999.

[42] Wang, C. Y.; Qin, J.; Wang, J. et al. (2013). Rockall score in predicting outcomes of elderly patients with acute upper gastrointestinal bleeding. *World Journal of Gastroenterology*, 19(22),3466-3472, 2013.

[43] Wang, X.; Yang, J.; Jensen, R. et al. (2006). Rough set feature selection and rule induction for prediction of malignancy degree in brain glioma. *Computer Methods & Programs in Biomedicine*, 83(2):147, 2006.

[44] Wei, Z.; Li, J.; Li, X. et al. (2019). Prediction and feature selection for fatal gastrointestinal bleeding recurrence in hospital via machine learning, *Chinese Critical Care Medicine*, 31(3), 359-362, 2019.

[45] Whitley, D. (1993). Cellular genetic algorithms, *Proceedings of the 5th International Conference on Genetic Algorithms, Morgan Kauffmann*, 658-662, 1993.

[46] Wu, T. T.; Li, H. (2016). Research progress of early warning scoring system for cardiac arrest in hospital. *Chinese Journal of Nursing*, 051(009), 1118-1123, 2016.

[47] Xu, F.F.; Miao, D.Q.; Wei, L.(2009). Fuzzy-rough attribute reduction via mutual information with an application to cancer classification, *Comput. Math. Appl.*, 57 (6), 1010–1017, 2009.

[48] Zhang, Y.; Liu, Z.; Zhang, H. et al. (2012). A Crowding Niche Cellular Genetic Algorithm. *Advanced Materials Research*, 482-484:1933-1936, 2012.

[49] Zhao, S. F.; Qu, Q. Y.; Feng, K. et al. (2017). Comparison of the AIMS65 and Glasgow Blatchford score for risk stratification in elderly patients with upper gastrointestinal bleeding. *European Geriatric Medicine*, 8(1), 37-41, 2017. .

[50] Zhou, T.; Lu H.L.; Wang W. W. et al. (2019). GA-SVM based feature selection and parameter optimization in hospitalization expense modeling. *Applied Soft Computing*, 75, 323-332, 2019.

[51] Zhu, X. Z.; Zhu, W.; Fan, X. N.(2017). Rough set methods in feature selection via submodular function. *Soft Computing*, 21(13), 3699-3711, 2017.

Journal's webpage: http://univagora.ro/jour/index.php/ijccc/

This journal is a member of, and subscribes to the principles of,
the Committee on Publication Ethics (COPE).
https://publicationethics.org/members/international-journal-computers-communications-and-control

Cite this paper as:
Jia, L.; Wang, H.; Zhuang, H.; Li, X. et al. (2020). A Rough Set and Cellular Genetic Fusion Algorithm for Acute Critical Disease Prediction, *International Journal of Computers Communications & Control*, 15(6), 3894, 2020.
https://doi.org/10.15837/ijccc.2020.6.3894.