Hindawi
Complexity
Volume 2018, Article ID 4824350, 12 pages
https://doi.org/10.1155/2018/4824350

WILEY | Hindawi

Research Article

Differential Diagnosis Model of Hypocellular Myelodysplastic Syndrome and Aplastic Anemia Based on the Medical Big Data Platform

Jianhui Wu,[1,2] Lu Zhang,[1] Sufeng Yin,[1,2] Haidong Wang,[1] Guoli Wang,[1,2] and Juxiang Yuan [1,2]

[1]*School of Public Health, North China University of Science and Technology, Tangshan, Hebei 063210, China*
[2]*Hebei Province Key Laboratory of Occupational Health and Safety for Coal Industry, North China University of Science and Technology, Tangshan, Hebei 063210, China*

Correspondence should be addressed to Juxiang Yuan; gwxyjxb@ncst.edu.cn

Received 10 July 2018; Revised 4 September 2018; Accepted 12 September 2018; Published 12 November 2018

Guest Editor: Zhihan Lv

The arrival of the era of big data has brought new ideas to solve problems for all walks of life. Medical clinical data is collected and stored in the medical field by utilizing the medical big data platform. Based on medical information big data, new ideas and methods for the differential diagnosis of hypo-MDS and AA are studied. The basic information, peripheral blood classification counts, peripheral blood cell morphology, bone marrow cell morphology, and other information were collected from patients diagnosed with hypo-MDS and AA diagnosed in the first diagnosis. First, statistical analysis was performed. Then, the logistic regression model, decision tree model, BP neural network model, and support vector machine (SVM) model of hypo-MDS and AA were established. The sensitivity, specificity, Youden index, positive likelihood ratio (+LR), negative likelihood ratio (−LR), area under curve (AUC), accuracy, Kappa value, positive predictive value (+PV), negative predictive value (−PV) of the four model training set and test set were compared, respectively. Finally, with the support of medical big data, using logistic regression, decision tree, BP neural network, and SVM four classification algorithms, the decision tree algorithm is optimal for the classification of hypo-MDS and AA and analyzes the characteristics of the optimal model misjudgment data.

1. Introduction

Myelodysplastic syndrome (MDS) is a clonal disease of acquired hematopoietic stem/progenitor cells, which is transformed into clinical features by myelocyte hematopoiesis and high risk to acute myeloid leukemia [1]. Some patients with MDS have low bone marrow hyperplasia, called hypocellular myelodysplastic syndrome (hypo-MDS). Hypo-MDS is a special type of MDS, accounting for 8.2%–29.0% of the total number of MDS, up to 38.0% [1]. Aplastic anemia (AA) refers to the primary bone marrow hematopoietic failure syndrome. The etiology is unknown, mainly manifested as low bone marrow hematopoietic function and complete blood cell reduction. Clinically, there may be bleeding and infection performance [2, 3].

At present, the differential diagnosis of hypo-MDS and AA is mainly carried out by hematology, cell morphology, bone marrow biopsy, and cytogenetics. In different stages of disease development, the peripheral blood of patients with hypo-MDS and AA may be reduced in one line, two lines, or three lines simultaneously [4, 5]. Pathological hematopoiesis is a major indicator of clinical diagnosis of hypo-MDS, but it has the disadvantages of poor reproducibility, poor specificity, and low sensitivity. Furthermore, pathological hematopoiesis can also be seen in some patients with AA [5]. Some studies have also found that there is no pathological hematopoietic MDS [6]. These show the nonspecificity of pathological hematopoiesis. Previously, cytogenetic abnormalities were considered to be reliable diagnostic criteria for hypo-MDS, but the detection rate of chromosomal

abnormalities in MDS patients ranged from 40% to 60% [7, 8] and even lower in hypo-MDS patients [9]. It can be seen that the abnormal cytogenetic ratio of MDS is not very high, suggesting that the index is not specific. In recent years, the value of flow cytometry (FCM) in the differential diagnosis of AA and hypo-MDS has become increasingly important [10, 11], but the differential diagnosis of hypo-MDS and AA with a single immunophenotypic marker is too low. The use of FCM to assess erythroid malignancies (a milestone in the diagnosis of MDS morphology) is difficult, limiting the widespread use of FCM in the diagnosis of MDS [12].

It can be seen that the pathological features and clinical manifestations of hypo-MDS and AA are very similar, and there are many differential diagnosis indicators, but the specificity is not high and the differential diagnosis of these two diseases is difficult in clinical practice. Every diagnosis process of disease will produce a large amount of data, and the data contain a lot of information about the disease. Therefore, using the collected big data for data mining, we can effectively analyze the disease.

Data mining refers to the process of extracting knowledge and information that has potential application value from large databases. It is a new type of the information processing system that has been rapidly developed in recent years [13]. Classification is a very important task in data mining. Commonly used methods include logistic regression, neural networks, decision trees, and SVM. Each of these methods has its own characteristics, has a strong representation in the classification algorithm, and has been widely and successfully applied in the medical field [14–16].

Some scholars have compared the classification effects of data mining classification methods in the medical field. For example, Agarwal et al. [17] compared the Bayesian, SVM, and decision tree classification results using medical data. The results show that the SVM has the highest classification accuracy. Heydari et al. [18] compared neural network, SVM, decision tree, and Bayesian methods in the diagnosis of type 2 diabetes and found that the highest accuracy of the neural network model is 97.44%, the decision tree is 95.03%, and the Bayesian network is 91.60%, while the accuracy of SVM is only 81.19%. Lui et al. [19] used SVM, Bayesian networks, radial basis neural networks, and multilayer perceptrons to establish a classification model of magnetic resonance features of mild traumatic brain injury. The highest accuracy rate is the radial basis neural networks (74%); the worst is the multilayer sensor (66%), and SVM and Bayesian network are 70%. Tseng et al. [20] used decision trees and neural network methods to analyze the prognosis of oral cancer patients and found that both methods had higher accuracy, but compared to neural networks, the results of the decision tree model are easier to explain and easier to accept. Wu et al. [21] compared the classification performance of the BP neural network and logistic regression and found that the classification accuracy of the BP neural network (93.5%) was higher than that of the logistic regression model (90.7%).

Based on the current clinical problems of differential diagnosis of hypo-MDS and AA, the case data of hypo-MDS patients and AA patients were analyzed, the data that did not conform to the actual and the errors were deleted, and the pure data was obtained. Then, applying pure big data to the data mining algorithm was done to compare the effects. In this study, logistic regression, decision tree, BP neural network, and SVM are used to establish the differential diagnosis model of two diseases. Through the evaluation of the model, a better classification model is finally obtained, combined with the clinical features of the misdiagnosed cases of the best differential diagnosis model, and the combined differential diagnosis is performed. This provides an effective new idea and method for the differential diagnosis of hypo-MDS and AA.

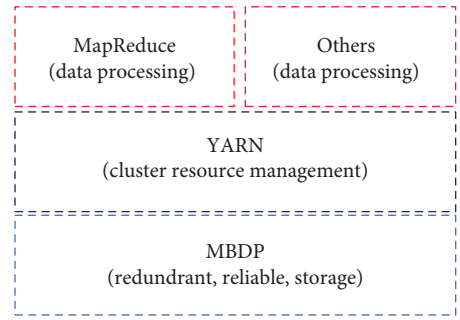

Figure 1: The core of the medical big data platform.

2. Medical Big Data Acquisition and Storage System Based on the Medical Big Data Platform

2.1. Medical Big Data Platform Software Architecture. Medical big data platform (MBDP) is a distributed system infrastructure developed by the Apache Foundation that allows users to develop distributed programs without knowing the underlying details of the distribution and take advantage of the power of clusters for high-speed computing and storage. Medical big data platform provides developers with a reliable, efficient, and scalable open source software framework for processing massive amounts of data. It realizes distributed computing of massive data in a cluster composed of a large number of computers. The medical big data platform open source distributed computing platform is mainly composed of two parts: medical big data platform distributed file system and MapReduce distributed computing framework (see Figure 1).

Medical big data platform is an open source, distributed storage, distributed computing platform that extends a single server to a cluster machine, with each node providing local computing and storage without relying on hardware for high availability. As the core component, MapReduce is used to implement task decomposition and scheduling. MBDP is used to store massive amounts of data. By storing the medical big data of clinical patients in real time and further effectively calculating and processing, the application value of the medical big data platform is fully utilized.

2.2. Research on Distributed Optimization of MBDP Based on Big Data. MBDP has been excellent enough in stability and performance, but it has low storage efficiency, cluster load

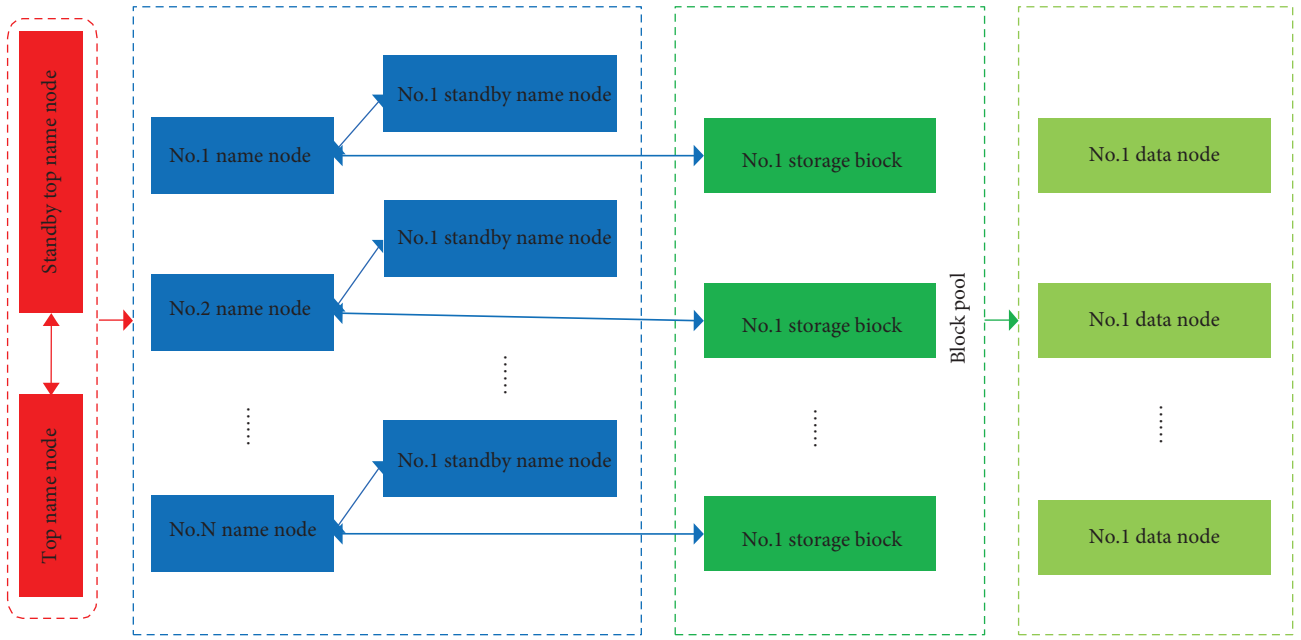

Figure 2: The structure diagram of the MBDP system for distributed nodes.

balancing ability is poor, NameNode single point failure, JobTracer load is too large, small file problem, hot spot problem, etc. Both seriously restrict the further development of MBDP. In order to achieve higher storage efficiency and more optimized load balancing capabilities for MBDP, an improved solution for MBDP is Noah. The management of the section is done by the mapping file to each node of the cluster, which solves the performance bottleneck problem of the central node (see Figure 2).

The experimental results show that Noah improves the data recovery speed of MBDP while ensuring the security of cluster data, optimizes the load balancing capability of MBDP, and reduces the overall storage cost of the medical big data platform. This has obvious implications for improving the actual operational efficiency of the medical big data platform and its associated cloud computing architecture.

2.3. Storage Platform Framework Based on Hypo-MDS and AA Case Big Data. According to the timeliness and large reserves of hypo-MDS and AA case data, the medical big data platform distributed storage system is placed in the virtualization pool of the resource management platform, with the medical big data platform slave node deployed dynamically, and the medical big data platform distributed storage is quickly built.

The newly built big data storage platform has good compatibility and long life cycle. The medical diagnosis process data is stored in the platform in real time to realize data analysis and processing. In the process of data storage, patient's basic information, peripheral blood classification count, peripheral blood cell morphology, bone marrow cell morphology, and other quantifiable data are included. Further interface with the classification system to achieve differential diagnosis of hypoplastic myelodysplastic syndrome and aplastic anemia is needed.

3. Big Data Based on Hypo-MDS and AA Cases

3.1. Storage Database Construction of Hypo-MDS and AA Cases Big Data

3.1.1. Data Collection for Hypo-MDS and AA Cases. Case data of hypo-MDS patients and AA patients were taken from the Affiliated Hospital of North China University of Technology and the Chinese Academy of Medical Sciences Blood Disease Hospital. A medical information database was made to collect basic information and medical history of eligible patients including the patient's gender, age, occupation, marital status, and smoking and drinking history. And the clinical examination data of the patients, including the peripheral blood classification count, peripheral blood cell morphology, and bone marrow cell morphology, were also collected.

3.1.2. Inclusion Criteria. The inclusion criteria include the following data:

(1) Newly diagnosed cases admitted from January 1, 2008, to December 31, 2016

(2) All hypo-MDS and AA cases met the 2008 revised WHO MDS classification criteria and blood disease diagnosis and efficacy criteria (third edition). The hypo-MDS also needs to meet the bone marrow tissue biopsy. The bone marrow cell volume is less than 30% for those under 60 years old or less than 20% for those over 60 years old and confirmed by a number of blood disease experts

(3) Patient case information was recorded using standard cases

3.1.3. Exclusion Criteria. The exclusion criteria include the following data:

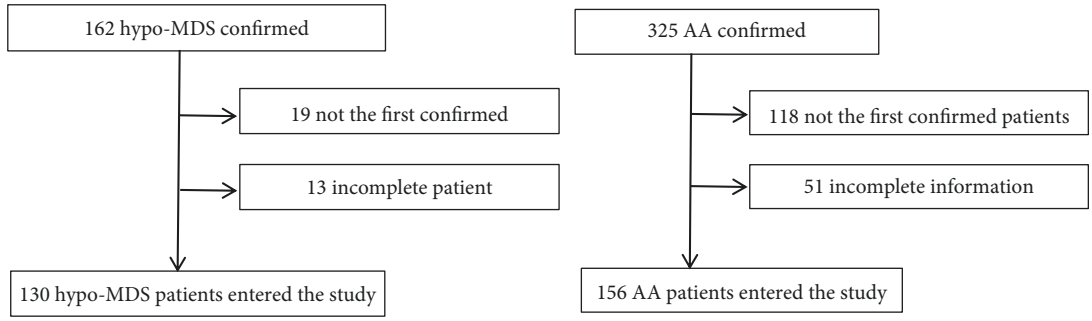

Figure 3: The selection of research object.

(1) Have severe wasting diseases

(2) Have a history of malignant tumors

3.2. Hypo-MDS and AA Diagnostic Criteria [2]. Hypo-MDS and AA disease diagnostic criteria overlap in hematology and cell morphology such as peripheral blood cell reduction and bone marrow hyperplasia [22]. How to distinguish between the two is often a big problem that plagues clinicians. The use of data mining methods to apply the collected data to the differential diagnosis of low proliferative myelodysplastic syndrome and aplastic anemia will greatly improve the accuracy of diagnosis.

3.2.1. Hypo-DMS Diagnostic Criteria. Hypo-MDS has so far no unified diagnostic criteria. The reference conditions for hypo-MDS diagnosis are as follows. (1) Peripheral blood showed more than two series of cytopenias, and the original cells or nucleated red blood cells could be seen in the classification. (2) Bone marrow smears show hyperplasia at more than two sites. (3) Bone marrow sections show a decrease in the bone marrow hematopoietic area and bone marrow cell volume, less than 30% for those under 60 years old and less than 20% for those over 60 years old. (4) The bone marrow has a pathological hematopoiesis in one or both blood cells, and the number of primitive cells varies depending on the MDS subtype.

3.2.2. AA Diagnostic Criteria. The AA diagnostic criteria include the following data: (1) the reduction of whole blood cells, the percentage of reticulocytes < 1%, and the increase of the proportion of lymphocytes; (2) generally without hepatosplenomegaly; (3) reduced hyperplasia of bone marrow (<normal 50%) or severe reduction (<normal 25%), decreased hematopoietic cells, increased proportion of non-hematopoietic cells, and empty bone marrow granules (bone marrow biopsy shows that hematopoietic tissue is reduced); (4) can exclude other diseases that cause pancytopenia, such as PNH, acute hematopoietic function arrest, megaloblastic anemia, myelofibrosis, and acute leukemia.

3.3. Analysis of Big Data in Hypo-MDS and AA Cases

3.3.1. General Situation of the Research Object. From January 1, 2008, to December 31, 2016, patients with hypo-MDS and AA diagnosed at the Institute of Hematology, Chinese Academy of Medical Sciences, and the Affiliated Hospital of North China University of Technology were selected as the study subjects. A total of 325 cases of AA patients were collected, among which 118 were not diagnosed for the first time and 51 cases were incomplete. A total of 156 AA patients entered the study. We collected 162 cases of patients with hypo-MDS, of which 19 were not first diagnosed and 13 cases were incomplete. In total, 130 patients with hypo-MDS entered the study (see Figure 3).

Of 156 patients with AA, 83 were men (53.20%) and 73 were women (46.80%). The age range was 6–80 years, and the average age was 28.51 ± 15.46 years. Of the 130 patients with hypo-MDS, 69 (53.08%) were males and 61 (46.92%) were females. The age range was 11–82 years, and the average age was 36.81 ± 16.42 years. The difference in age between the two diseases was statistically significant ($\chi^2 = 16.74$, $P = 0.001$). The two diseases are in gender composition ($\chi^2 < 0.001$, $P = 0.983$), marital status ($\chi^2 = 1.26$, $P = 0.261$), history of smoking ($\chi^2 = 1.70$, $P = 0.193$), ethnic ($\chi^2 = 0.79$, $P = 0.374$), and drinking history ($\chi^2 = 3.80$, $P = 0.051$); there were no statistically significant differences in such aspects.

In terms of occupational composition of patients, the patient populations of the two diseases are mainly concentrated in workers, farmers, and students. The difference in occupational composition between the two diseases was statistically significant ($\chi^2 = 49.87$, $P < 0.001$). The proportion of farmers with hypo-MDS is the highest (38.46%), while the percentage of students with AA is the highest (51.92%) (see Table 1).

3.3.2. Results of Laboratory Tests in Two Groups of Patients. Peripheral blood cell counts, blood smears, and bone marrow smears were analyzed in 130 patients with hypo-MDS and 156 patients with AA. Blood cell counts showed that the red blood cell content and hemoglobin content in hypo-MDS patients were lower than those in AA patients, and the difference was statistically significant ($P < 0.05$). The platelet content of patients with hypo-MDS was lower than that of patients with AA, but there was no significant difference between the two groups ($P > 0.05$).

Blood smear showed that the proportion of neutrophils in rod-shaped nucleus was lower in hypo-MDS patients than in AA patients, and the proportion of mature lymphocytes was lower in AA patients than in AA patients ($P < 0.05$). The proportion of neutrophilic neutrophils and mature mononuclear cells in patients with hypo-MDS was higher

Table 1: The basic characteristics of the patients.

General information	Category	AA (%)	Hypo-MDS (%)	χ^2	P
Age	0~14	26 (16.67)	5 (3.85)	18.04	<0.001
	15~29	70 (44.87)	50 (39.23)		
	30~59	54 (33.97)	63 (48.46)		
	≥60	6 (4.49)	12 (9.23)		
Sex	Male	83 (53.21)	69 (53.08)	<0.01	0.983
	Female	73 (46.79)	61 (46.92)		
Marital status	Unmarried	80 (51.28)	58 (44.62)	1.26	0.261
	Married	76 (48.72)	72 (55.38)		
Ethnic	Minority	8 (5.13)	10 (7.69)	0.79	0.374
	Ethnic Han	148 (94.87)	120 (92.31)		
Profession	Cadres	10 (6.41)	0 (0.00)	60.93	<0.001
	Workers	36 (20.08)	19 (14.62)		
	Farmers	24 (15.39)	50 (38.46)		
	Self-employed	2 (1.28)	4 (3.08)		
	Students	81 (51.92)	36 (27.69)		
	No work	3 (1.92)	21 (16.15)		

Table 2: Laboratory results ($\bar{x} \pm s$).

Item	Index	Hypo-MDS ($n = 130$)	AA ($n = 156$)	t	P
Blood cell count	WBC ($\times 10^{12}$/L)	2.87 ± 1.25	3.05 ± 1.22	1.21	0.229
	RBC ($\times 10^{12}$/L)	2.08 ± 0.71	2.57 ± 0.78	5.45	<0.001
	HGB (g/L)	68.89 ± 22.13	83.13 ± 24.85	5.07	<0.001
	PLT ($\times 10^{9}$/L)	44.12 ± 77.81	34.31 ± 35.32	1.33	0.186
Blood smear (%)	Rod-like nuclear neutrophils	8.26 ± 11.28	7.13 ± 8.42	0.97	0.335
	Lobular nuclear neutrophils	25.88 ± 14.63	21.54 ± 16.23	2.34	0.020
	Mature lymphocyte	59.07 ± 18.51	65.62 ± 20.78	2.79	0.006
	Mature monocyte	3.35 ± 3.70	2.73 ± 2.68	1.57	0.117
Marrow smear (%)	Progranulocyte	0.55 ± 0.94	0.34 ± 0.59	2.14	0.033
	Neutrophil neutrophils	3.18 ± 2.65	3.27 ± 3.61	0.25	0.804
	Neutrophil metamylocyte	2.90 ± 2.31	3.93 ± 4.06	2.70	0.008
	Rod-like nuclear neutrophils	7.65 ± 3.54	8.86 ± 6.19	1.97	0.050
	Lobular nuclear neutrophils	4.79 ± 4.67	7.14 ± 7.73	3.17	0.002
	Basophilic normoblast	0.81 ± 1.11	0.47 ± 0.92	2.75	0.006
	Polychromatic normoblast	11.16 ± 11.41	5.53 ± 7.48	4.83	<0.001
	Orthochromatic normoblast	21.85 ± 15.61	12.75 ± 10.95	5.59	<0.001
	Mature lymphocyte	43.42 ± 25.43	52.27 ± 24.75	2.97	0.003
	Mature monocyte	2.00 ± 6.50	1.38 ± 1.56	1.16	0.249
	Mature plasma cell	0.64 ± 0.92	0.98 ± 1.49	2.37	0.019

than that in patients with AA, but there was no significant difference between the two groups ($P > 0.05$).

The morphology of myeloid cells showed that the proportion of precocious neutrophils, late neutrophils, neutrophils, polymorphonuclear neutrophils, and mature lymphocytes was lower in patients with hypo-MDS than in patients with AA. The proportion of early red blood cells, medium and young red blood cells, late young red blood cells, and mature plasma cells is higher in hypo-MDS patients than in AA patients. And the difference was statistically significant ($P < 0.05$). The proportion of mature monocytes in patients with hypo-MDS is higher than that of patients with AA. The proportion of neutrophils and rod-shaped nuclear neutrophils is lower than that of AA patients, but the difference was not statistically significant ($P > 0.05$) (see Table 2).

3.3.3. Variable Selection and Assignment. Although the difference in occupational composition between hypo-MDS and

Table 3: Variable assignment table.

Variable	Definition	Evaluation
Y	Type of disease	AA = 0, hypo-MDS = 1
X1	Age	Continuous variable
X2	RBC ($\times 10^{12}$/L)	Continuous variable
X3	HGB (g/L)	Continuous variable
X4	Lobular nuclear neutrophils of blood smear (%)	Continuous variable
X5	Mature lymphocyte of blood smear (%)	Continuous variable
X6	Progranulocyte of marrow smear (%)	Continuous variable
X7	Neutrophil metamylocyte of marrow smear (%)	Continuous variable
X8	Lobular nuclear neutrophils of marrow smear (%)	Continuous variable
X9	Basophilic normoblast of marrow smear (%)	Continuous variable
X10	Polychromatic normoblast of marrow smear (%)	Continuous variable
X11	Orthochromatic normoblast of marrow smear (%)	Continuous variable
X12	Mature lymphocyte of marrow smear (%)	Continuous variable
X13	Mature plasma cell of marrow smear (%)	Continuous variable

AA is statistically significant, there is no evidence that the prevalence of hypo-MDS and AA is related to occupational factors, so occupational factors are not included in the establishment of the model. Red blood cells and hemoglobin in blood cell counts were included in the establishment of the model as a basic reference for the differential diagnosis of clinical hypo-MDS and AA. There is a literature supporting [23] that neutrophils, precocious erythroblasts, medium and young erythrocytes, late erythroblasts, mature lymphocytes, and mature plasma cells contribute to the identification of hypo-MDS and AA, so these indicators were also included in the model (see Table 3 for variable assignments).

4. Decision Tree-Based Differential Diagnosis Model

4.1. The Establishment of a Decision Tree Model. The decision tree [24] is a layered rule of a tree structure formed by a top-down transfer method by determining a series of logical branch relationships. The root node, intermediate nodes, and leaf nodes are generated in the decision tree generation process. The root node, intermediate nodes, and leaf nodes are generated in the decision tree generation process. The root node of the decision tree is the beginning of the decision tree. It represents the most distinguishing feature variable of the sample data. Then, the feature classification point of the node was selected to split the node until the data of a certain node only belongs to one category or the variance is the smallest, and the node will not split.

The key issue of decision tree generation is the selection of the most partitioned attributes, namely, the selection of node features and feature splitting points. As the decision tree continues to grow downwards to generate various branch nodes, we hope that the samples contained in each node belong to the same category as much as possible, that is, the impurity of the growing nodes of the tree is getting lower and lower. According to different decision tree algorithms, there are three methods used to measure the degree of node impurity [25, 26]: information gain, gain ratio, and Gini index.

Table 4: Classification accuracy of training set and test set (n (%)).

Subarea	Training set (%)	Test set (%)
Correct	199 (95.22)	62 (80.52)
Error	10 (4.78)	15 (19.48)
Total	209 (100.00)	77 (100.00)

Table 5: Decision tree model evaluation.

Aspect	Index	Training set result	Test set result
Authenticity	Sensitivity (%)	98.96	76.47
	Specificity (%)	92.04	83.72
	Youden index	0.91	0.60
	+LR	12.42	4.70
	−LR	0.01	0.28
	AUC (95% CI)	0.96 (0.92, 0.98)	0.80 (0.70, 0.88)
Reliability	Accuracy (%)	95.22	80.52
	Kappa	0.90	0.60
Benefit	+PV (%)	91.35	78.79
	−PV (%)	99.05	81.82

The C5.0 algorithm in the decision tree model often uses information gain to select node features and feature split points. The calculation method is as follows. Information entropy is an indicator used to describe the purity of sample data. Assume that the relative frequency of c samples in sample data set I is p_c $(c = 1, 2, \ldots, |\gamma|)$, then, the information entropy of I is

$$\mathrm{Ent}(I) = -\sum_{c=1}^{|\gamma|} p_c \log_2 p_c. \quad (1)$$

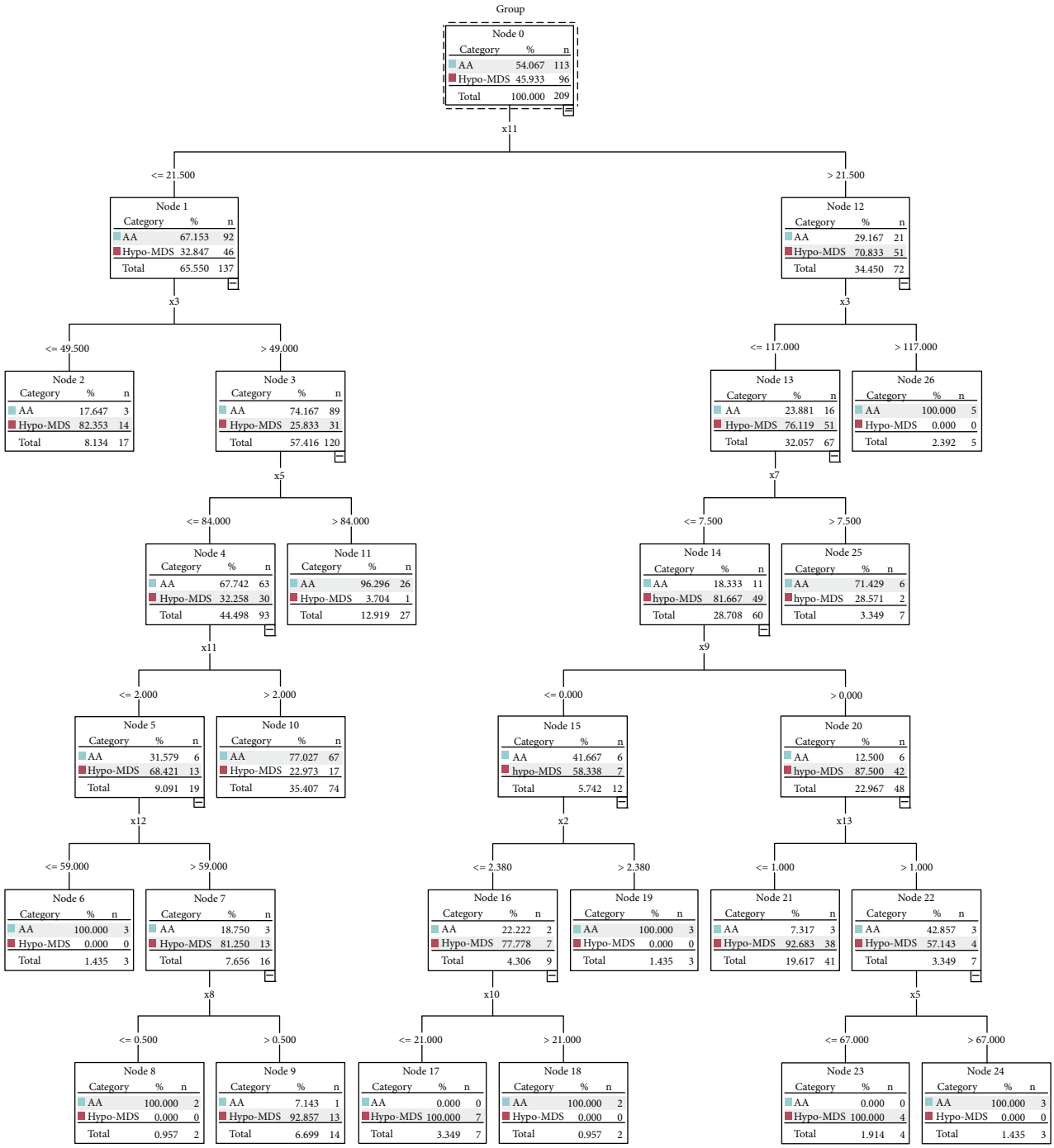

Figure 4: Decision tree model of hypo-MDS and AA.

The smaller the $\mathrm{Ent}(I)$, the higher the purity of I. When sample data is evenly distributed in each category, the maximum entropy $\log_2 C$ is used to indicate the lowest purity. When all samples belong to the same category, the information entropy has a minimum value of 0, indicating the highest purity.

Assume that a is the attribute of the sample data set I, there are V possible values $\{a^1, a^2, \ldots, a^V\}$; then, we can use the attribute a to make a V branch nodes after zapping the sample data set I. We note that in sample data set I contained in the v branch node, all samples on a that have an a^v value are I^v. Therefore, the information gain obtained by dividing attribute data set I with attribute a is

$$\mathrm{Gain}(I, a) = \mathrm{Ent}(I) - \sum_{v=1}^{V} \frac{|I^v|}{|I|} \mathrm{Ent}(I^v). \tag{2}$$

In general, the greater the information gain, the greater the purity of the division of the sample data set by the attribute a. Therefore, the information gain can be used to select the division attribute of the decision tree.

The common gain rate of the C4.5 algorithm in the decision tree model is used to select node features and feature splitting points. Using the same sign as the information gain calculation, the gain rate is defined as

$$\text{Gain\_ratio}(I, a) = \frac{\text{Gain}(I, a)}{DV(a)}, \tag{3}$$

where

$$DV(a) = -\sum_{v=1}^{V} \frac{|I^v|}{|I|} \log_2 \frac{|I^v|}{|I|}. \tag{4}$$

This is called the intrinsic value of the attribute a. The more possible the value V of the attribute a, the larger the value of the $DV(a)$ will generally be.

The CART algorithm in the decision tree model uses the Gini index to select node features and feature splitting points. Using the same sign as the information gain calculation, the Gini index of sample data set I can be expressed as

$$\text{Gini}(I) = \sum_{c=1}^{|\gamma|} \sum_{c \neq c} p_c p_{c'} = 1 - \sum_{c=1}^{|\gamma|} p_c^2. \tag{5}$$

From the sample data set I, randomly selected two samples, according to the above formula, can be obtained, $\text{Gini}(D)$ reflects the probability of inconsistency between the two random sample categories. Thus, the smaller the $\text{Gini}(D)$, the higher the purity of the sample data set I.

The Gini index for attributes is defined as

$$\text{Gini\_index}(D, a) = \sum_{v=1}^{V} \frac{|I^v|}{|I|} \text{Gini}(I^v). \tag{6}$$

Therefore, we choose the attribute with the smallest Gini index as the optimal partition attribute in the candidate attribute set A, namely,

$$a_* = \arg\min_{a \in A} \text{Gini\_index}(I, a). \tag{7}$$

4.2. Pruning of Decision Trees. In the top-down generation process of decision trees, overfitting often occurs if there is no restriction on its growth. At this point, the decision tree needs to be pruned to correct overfitting. The pruning of decision trees cannot be arbitrarily done, and it often needs to take into account the prediction accuracy and complexity of the decision tree; otherwise, it will cause decision loss. Pruning is divided into prepruning and postpruning according to the time of pruning [27]. Prepruning occurs during the growth of the decision tree and is estimated before the node is divided. If the division at this time does not improve the performance of the decision tree, then the partitioning is stopped and the decision branch of the decision tree is reduced. After the pruning occurs after the completion of the growth of the decision tree, the nonleaf node is evaluated. If the subtree under the node can replace the leaf node to improve the performance of the decision tree, it is pruned to prevent overfitting.

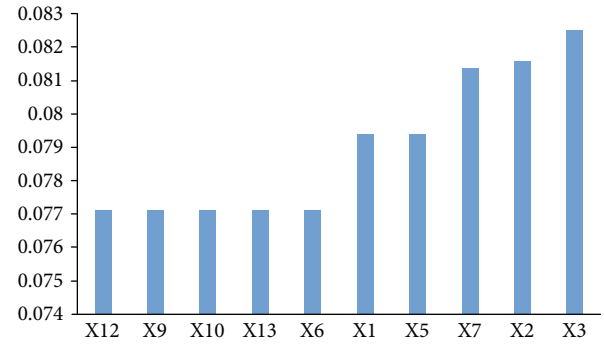

Figure 5: Analysis of the importance of independent variables to the model.

5. Decision Tree Model Establishment and Solution

5.1. Establishment of Hypo-MDS and AA Decision Tree Models. The sample big data is partitioned; the training partition is 73% in the model establishment process, and the test partition accounts for 27%. The C5.0 algorithm is used to select the boosting method and cross-validation. The pruning severity is set to 75, and the minimum number of records per subbranch is 2. The global pruning is chosen to establish a decision tree model for the two diseases. The model of the training set was 209 cases: 199 cases were correctly classified and 10 cases were misclassified. The test set samples were 77 cases: 62 cases were correctly classified and 15 cases were misclassified (see Table 4). Sensitivity, specificity, Youden index, positive likelihood ratio, negative likelihood ratio, AUC, accuracy, Kappa value, positive predictive value, and negative predictive value of the model classification were evaluated (see Table 5).

The dendrogram depth is 8, and there are 9 layered nodes. The proportion of late erythroblasts in bone marrow cells is used as the root node to develop the growth of the decision tree. After the growth of the decision tree is completed, we can extract valid information according to the decision rules of the decision tree, in order to achieve the purpose of identifying hypo-MDS and AA (see Figure 4). For example, the decision message passed to us by node 4 is that the percentage of late erythroblasts in bone marrow cells is less than 26.50% and that of peripheral blood red blood cells is greater than 1.36%. When the age is less than 39 years old, the likelihood of the patient being AA is 76.92% and the probability of the patient being hypo-MDS is 23.08%. The analysis of the effect of independent variables on the model showed that peripheral blood red blood cells had the greatest influence on model classification, followed by medium and young red and late young red blood cells in bone marrow cells (see Figure 5).

Table 6: The comparison of the training set evaluation index.

Aspects	Index	Logistic	Decision tree	BP neural network	SVM
Authenticity	Sensitivity (%)	68.75	98.96	84.38	69.79
	Specificity (%)	78.76	92.04	78.76	82.30
	Youden index	0.48	0.91	0.63	0.52
	+LR	3.24	12.42	3.97	3.94
	−LR	0.40	0.01	0.20	0.37
	AUC (95% CI)	0.74 (0.67, 0.80)	0.96 (0.92, 0.98)	0.82 (0.76, 0.87)	0.76 (0.70, 0.82)
Reliability	Accuracy (%)	74.16	95.22	81.34	76.56
	Kappa	0.48	0.90	0.63	0.53
Benefit	+PV (%)	73.33	91.35	77.14	77.01
	−PV (%)	74.79	99.05	85.58	76.23

Table 7: The comparison of the test set evaluation index [28].

Aspects	Index	Logistic	Decision tree	BP neural network	SVM
Authenticity	Sensitivity (%)	70.59	76.47	76.47	67.65
	Specificity (%)	72.09	83.72	72.09	76.74
	Youden index	0.43	0.60	0.49	0.44
	+LR	2.53	4.70	2.74	2.91
	−LR	0.41	0.28	0.33	0.42
	AUC (95% CI)	0.71 (0.60, 0.81)	0.80 (0.70, 0.88)	0.74 (0.63, 0.84)	0.72 (0.61, 0.82)
Reliability	Accuracy (%)	71.43	80.52	74.03	72.73
	Kappa	0.42	0.60	0.48	0.45
Benefit	+PV (%)	66.67	78.79	68.42	69.70
	−PV (%)	75.61	81.82	79.49	75.00

5.2. Comparison of Hypo-MDS and AA Classification Effects by Four Models

5.2.1. Results for Training Set Samples. Combining the above results, logistic regression, decision tree, BP neural network, and SVM are used to evaluate the classification models of hypo-MDS and AA big data from three aspects: authenticity, reliability, and benefit. The results show that, in terms of the comparison of authenticity evaluation, logistic regression, decision tree, BP neural network, and SVM, the decision tree model has the best authenticity. In terms of reliability evaluation, the reliability of the decision tree model is best compared with logistic regression, decision trees, BP neural networks, and SVM. In terms of model benefits, logistic regression, decision tree, BP neural network, and support vector machine have the highest benefit compared to the decision tree model (see Table 6).

After comparison, the sensitivity difference between logistic regression model and decision tree model and between decision tree model and support vector machine has statistical significance ($P < 0.05$) (Table 7). There is no statistically significant difference among other models ($P > 0.05$). The difference in specificity between logistic regression model and decision tree model, decision tree model and BP neural network, and decision tree model and support vector machine has statistical significance ($P < 0.05$). There is no statistically significant difference among other models ($P > 0.05$). The difference in accuracy between logistic regression model and decision tree model, decision tree model and BP neural network, and decision tree model and support vector machine has statistical significance ($P < 0.05$). There is no statistically significant difference among other models ($P > 0.05$). There was a statistically significant difference in the ROC curve area between logistic regression model and decision tree model, between decision tree model and BP neural network, and between decision tree model and support vector machine ($P < 0.05$). There is no statistically significant difference among other models ($P > 0.05$). Through the distribution map of AUC, it can be found that the area under the curve of the decision tree is the largest, indicating that the effect is the best, as shown in Figure 6.

Combining the above model evaluation indicators, the decision tree model is the optimal model for classifying big data of hypo-MDS and AA in terms of model authenticity, reliability, and benefit evaluation.

5.2.2. Results for Test Set Samples. Combined with the above results, the logistic regression, decision tree, BP neural network, and support vector machine hyper-MDS and AA big data classification model are evaluated from three aspects: authenticity, reliability, and benefit. The results show that, in terms of the comparison of authenticity evaluation, logistic regression, decision tree, BP neural network, and SVM, the decision tree model has the best authenticity. In terms of

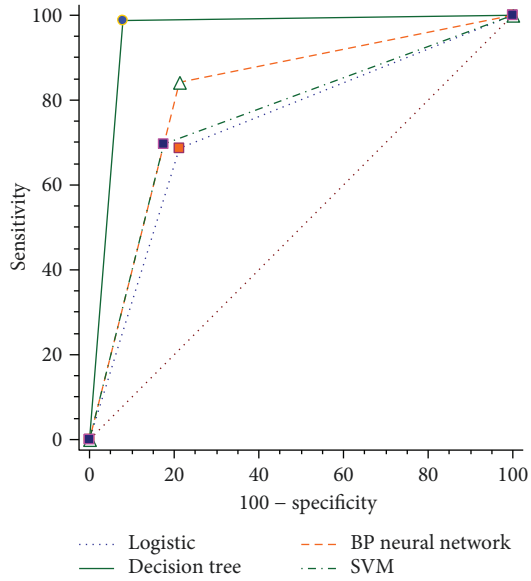

Figure 6: The ROC curve of four prediction models.

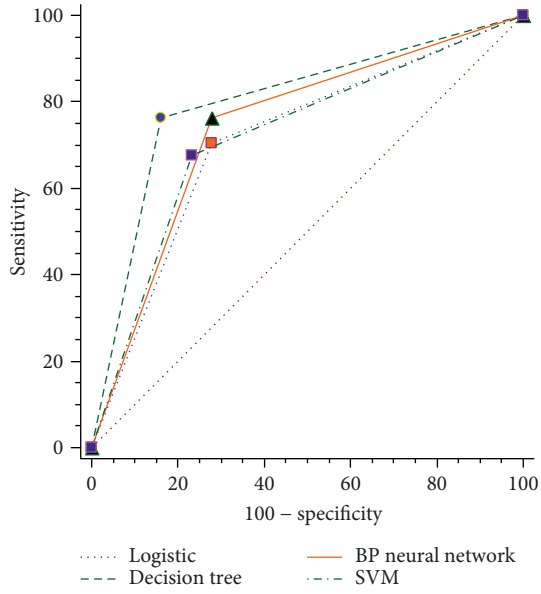

Figure 7: The ROC curve of four prediction models.

reliability evaluation, the reliability of the decision tree model is best compared with logistic regression, decision trees, BP neural networks, and SVM. In terms of model benefits, logistic regression, decision trees, BP neural networks, and SVM compare the decision tree models with the highest returns. After comparison, the sensitivity, specificity, accuracy, and area under the ROC curve of the four models were not statistically significant ($P > 0.05$) (see Figure 7). Although the results of the two comparisons show that the differences between the models are not statistically significant, the performance of the decision tree model is significantly better than the other three models in terms of various indicators of model evaluation. In summary, the decision tree model is the optimal model for classifying hypo-MDS and AA big data, both in terms of model authenticity, reliability, and benefit evaluation. Through the distribution map of AUC, it can be found that the area under the curve of the decision tree is the largest, indicating that the effect is the best, as shown in Figure 7.

5.3. Analysis of Cases of Hypojudgment of Hypo-MDS and AA. Through the model evaluation, we find that the decision tree model is the optimal classification model. Although the decision tree model has a good prediction effect, this model still has the potential to misjudge hypo-MDS and AA. Therefore, it is more conducive to the differential diagnosis of these two diseases of the in-depth analysis of misdiagnosed cases.

5.3.1. Hypo-MDS Misjudgment Case Analysis. The optimal model decision tree model classified 130 patients with hypo-MDS and classified 13 patients with hypo-MDS as AA patients. Comparing the misjudgment cases with the positive cases, it was found that the red blood cell content and hemoglobin content in the misjudged cases in the peripheral blood cell count were higher than the positive cases. The proportion of mature lymphocytes in misdiagnosed cases in bone marrow smear is higher than that in positive cases. The proportion of early erythroblasts and late erythroblasts was lower than that of positive culprit cases, and the difference was statistically significant ($P < 0.05$). There was no significant difference among other indicators ($P > 0.05$).

5.3.2. AA Misjudgment Case Analysis. The optimal model decision tree model classified 156 patients with AA, and 15 patients with AA were misclassified as hypo-MDS patients. Comparing the erroneously judged case with the positive case, it was found that the erythrocyte content and hemoglobin content in the erroneously judged cases in the peripheral blood cell count were lower than the positive case. The proportion of early erythroblasts, the ratio of red blood cells to young erythroblasts, and the proportion of late erythroblasts in misdiagnosed cases in bone marrow smears are higher than that in positive cases. The proportion of mature lymphocytes was lower than that of positive cases, and the difference was statistically significant ($P < 0.05$). However, there was no significant difference in other indicators ($P > 0.05$).

6. Conclusion

According to the analysis of basic patient data and disease index data, the difference in age and occupational composition between patients with hypo-MDS and AA was statistically significant ($P < 0.05$). There was no significant difference in other basic data ($P > 0.05$). For training set, logistic regression, BP neural network, support vector machine and decision tree sensitivity, Youden index, positive likelihood ratio, classification accuracy, positive predictive value, and negative predictive value were evaluated. There was a statistically significant difference in sensitivity between logistic regression model and decision tree model and between decision tree model and support vector machine ($P < 0.05$). The specificity, accuracy, and area under ROC curve between decision tree model and logistic regression model, decision tree model and BP neural network, and

decision tree model and support vector machine were statistically significant ($P < 0.05$). For the test set, logistic regression, BP neural network, support vector machine and decision tree sensitivity, Youden index, positive likelihood ratio, classification accuracy, positive predictive value, negative predictive value, the sensitivity, specificity, accuracy, and area under the ROC curve of the four models were not statistically significant ($P > 0.05$).

The classification effects of logistic regression, decision tree, BP neural network, and support vector machine are compared. The decision tree algorithm has the best classification effect on hypo-MDS and AA, which can help the clinicians to identify and diagnose the two diseases.

Data Availability

From January 1st, 2008, to December 31st, 2016, the patients diagnosed with hypo-MDS and AA were diagnosed at the First Affiliated Hospital of Chinese Academy of Medical Sciences and the Affiliated Hospital of North China University of Science and Technology. All the cases were completely recorded.

Conflicts of Interest

All the authors do not have any possible conflicts of interest.

Acknowledgments

This study was funded by Hebei Provincial Natural Science Foundation (H2017209172) (to Jianhui Wu).

References

[1] J. R. Krause, *WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues*, IARC Press, Lyon, 2008.

[2] J. W. Vardiman, J. Thiele, D. A. Arber et al., "The 2008 revision of the World Health Organization (WHO) classification of myeloid neoplasms and acute leukemia: rationale and important changes," *Blood*, vol. 114, no. 5, pp. 937–951, 2009.

[3] J. Shi and Y. Z. Zheng, "Thoughts on differential diagnosis between aplastic anemia and hypoplastic myelodysplastic syndrome," *Chinese Journal of Hematology*, vol. 34, no. 10, pp. 910–912, 2013.

[4] S. B. Killick, N. Bown, J. Cavenagh et al., "Guidelines for the diagnosis and management of adult aplastic anaemia," *British Journal of Haematology*, vol. 172, no. 2, pp. 187–207, 2015.

[5] J. Huang, M. F. Deng, Y. L. Chen, Y. Y. Tang, and Z. P. Huang, "Diagnosis and differential diagnosis between hypoplastic myelodysplastic syndrome (Hypo-MDS) and aplastic anemia (AA)," *Chinese Journal of Health Laboratory Technology*, vol. 24, no. 16, pp. 2371–2373, 2014.

[6] R. Hast, M. Eriksson, S. Widell, I. Arvidsson, and P. Bemell, "Neutrophil dysplasia is not a specific feature of the abnormal chromosomal clone in myelodysplastic syndromes," *Leukemia Research*, vol. 23, no. 6, pp. 579–584, 1999.

[7] A. Rashid, M. Khurshid, U. Shaikh, and S. Adil, "Chromosomal abnormalities in primary myelodysplastic syndrome," *Journal of the College of Physicians and Surgeons Pakistan*, vol. 24, no. 9, pp. 632–635, 2014.

[8] L. Wu, W. Shi, X. Li et al., "High expression of the human equilibrative nucleoside transporter 1 gene predicts a good response to decitabine in patients with myelodysplastic syndrome," *Journal of Translational Medicine*, vol. 14, no. 1, p. 66, 2016.

[9] D. C. de Souza, C. de Souza Fernandez, A. Camargo et al., "Cytogenetic as an important tool for diagnosis and prognosis for patients with hypocellular primary myelodysplastic syndrome," *BioMed Research International*, vol. 2014, no. 1, Article ID 542395, 10 pages, 2014.

[10] J. Huang, M. Ge, S. Lu et al., "Impaired autophagy in adult bone marrow $CD34^+$ cells of patients with aplastic anemia: possible pathogenic significance," *PLoS One*, vol. 11, no. 3, article e0149586, 2016.

[11] F. Jiang, Y. Y. Wang, J. N. Cen et al., "Autophagy activity and clinical significance of CD34(+) cells in myelodysplastic syndromes," *Chinese Journal of Experimental Hematology*, vol. 24, no. 3, pp. 779–783, 2016.

[12] A. A. van de Loosdrecht, C. Alhan, M. C. Bene et al., "Standardization of flow cytometry in myelodysplastic syndromes: report from the first European LeukemiaNet working conference on flow cytometry in myelodysplastic syndromes," *Haematologica*, vol. 94, no. 8, pp. 1124–1134, 2009.

[13] F. Lu, H. Bi, M. Huang, and S. Duan, "Simulated annealing genetic algorithm based schedule risk management of IT outsourcing project," *Mathematical Problems in Engineering*, vol. 2017, Article ID 6916575, 17 pages, 2017.

[14] Y. Le Manach, G. Collins, R. Rodseth et al., "Preoperative score to predict postoperative mortality (POSPOM)," *Anesthesiology*, vol. 124, no. 3, pp. 570–579, 2016.

[15] F. M. Santin, R. V. da Silva, and J. M. V. Grzybowski, "Artificial neural network ensembles and the design of performance-oriented riparian buffer strips for the filtering of nitrogen in agricultural catchments," *Ecological Engineering*, vol. 94, pp. 493–502, 2016.

[16] M. Stoia, Z. Kurtanjek, and S. Oancea, "Reliability of a decision-tree model in predicting occupational lead poisoning in a group of highly exposed workers," *American Journal of Industrial Medicine*, vol. 59, no. 7, pp. 575–582, 2016.

[17] V. Agarwal, S. Thakare, and A. Jaiswal, "Survey on classification techniques for data mining," *International Journal of Computer Applications*, vol. 132, no. 4, pp. 13–16, 2015.

[18] M. Heydari, M. Teimouri, Z. Heshmati, and S. M. Alavinia, "Comparison of various classification algorithms in the diagnosis of type 2 diabetes in Iran," *International Journal of Diabetes in Developing Countries*, vol. 36, no. 2, pp. 167–173, 2016.

[19] Y. W. Lui, Y. Xue, D. Kenul, Y. Ge, R. I. Grossman, and Y. Wang, "Classification algorithms using multiple MRI features in mild traumatic brain injury," *Neurology*, vol. 83, no. 14, pp. 1235–1240, 2014.

[20] W.-T. Tseng, W.-F. Chiang, S.-Y. Liu, J. Roan, and C.-N. Lin, "The application of data mining techniques to oral cancer prognosis," *Journal of Medical Systems*, vol. 39, no. 5, pp. 59–57, 2015.

[21] J. H. Wu, G. L. Wang, X. M. Li, and S. F. Yin, "Comparison of BP neural network model and logistic regression in the analysis of influencing factors of violence in hospitals," *Applied Mechanics and Materials*, vol. 50-51, pp. 964–967, 2011.

[22] Z. N. Zhang, *Hematological Diagnosis and Efficacy Criteria*, Science Press, 2007.

[23] X. Wang, *Clinical Comparison of Hypoproliferative Myelodysplastic Syndrome and Aplastic Anemia*, Jilin University, 2007.

[24] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.

[25] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 21, no. 3, pp. 660–674, 1991.

[26] F. Q. Lu, M. Huang, W. K. Ching, and T. K. Siu, "Credit portfolio management using two-level particle swarm optimization," *Information Sciences*, vol. 237, no. 13, pp. 162–175, 2013.

[27] S. Shan, "Decision tree learning," in *Machine Learning Models and Algorithms for Big Data Classification*, pp. 1–28, Springer, US, 2016.

[28] S. H. Hwang, D. H. Ham, and J. H. Kim, "Forecasting performance of LS-SVM for nonlinear hydrological time series," *KSCE Journal of Civil Engineering*, vol. 16, no. 5, pp. 870–882, 2012.