Array 14 (2022) 100140

Contents lists available at ScienceDirect

Array

journal homepage: www.elsevier.com/locate/array

GARD: Gender difference analysis and recognition based on machine learning☆

Shiwen He [a,b,c,*], Jian Song [a], Yeyu Ou [a], Yuanhong Yuan [d,**], Xiaojie Zhang [e,f], Xiaohua Xu [g]

[a] School of Computer Science and Engineering, Central South University, Changsha 410083, China
[b] National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China
[c] The Purple Mountain Laboratories, China
[d] Emergency Center of Hunan Children's Hospital, Changsha 410007, China
[e] Department of Psychiatry, The Second Xiangya Hospital, Central South University, Changsha 410011, China
[f] National Clinical Research Center for Mental Disorders, Changsha, 410011, China
[g] Department of Endocrinology, Nanjing First Hospital, Nanjing Medical University, Nanjing 210012, China

ARTICLE INFO

Keywords:
Gender difference analysis
Gender recognition
Medical examination data
Machine learning

ABSTRACT

In recent years, intelligent diagnosis and intelligent medical treatment based on big data of medical examinations have become the main trend of medical development in the future. In this paper, we propose a method for analyzing the difference between males and females in medical examination items (medical attributes) and find that males and females of different ages have differences in medical attributes. Then, the cluster analysis method is used to further analyze the differences between male and female in medical examination items, such that some common important attributes (CIAs) that can be used for gender recognition are found within a specific age range. Following, we propose two gender recognition models (GRMs) by using the found CIAs to identify the gender. A large number of experimental results are provided to validate the effectiveness of the proposed GRMs. Experimental results show that the medical attributes with a large value of difference really contribute to gender recognition. Within a certain age range, such as 17 to 51 years old, the proposed GRM can reach 92.8% accuracy using only six medical attributes.

1. Introduction

With the rapid development of computer technology and hospital information systems, electronic medical record (EMR) has been popularized to replace the traditional handwritten medical records [1]. Furthermore, the establishment of an advanced EMR system will bring the hospital into a new era of digital hospitals and provide proactive, convenient, and efficient data services for the hospital's medical, scientific research and teaching as well as hospital management. However, during collecting the EMR data, the values of some medical examination items cannot be avoided to lose, such as the value of gender. For EHR analysis, patient ender information plays a very essential role in referring some useful information, and the lack of this information will affect the data quality of EMR. However, in some Chinese EHRs, gender information is either missing [2], or it is hidden and deleted due to privacy concerns [3]. Therefore, it is necessary to investigate the filling miss value algorithm with the non-missing value of medical data, and predict gender from data that does not involve privacy risks.

It is well-known that accurate analysis of medical examination data, which depends on data integrity, is conducive to early disease detection, patient care, and community service. However, the incompleteness of medical examination data will reduce the analysis accuracy. Yoon et al. proposed a generative adversarial imputation network (GAIN) for imputing missing data by adopting the well-known generative adversarial network (GAN) framework [4]. Yang et al. pointed out that GAIN is not suitable for analyzing medical examination data, because GAN itself adapts to pixel data, so they combined fuzzy coding to adjust the GAIN [5]. As most of the medical examination data in the intensive care unit are time series, Luo et al. combined GAN

☆ This work was supported in part by National Natural Science Foundation of China under Grants 62171474, in part by the open research fund of National Mobile Communications Research Laboratory Southeast University under Grants No. 2022D03, in part by OPPO research fund under Grants No. CN05202112160224.

* Corresponding author at: School of Computer Science and Engineering, Central South University, Changsha 410083, China.
** Corresponding author.
E-mail addresses: shiwen.he.hn@csu.edu.cn (S. He), 1301140117@csu.edu.cn (J. Song), Ouyeyu@csu.edu.cn (Y. Ou), yyhong120@163.com (Y. Yuan), Xiaojiezhang2014@163.com (X. Zhang), xxh7812@163.com (X. Xu).

https://doi.org/10.1016/j.array.2022.100140
Received 8 January 2022; Accepted 7 March 2022
Available online 25 March 2022

with the gated recurrent unit to construct the gated recurrent unit for data imputation for filling missing values in time series data [6]. Similarly, the time dimension information is used to fill missing values by multidirectional recurrent neural network [7]. Machine learning and deep learning are also widely used to fill in missing values in EMR. McLean et al. introduced Predictive Mean Matching, K Nearest Neighbors, Iterative Imputer, and MICE to generate values for missing data in EMR centered on oncology [8]. Farnaz et al. developed a time series medical generative adjunctive network (GANs) for a wound prognosis model that can generate continuous and categorical features from EMR [9].

When solving the problem of data incompleteness, filling gender information, i.e., gender recognition is also a research hotspot. For gender identification, Li et al. introduced a multiple support vector machine (SVM) gender classifier based on five facial features in addition to human hair and clothes [10]. The Gabor filter responses of face images were used to identify gender and age [11]. A RestNet based gender recognizer was developed by detecting and locating the face landmarks on the regression tree set [12]. In addition to face images, there are studies that use text, mobile phone behaviors, and neural signals to distinguish males and females. Moumita et al. effectively identified the male and female by extracting the Euler Numbers of names [13]. The gender information of users is predicted by matching the mobile text data of users with the gender representative word set based on Web documents [14]. Dongxu et al. used reposing behaviors on social networks to distinguish male and female by combining statistical knowledge and sociological knowledge [15]. Kaur et al. tried to mine useful information from neural signals, they designed a prediction framework to use neural signals collected by brain wave sensors to identify age and gender [16]. In the medical field, Junmei et al. developed a data management method using convolutional neural network(CNN) to predict missing values of patient gender from EMR [2]. Serkan et al. proposed a machine learning algorithm based on multidetector computed tomography (MDCT) image measurement of the patella, using a decision tree (DT) method to determine gender [17]. Yuichiro et al. used brain MRI(magnetic resonance imaging) to construct a multichannel 3D-CNN and verified the effect of brain structure image patterns on gender recognition [18].

In this paper, we first analyze the differences between males and females in medical examination items, and then build two gender recognition models based on the results of the analysis to identify males and females. The main contributions are summarized as follows:

1. An algorithm based on overlapping areas is proposed to analyze the difference between males and females in medical examination;
2. Some common important attributions (CIAs) in a certain age ranges are found to facilitate the recognition of males and females via using cluster analysis, and an age range division algorithm is proposed as a by-product;
3. Two gender recognition models (GRMs), i.e., GRM A for the sample to be identified with known age, and GRM B for the sample to be identified with unknown age, are developed to identify the gender of samples. Experimental results show that in certain age ranges, such as 17 to 51 years old, the proposed GRM A and GRM B can reach 92.8% and 93.8% accuracy, respectively.

The rest of the paper is organized as follows. In Section 2, we describe the dataset and analyze the differences between males and females in medical examination. In Section 3, we further analyze through cluster analysis and determine the medical examination items that play an important role in gender recognition in each age subrange. In Section 4, two gender recognition models are proposed according to whether the age of the examples is known or not. Then, we evaluate the proposed models and the role of data analysis in gender recognition in Section 5. Finally, we conclude in Section 6.

2. Data description and analysis

2.1. Data description

To analyze the differences between males and females in medical examination data, we collected 62,072 samples from the Xiangya Medical Big Data System after desensitization. For ease of presentation, let D be the set of samples. Each sample contains $N = 39$ medical examination items (medical attributes) and 2 basic attributes ("AGE": Age, "SEX": Sex). In particular, Table 1 lists the medical attribute types, index, attributes, and full name of attributes. The medical attributes include the routine blood test, liver function test, renal function test, and examination of plasma prothrombin time measurements. To grasp the data structure in set D, we list the common statistics of each attribute, including the mean value (Mean), minimum (MIN), and maximum (MAX). For the basic attributes, the minimum age and the maximum age of samples in set D are 1 and 70, respectively, i.e., the age range is $\mathcal{T} = \{1, 2, \ldots, 69, 70\}$. For the male samples, the value of attribute "SEX" is "0", while for the female samples, the value of attribute "SEX" is "1".

2.2. Difference analysis

The existing studies have shown that there are differences between males and females in some medical attributes listed in Table 1. For example, the medical attributes RBC, HGB, and HCT have a significant differences between males and females after 2 years old [19,20]. Through the statistical analysis of samples between 1 and 17 years old, CREA and UA have a significant differences between males and females within 12 to 17 years old [21]. However, the statistical methods used in the aforementioned literature can only determine whether the distributions of medical attributes γ of males and females are the same or not. If they are not the same, the Difference between malEs and Females in mEdical Attribute (DEFEA) γ is considered significant. However, these aforementioned literature has not revealed the change trend of DEFEA γ, what we only know is there is the DEFEA γ. In other words, they do not compare the magnitude of DEFEA γ. In short, these statistical analysis cannot tell us the specific value of DEFEA γ [22]. In what follows, we analyze the DEFEA γ through the distribution of medical attribute γ of male and female, respectively.

Fig. 1 shows the numerical distribution of three medical attributes at different ages of males and females. Specifically, Figs. 1(a), 1(b), and 1(c) illustrate the numerical distribution of medical attributes CREA, HGB, and PT%, respectively. It is not difficult to find that for medical attributes CREA and HGB, the numerical distribution of males and females is almost the same at the age of 5. However, there are large differences between the numerical distribution of males and females over 5 years old. For the medical attribute PT%, the difference between the numerical distribution of males and females is not obvious at these ages. From the statistical viewpoint, for the medical attribute γ, the closer the numerical distribution of males and females is, the larger the overlap area of the numerical distribution is. This implies that the DEFEA γ at age t can be evaluated with the overlap area $\mathbf{O}(t, \gamma)$ of the numerical distribution of medical attribute γ of males and females at age t, given by

$$\mathbf{D}(t, r) = -\ln(\mathbf{O}(t, r)), \tag{1}$$

where r represents the index of the corresponding medical attribute γ, $r \in \mathcal{R} = \{1, 2, \ldots, 39\}$, and $\mathbf{O}(t, \gamma)$ is calculated by

$$\mathbf{O}(t, r) = \sum_{j \in \Lambda_\gamma} \min\left(P_m(t, r, j), P_f(t, r, j)\right), \tag{2}$$

where Λ_γ is a set of subintervals generated by dividing uniformly the value range of medical attribute γ into multiple intervals, $P_m(t, r, j)$ and $P_f(t, r, j)$ represent the frequency of samples of medical attribute γ of males and females belonging to the jth subinterval at age t,

Table 1
Data structure.

Attribute types	Index	Attribute	Full name of attribute	Mean	MIN	MAX	Unit
Blood routine examination	1	RBC	Erythrocyte count	4.46	3.27	5.60	10^12/L
	2	HCT	Hematocrit	40.05	29.40	49.00	%
	3	MCV	Erythrocyte mean corpuscular volume	89.97	73.40	100.00	fL
	4	MCH	Mean corpuscular hemoglobin	29.52	22.30	33.10	Pg
	5	MCHC	Mean corpuscular hemoglobin concentration	328.27	301.00	351.00	g/L
	6	RDW-CV	Red blood corpuscular volume distribution width-CV	12.77	11.60	17.00	%
	7	WBC	Leukocyte count	6.29	3.54	12.63	10^9/L
	8	LYM#	Lymphocyte count	2.02	0.77	5.47	10^9/L
	9	LYM%	Lymphocyte concentration	32.80	10.50	61.70	%
	10	MONO#	Monocyte count	0.33	0.05	0.64	10^9/L
	11	MONO%	Monocyte concentration	5.27	0.70	8.30	%
	12	EO	Eosinophil concentration	2.50	0.10	9.00	%
	13	BASO	Basophil concentration	0.40	0.00	0.90	%
	14	PCT	Plateletcrit	0.25	0.13	0.42	%
	15	MPV	Mean platelet volume	10.96	9.00	13.50	fL
	16	HGB	Hemoglobin	131.48	92.00	164.00	g/L
	17	PLT	Blood platelet count	231.93	111.00	425.00	10^9/L
	18	P_LCR	Platelet-larger cell ratio	32.50	16.00	52.70	%
	19	PDW	Platelet distributionwidth	13.27	9.10	20.10	%
	20	EO#	Eosinophil count	0.15	0.01	0.62	10^9/L
	21	BASO#	Basophil count	0.02	0.00	0.06	10^9/L
	22	RDW-SD	Red blood corpuscular volume distribution width-SD	12.77	11.60	17.00	%
Liver function test	23	TBA	Total bile acid	4.46	0.10	15.50	μmol/L
	24	TBIL	Total bilirubin	10.25	3.60	24.90	μmol/L
	25	TP	Total Protein	66.98	55.60	78.90	g/L
	26	GLO	Globulin	26.59	19.00	36.00	g/L
	27	A/G	Albumin globulin ratio	1.54	0.98	2.18	
	28	ALB	Albumin	40.39	30.90	47.90	g/L
	29	ALT	Alanine aminotransferase	16.77	3.30	52.90	μ/L
	30	AST	Aspartate aminotransferase	20.03	5.30	45.00	μ/L
	31	AST/ALT	Aspartate aminotransferase/alanine aminotransferase	1.38	0.52	3.24	
	32	DBIL	Direct Bilirubin	3.20	0.10	7.00	μmol/L
Renal function test	33	CREA	Creatinine	61.02	22.30	113.90	μmol/L
	34	BUN	Urea nitrogen	4.96	2.50	9.30	mmol/L
	35	UA	Uric acid	297.26	162.40	500.00	μmol/L
Plasma prothrombin time detection	36	PT_sec	Prothrombin time	11.94	10.00	14.40	s
	37	PT%	Prothrombin time activity	114.34	76.00	166.00	%
	38	PT_Ratio	Prothrombin time ratio	0.95	0.82	1.14	
	39	INR	International normalized ratio	0.95	0.78	1.17	
Personal information	40	SEX	Sex	0.56	0	1	
	41	AGE	Age	41	1	70	

respectively. Note that $\mathbf{D}(t,r)$ is negatively correlated with $\mathbf{O}(t,r)$, and approximates 0 when $\mathbf{O}(t,r)$ is close to 1. In other words, the overlap of the two numerical distributions means that there is no difference between females and males in a medial attribute.

Fig. 2 shows the change of the value of DEFEA as age increases. It can be seen that for most medical attributes, the value of DEFEA is smaller than 0.5 at all ages and fluctuates with age. However, with an increasing age, the value of DEFEA has obvious changes for some medical attributes, such as CREA, HGB, HCT, RBC, UA, AST/ALT, and ALT. In particular, one can see that for these medical attributes, the value of DEFEA grows rapidly from the age of 12, and reaches the highest point between the ages of 15 and 50 years old. After 50 years old, the values of DEFEA CREA, HGB, HCT, RBC, and UA are still large, while the values of DEFEA AST/ALT and ALT drop to a lower level compared with other medical attributes, such as MONO#, MCH, and LYM.

3. Cluster analysis

After the aforementioned analysis, one can find that only a few medical attributes, such as CREA, ALT, have an obvious differences between males and females in a certain age range in terms of the value of DEFEA. The difference between the values of other medical attributes is not obvious at each age. Can we find medical attributes, which play an important role in distinguishing males and females with a higher value of DEFEA at each age? To answer this question, in what follows, we focus on designing an effective method to divide age range $\mathcal{T}$ into an age subrange $\mathcal{C}_i$, i.e., $\cup_i \mathcal{C}_i = \mathcal{T}$. As a result, in the age subrange $\mathcal{C}_i$, some common medical attributes can be used to distinguish males and females. Note that to reduce the computational complexity and to take into account the fact that the age itself is ordered, the age contained in age subrange $\mathcal{C}_i$ should be adjacent, such as $\mathcal{C}_i = \{1,2,3\}$.

For easy of presentation, we define the mean value of the values of DEFEAs at age t as

$$\pi(t) = \frac{\sum_{r\in\mathcal{R}} \mathbf{D}(t,r)}{N}. \tag{3}$$

If $\mathbf{D}(t,r)$ is higher than $\pi(t)$, the medical attribute γ is regarded as an important attribute (IA) in distinguishing males and females at age t. We introduce a matrix $\mathbf{M}$ whose element is given by

$$\mathbf{M}(t,r) = \begin{cases} 1, & \text{if } \mathbf{D}(t,r) > \pi(t) \\ 0, & \text{other,} \end{cases} \tag{4}$$

where $t \in \mathcal{T}$, $r \in \mathcal{R}$. The cohesion of IAs in the age subrange $\mathcal{C}_i$ is defined as

$$\Phi(i) = \sum_{t_1,t_2\in\mathcal{C}_i} H(t_1,t_2), \tag{5}$$

where $H(t_1,t_2)$ is calculated by

$$H(t_1,t_2) = \sum_{r\in\mathcal{R}} \left|\mathbf{M}(t_1,r) - \mathbf{M}(t_2,r)\right|. \tag{6}$$

As can be seen from Eqs. (5) and (6), the smaller the cohesion, the larger similarity of IAs in age subrange $\mathcal{C}_i$.

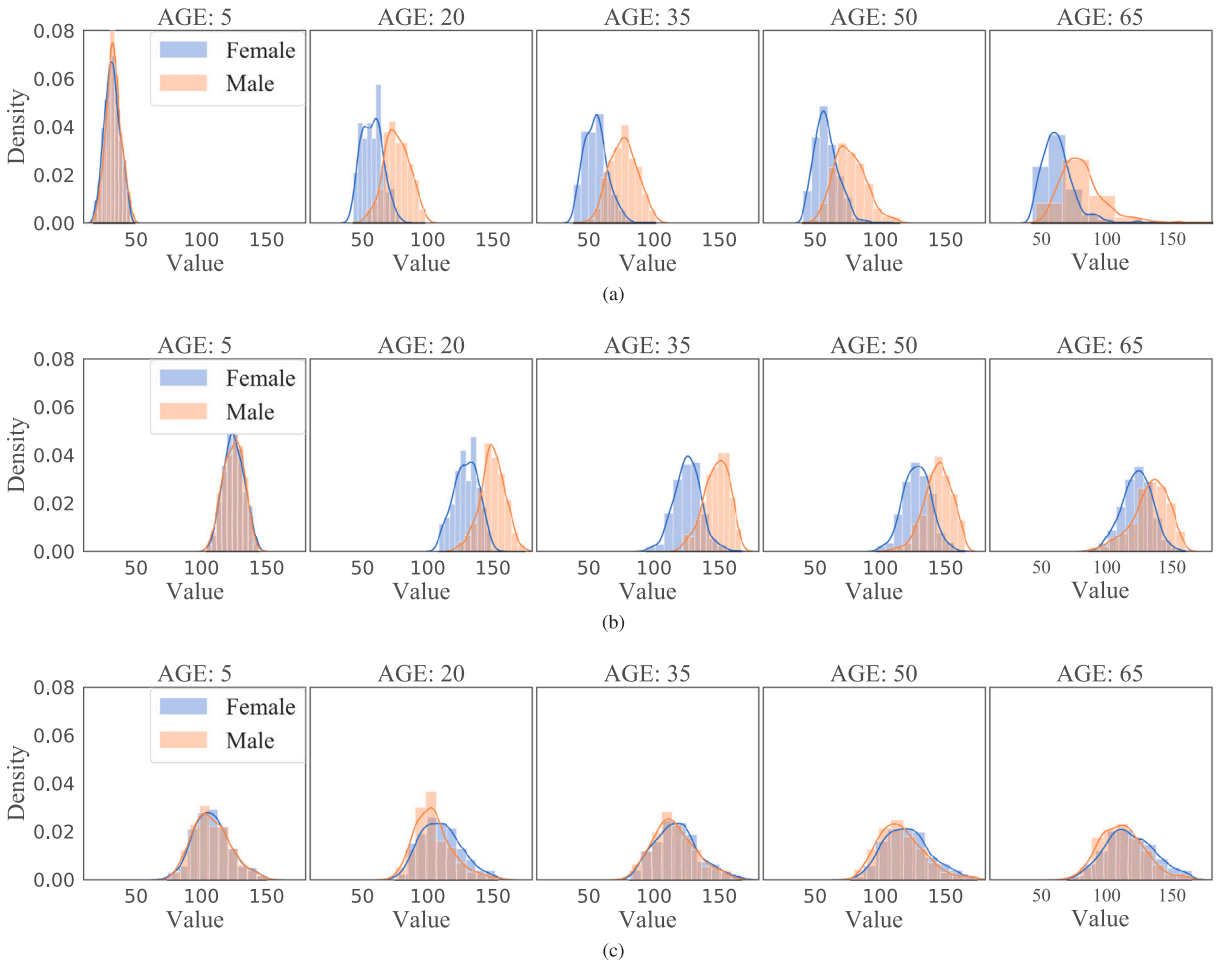

Fig. 1. Numerical distribution of (a) CREA, (b) HGB, (c) PT% in different ages.

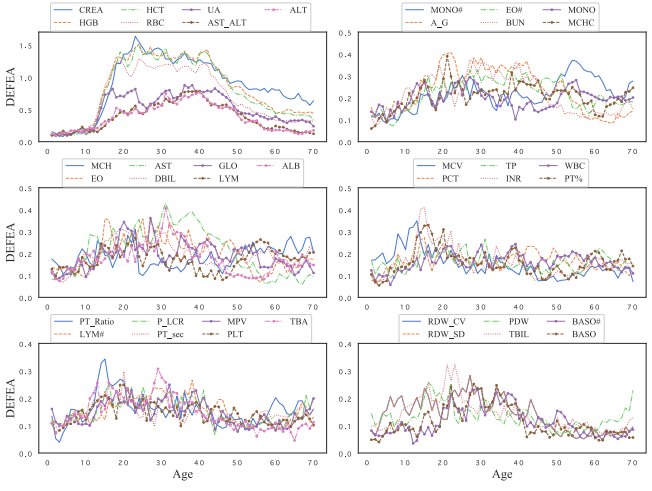

Fig. 2. The value of DEFEAs.

By using the concept of cohesion, we use division hierarchical clustering (DHC) to divide the age range $\mathcal{T}$ into some age subranges with the goal of minimizing the sum of cohesion of age subranges, such that the IAs between ages in age subrange C_i is as the same as possible. The detailed DHC is summarized as Algorithm 1. To illustrate the process of running Algorithm 1, Fig. 3 gives out an example of the number of age subranges is 10. One can see that when $n = 10$, each age subrange contains a small number of age before the age of 16, and contains more ages after the age of 16. This implies that the common IAs changes frequently between the ages of 1 and 16 years, and tends to be stable after the age of 16. Furthermore, the IAs of different ages within the same age subrange obtained by DHC may also be different. In the age subrange C_i, we regard the mean value of DEFEA γ at all ages as the value of DEFEA γ, given by

$$\mathbf{V}(i,r) = \frac{\sum_{t\in C_i} \mathbf{D}(t,r)}{|C_i|}, \tag{7}$$

where $r \in \mathcal{R}$, $|C_i|$ means the number of ages in the age subrange C_i. Similarly, for the age subrange C_i, we define the mean value of the values of DEFEAs as

$$\Pi(i) = \frac{\sum_{r\in\mathcal{R}} \mathbf{V}(i,r)}{N}. \tag{8}$$

If $\mathbf{V}(i,r)$ is higher than $\Pi(i)$, the medical attribute γ is regarded as a common important attribute (CIAs) in distinguishing males and females in the age subrange C_i.

Algorithm 1 Divisive Hierarchical Clustering

Input: n (the number of age subranges ($n \le 70$));
Output: Cell array $Cluster$, n cells of $Cluster$ represent n age subranges;
1: Using Eq. Eq. (4) to calculate $\mathbf{M}$;
2: Initialize cell array $Cluster = \left\{\{1,2,\cdots,70\}, \underbrace{\varnothing,\varnothing,\cdots,\varnothing}_{n-1}\right\}$;
3: **while** $Cluster\{n\} = \varnothing$ **do**

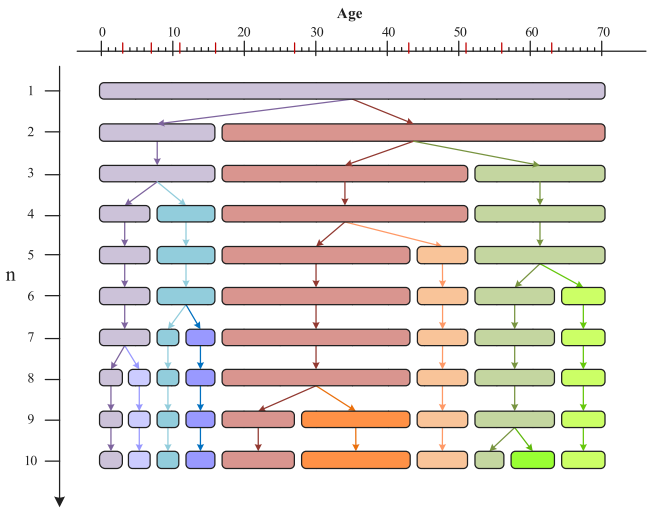

Fig. 3. Change of cohesion and number of IAs with division number.

Initialize array LC, the size of LC is n, the elements of LC are 0;
for $j = 1 : n$ **do**
  if $Cluster\{j\} \neq \varnothing$ **then**
    $LC(j) \leftarrow$the cohesion of $Cluster\{j\}$;
  end if
end for
$k \leftarrow \arg\max LC$;
$a \leftarrow \min Cluster\{k\}$;
$b \leftarrow \max Cluster\{k\}$;
Initialize array SC, the size of SC is $(b-a)$, the elements of SC are 0;
for $l = 1 : b - a$ **do**
  $SC(l) \leftarrow$the sum of the cohesion of $\{a, a+1, \cdots, a+l-1\}$ and $\{a+l, \cdots, b\}$;
end for
$u \leftarrow \arg\min SC + a - 1$;
for $i = n-1 : -1 : k+1$ **do**
  $Cluster\{i+1\} \leftarrow Cluster\{i\}$;
end for
$Cluster\{k\} \leftarrow \{a, a+1, \cdots, u\}$;
$Cluster\{k+1\} \leftarrow \{u+1, u+2, \cdots, b\}$;
end while
return $Cluster$

Through Algorithm 1, the age range $\mathcal{T}$ is divided into many age subranges, but excessive division is not necessary. To determine the optimal number of age subranges, two methods are used in this paper. One is to observe the changing trend of the sum of cohesion of all age subranges as the number of age subranges increases. The other is to observe the changing trend of the number of different CIAs in all age subranges as the number of age subranges increases. Fig. 4 illustrates the aforementioned two trends with an increasing number of age subranges. One observes that the total cohesion decreases with the number of age subranges increases and tends to saturation. While, the number of different CIAs in all age subranges increases with an increasing number of age subranges and also tends to saturation. To make the division results as accurate as possible, we need to make trade-offs. On the one hand, it is necessary to avoid unnecessary divisions. On the other hand, it is necessary to ensure that ages within the same age subrange have similar CIAs. Combining these two factors, in what follows, we divide age ranges $\mathcal{T}$ into 4 age subranges is a better

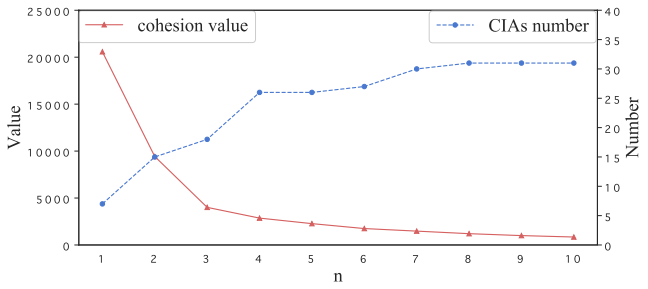

Fig. 4. Change of cohesion and number of CIAs with an increasing number of age subranges.

choice,[1] i.e., $C_1 = \{1, 2, \ldots, 7\}$, $C_2 = \{8, 9, \ldots, 16\}$, $C_3 = \{17, 18, \ldots, 51\}$, $C_4 = \{52, 53, \ldots, 70\}$. Table 2 lists the CIAs in age subrange C_i and the ratio of DEFEA γ calculated by

$$r(i,r) = \frac{\mathbf{D}(i,r)}{\sum_{k \in \mathcal{R}} \mathbf{D}(i,k)} \cdot 100\%. \tag{9}$$

In Table 2, there is a total of 27 different medical attributes. The number of CIAs varies greatly across age subranges. Specifically, there are 18 CIAs in the age subrange C_1, which is the most. The number of CIAs in the age subrange C_3 is 7, which is the least. In addition, the number of CIAs in age subrange C_2 and C_4 are 13 and 11, respectively. This implies that the values of DEFEA CIAs and the values of DEFEA non-CIAs in age subrange C_1 are not much different, on the contrary, the values of DEFEA CIAs in age subrange C_3 are much higher than that of non-CIAs, which is consistent with Fig. 2. For the most medical attributes, the value of DEFEA remains below 0.4 at any age, we can think that the smaller the number of CIAs in age subrange C_i, the higher the value of DEFEA of CIAs in age subrange C_i.

Note that the correlation between CIAs is not considered in Algorithm 1, therefore, the Pearson correlation method is firstly used to reduce some correlated medical attributes for facilitating the construction of GRMs. Generally speaking, if the absolute value of Pearson correlation coefficient corr (x, y) is greater than $\delta = 0.80$, the variables x and y have a strong correlation. Table 3 lists all medical attributes with the absolute value of Pearson correlation coefficients larger than δ. When corr (x, y) $> \delta$, we discard the medical attribute with a low value of DEFEA. For a more intuitive display, Table 2 shows whether the attribute is retained or not, in which "~~~~" means that the medical attribute is discarded, otherwise the medical attribute is used to construct the GRMs.

4. Gender recognition

To further assess the effectiveness of CIAs, in this section, according to whether the age is known or not, we construct two GRMs to distinguish the gender of samples. In particular, we study an effective GRM for case I in which the age is known. While, for the age is not known, called case II, we design another GRM. For the sake of explanation, we represent the reserved CIAs in the age subrange C_i as the set σ_i.

Case I: When the age is known, it is easy to know that the sample belongs to which of the four age subranges. This implies that we can use the CIAs of each age subrange to construct the corresponding GRM. Let the samples whose age belongs to age subrange C_i constitute a sample subset $\mathcal{X}_i$, $i \in \varpi = \{1, 2, 3, 4\}$. It is not difficult to find that the gender recognition problem belongs to two-classification problem. Therefore, for the age subrange C_i, the two-classification learning algorithm such as Logistic Regression (LR) [23], Random Forest (RF) [24],

[1] The proposed gender recognition can be used in other cases, i.e. other number of age subranges.

Table 2
CIAs and their DEFEA ratio.

Age subrange	CIAs and the DEFEA ratios
$C_1 = \{1, 2, \ldots, 7\}$	MCV (4.00%), MCH (3.14%), ALT (3.04%), MONO (3.01%), A_G (2.95%), RDW-SD (2.95%), RDW-CV (2.95%), EO (2.95%), TP (2.93%), AST/ALT (2.89%), HGB (2.87%), RBC (2.83%), PDW (2.82%), LYM (2.81%), MPV (2.80%), LYM# (2.74%), CREA (2.68%), BUN (2.62%)
$C_2 = \{8, 9, \ldots, 16\}$	RBC (5.44%), HGB (5.14%), HCT (4.53%), UA (4.35%), CREA (3.94%), MCV (3.34%), INR (3.01%), BUN (2.85%), ALT (2.76%), PT% (2.68%), AST (2.62%), PT_Ratio (2.58%), MONO (2.56%)
$C_3 = \{17, 18, \ldots, 51\}$	HGB (9.79%), CREA (9.77%), HCT (9.60%), RBC (8.26%), UA (5.51%), AST/ALT (4.52%), ALT (4.40%)
$C_4 = \{52, 53, \ldots, 70\}$	CREA (8.76%), HGB (6.02%), HCT (5.40%), RBC (4.34%), UA (4.11%), MONO# (3.50%), MONO (3.00%), MCH (2.86%), DBIL (2.75%), LYM (2.74%), MCHC (2.58%)

Table 3
Correlation coefficient of the removed medical attributes.

ID	Age subrange				x	y	corr (x, y)
	C_1	C_2	C_3	C_4			
1	✓				PDW	MPV	0.83
2	✓				RDW-CV	RDW-SD	1.00
3		✓			PT_Ratio	INR	0.96
4		✓	✓	✓	HCT	HGB	0.82

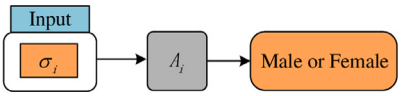

Fig. 5. Gender recognition process of GRM A.

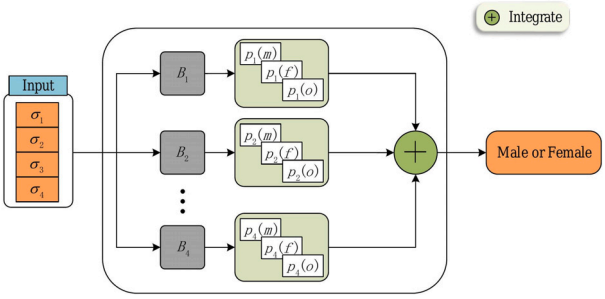

Fig. 6. GRM B.

can be used to construct the recognition model A_i. The recognition model A_i is trained with the samples in sample subset $\mathcal{X}_i$ by regarding the elements in the set σ_i as the features and the basic attribute SEX as the label of samples, respectively. Thus, all recognition models A_i constitute GRM A. When using GRM A for gender recognition, if the age of sample to be identified is within the age subrange C_i, the sample is identified by the corresponding recognition model A_i is shown in Fig. 5.

Case II: When the age is not known, we do not know which age subrange the age of sample to be identified belongs to, it means that the ideas of designing GRM A cannot be directly used. To effectively exploit the CIAs of each age subranges, in what follows, we design a three-classification learning for each age subrange. Then, the integration learning method is used to construct GRM B to identify the gender of the samples to be identified. Specifically, recognition model B_i for age subrange C_i has the ability of identifying the gender of the sample and judging the age range of the sample by outputting three probabilities $p_i(m)$, $p_i(f)$, and $p_i(o)$. $p_i(m)$ is the probability that the age of sample to be identified with the gender being males is within the age subrange C_i. $p_i(f)$ is the probability that the age of sample to be identified with the gender being females is within age subrange C_i. $p_i(o)$ is the probability that the age of sample to be identified is not in age subrange C_i. Unlike the training process of recognition model A_i, recognition model B_i is trained using samples at all ages. In particular, to train recognition B_i, the CIAs in set $\cup_{i\in\varpi}\sigma_i$ are regarded as features. At the same time, the males sample and females sample in sample subset $\mathcal{X}_i$ are labeled as "m" and "f", respectively, and the sample in set $\cup_{j\neq i}\mathcal{X}_j$ is labeled as "o". Then, the three probabilities of each age subrange are jointly used to identify the gender of samples via computing the integrated probabilities of males and females, i.e.,

$$\begin{aligned}\rho(m) &= \sum_{i\in\varpi}\frac{p_i(m)}{p_i(o)},\\ \rho(f) &= \sum_{i\in\varpi}\frac{p_i(f)}{p_i(o)}.\end{aligned} \tag{10}$$

If $\rho(m) > \rho(f)$, the gender of sample is males, otherwise is females. To clearly describe GRM B for case II, the detailed structure of GRM B is illustrated in Fig. 6. Note that even the age of samples to be identified is known, we also can use GRM B to identify the gender of samples via ignoring the age information. This implies that GRM B has a wider range of applications compared to GRM A.

5. Experiment

In this section, we evaluate the performance of the GRMs and discuss further the role of CIAs as well as the important of the division of age range. For each experiment, five learning algorithms, i.e., LR, linear discriminant analysis (LDA) [25], naive Bayes (NB) [26], RF, and gradient boosting decision tree (GBDT) [27], are used. In addition, for comparison, we also evaluate several simple GRM, i.e., GRM_0, GRM_1, and GRM_2. In particular, GRM_0 directly identifies the gender of the sample via two-classification method, i.e., without considering the age factor. GRM_0 is trained with samples at all ages and uses $\cup_{i\in\varpi}\sigma_i$ as the training feature. GRM_1 and GRM_2 regard the CIAs in $\cup_{i\in\varpi}\sigma_i$ and in $\cup_{j\neq i}\sigma_j$ as features to train GRM A_i, respectively. We take the recognition result of recognition model A_i as the result of GRM A (GRM_1 and GRM_2) in age subrange C_i. The recognition results of GRM B and GRM_0 in age subrange C_i are computed via the sample whose age belongs to age subrange C_i.

Table 4 shows the recognition accuracy of GRM A and GRM B. It can be seen that when we use the classical learning algorithm such as LR, LDA, and NB, there is a comparable performance difference between GRM A and GRM B in age subrange C_1 and C_2. Specifically, GRM A obtains about 6% accuracy gain compared to GRM B. However, there is a tiny difference between GRM A and GRM B in age subrange C_3 and C_4. When using learning algorithms with strong learning ability, such as RF and GBDT are adopted, the recognition accuracy of GRM B can reach the level of GRM A in age subrange C_1, and in age subrange C_3 and C_4, GRM B outperforms GRM A. Experimental results show that the lack of age information has a more severe impact on gender recognition in age subrange C_1 and C_2. In addition, no matter GRM A or GRM B, the accuracy of five algorithms is the lowest in age subrange C_1, and is the highest in age subrange C_3. This is because the smaller the number of CIAs in age subrange C_i, the larger the value of DEFEA in age subrange C_i. We also find that the smaller the number of CIAs in age subrange C_i,

Table 4
Recognition performance of GRM A and GRM B.

Model	Age subrange	Algorithm (Mean±Std of Accuracy)				
		LR	LDA	NB	RF	GBDT
GRM A	$C_1 = \{1, 2, \dots, 7\}$	0.6069 ± 0.0203	0.6254 ± 0.0206	0.6152 ± 0.0300	0.6134 ± 0.0230	0.6192 ± 0.0136
	$C_2 = \{8, 9, \dots, 16\}$	0.6999 ± 0.0356	0.7308 ± 0.0337	0.6807 ± 0.0287	0.7276 ± 0.0192	0.7295 ± 0.0273
	$C_3 = \{17, 18, \dots, 51\}$	0.9222 ± 0.0060	0.9276 ± 0.0049	0.9116 ± 0.0072	0.9234 ± 0.0059	0.9287 ± 0.0058
	$C_4 = \{52, 53, \dots, 70\}$	0.8419 ± 0.0117	0.8391 ± 0.0124	0.8235 ± 0.0108	0.8475 ± 0.0087	0.8460 ± 0.0077
GRM B	$C_1 = \{1, 2, \dots, 7\}$	0.5460 ± 0.0217	0.5351 ± 0.0159	0.5607 ± 0.0204	0.6094 ± 0.0222	0.5919 ± 0.0199
	$C_2 = \{8, 9, \dots, 16\}$	0.6475 ± 0.0209	0.6216 ± 0.0262	0.6324 ± 0.0275	0.6797 ± 0.0148	0.6541 ± 0.0259
	$C_3 = \{17, 18, \dots, 51\}$	0.9116 ± 0.0034	0.9236 ± 0.0034	0.9055 ± 0.0052	0.9371 ± 0.0068	0.9381 ± 0.0059
	$C_4 = \{52, 53, \dots, 70\}$	0.8361 ± 0.0061	0.8400 ± 0.0055	0.8104 ± 0.0071	0.8624 ± 0.0060	0.8614 ± 0.0059

Table 5
Comparison between GRM_0, GRM B and GRM_1.

Age subrange	Model	Algorithm (Mean±Std of Accuracy (Gain (%)))				
		LR	LDA	NB	RF	GBDT
$C_1 = \{1, 2, \dots, 7\}$	GRM_0	0.5365 ± 0.0205 (–)	0.5237 ± 0.0174 (–)	0.5543 ± 0.0234 (–)	0.5992 ± 0.0237 (–)	0.5871 ± 0.0190 (–)
	GRM B	0.5460 ± 0.0217 (1.77%)	0.5351 ± 0.0159 (2.18%)	0.5607 ± 0.0204 (1.15%)	0.6094 ± 0.0222 (1.70%)	0.5919 ± 0.0199 (0.82%)
	GRM_1	0.6069 ± 0.0248 (13.12%)	0.6217 ± 0.0224 (18.71%)	0.6120 ± 0.0312 (10.41%)	0.6051 ± 0.0150 (0.98%)	0.6203 ± 0.0177 (5.65%)
$C_2 = \{8, 9, \dots, 16\}$	GRM_0	0.6502 ± 0.0240 (–)	0.6390 ± 0.0215 (–)	0.6279 ± 0.0296 (–)	0.6789 ± 0.0198 (–)	0.6296 ± 0.0232 (–)
	GRM B	0.6475 ± 0.0209 (−0.42%)	0.6216 ± 0.0262 (−2.72%)	0.6324 ± 0.0275 (0.72%)	0.6797 ± 0.0148 (0.12%)	0.6541 ± 0.0259 (3.89%)
	GRM_1	0.6992 ± 0.0433 (7.54%)	0.7321 ± 0.0202 (14.57%)	0.6807 ± 0.0295 (8.41%)	0.7270 ± 0.0154 (7.08%)	0.7315 ± 0.0258 (16.18%)
$C_3 = \{17, 18, \dots, 51\}$	GRM_0	0.9033 ± 0.0039 (–)	0.9138 ± 0.0030 (–)	0.9017 ± 0.0067 (–)	0.9288 ± 0.0140 (–)	0.9285 ± 0.0077 (–)
	GRM B	0.9116 ± 0.0034 (0.92%)	0.9236 ± 0.0034 (1.07%)	0.9055 ± 0.0052 (0.42%)	0.9371 ± 0.0068 (0.89%)	0.9381 ± 0.0059 (1.03%)
	GRM_1	0.9390 ± 0.0078 (3.95%)	0.9428 ± 0.0045 (3.17%)	0.9155 ± 0.0057 (1.53%)	0.9377 ± 0.0044 (0.96%)	0.9413 ± 0.0060 (1.38%)
$C_4 = \{52, 53, \dots, 70\}$	GRM_0	0.8211 ± 0.0059 (–)	0.8244 ± 0.0044 (–)	0.8028 ± 0.0055 (–)	0.8627 ± 0.0068 (–)	0.8561 ± 0.0070 (–)
	GRM B	0.8361 ± 0.0061 (1.83%)	0.8400 ± 0.0055 (1.89%)	0.8104 ± 0.0071 (0.95%)	0.8624 ± 0.0060 (−0.03%)	0.8614 ± 0.0059 (0.62%)
	GRM_1	0.8678 ± 0.0151 (5.69%)	0.8656 ± 0.0126 (5.00%)	0.8367 ± 0.0175 (4.22%)	0.8677 ± 0.0086 (0.58%)	0.8674 ± 0.0105 (1.32%)

the better the recognition accuracy of GRMs in age subrange C_i. This shows that the medical attributes with a large value of DEFEA really contribute to gender recognition.

Table 5 gives out the comparison between GRM_0, GRM B, and GRM_1 to further evaluate the effectiveness of GRM B and analyze the role of age range division. Note that in most age subranges, GRM B with the five learning algorithms obtains up to 1% gain. However, in the age subrange C_2, GRM B does not show an overwhelming advantage. However, it is worth mentioning that if the learning algorithms NB, RF or GBDT are used, the recognition accuracy of GRM B is still higher than that of GRM_0 in age subrange C_2. This means that GRM B is effective even for age subrange C_2 which is most affected by the lack of age information. At the same time, it also shows that if we do not know the age of samples and divide the age range, we can still improve the accuracy of gender recognition. Let us look at the results of another group of control experiments, it can be seen that in all age subranges, GRM_1 outperforms GRM_0 in terms of the recognition accuracy for the five learning algorithms. In particular, in age subrange C_1 and C_2, the average gain is 10.3%, while in age subrange C_3 and C_4, the average gain is 2.8%. This means that if we know the age of samples, the age range division will improve the accuracy of gender recognition more, which is much higher than the case of unknown sample age. Furthermore, the division of age range is more helpful for gender recognition in age subrange C_1 and C_2. This is because the gender differences are different in different age subranges, and learning their gender difference information separately helps improve recognition accuracy. If the age range is not divided, GRM is more inclined to learn the gender difference information of medical attributions in age subrange C_3 and C_4 in which the gender difference is obvious.

Table 6 lists the comparison between GRM A, GRM_1, and GRM_2 to reflect the role of CIAs. We can see that in age subrange C_1 and C_2, GRM_1 has no significant improvement in terms of the recognition accuracy compared to GRM A. This means that for age subranges C_1 and C_2, the selection of CIAs is appropriate, and we have not left out any useful medical attributes for distinguishing males and females. For age subranges C_3 and C_4, the participation of more medical attributes leads to a small improvement in recognition accuracy. In particular, in age subrange C_3 and C_4, GRM_1 obtains about 2% gain in terms of the recognition accuracy at the cost of large computation complexity. This means that the CIAs cover the vast majority of useful information for distinguishing males and females. For example, in age subrange C_3, the five learning algorithms can achieve about 92% recognition accuracy with only six CIAs. In age subrange C_1 and C_2, compared to GRM A, GRM_2 has 10% recognition accuracy loss. While, the recognition accuracy loss will be more in age subrange C_3 and C_4, close to 20%. The values of DEFEA CIAs and the values of DEFEA non-CIAs in age subrange C_1 and C_2 are not much different, on the contrary, the values of DEFEA CIAs in age subrange C_3 and C_4 are much higher than that of non-CIAs, which leads to the above situation. The above two comparisons show that the CIAs are indeed more conducive to gender

Table 6
Controlled experiments on the role of CIAs.

Age subrange	Model (Attribute number)	Algorithm (Mean±Std of Accuracy (Gain (%)))				
		LR	LDA	NB	RF	GBDT
$C_1 = \{1, 2, \ldots, 7\}$	GRM A (16)	0.6069 ± 0.0203 (–)	0.6254 ± 0.0206 (–)	0.6152 ± 0.0300 (–)	0.6134 ± 0.0230 (–)	0.6192 ± 0.0136 (–)
	GRM_1 (23)	0.6069 ± 0.0248 (0.00%)	0.6217 ± 0.0224 (−0.59%)	0.6120 ± 0.0312 (−0.52%)	0.6051 ± 0.0150 (−1.35%)	0.6203 ± 0.0177 (0.18%)
	GRM_2 (7)	0.5138 ± 0.0225 (−15.34%)	0.5917 ± 0.0278 (−5.39%)	0.5895 ± 0.0295 (−4.18%)	0.5536 ± 0.0227 (−9.75%)	0.5783 ± 0.0259 (−6.61%)
$C_2 = \{8, 9, \ldots, 16\}$	GRM A (11)	0.6999 ± 0.0356 (–)	0.7308 ± 0.0337 (–)	0.6807 ± 0.0287 (–)	0.7276 ± 0.0192 (–)	0.7295 ± 0.0273 (–)
	GRM_1 (23)	0.6992 ± 0.0433 (−0.10%)	0.7321 ± 0.0202 (0.18%)	0.6807 ± 0.0295 (0.00%)	0.7270 ± 0.0154 (−0.08%)	0.7315 ± 0.0258 (0.27%)
	GRM_2 (12)	0.5969 ± 0.0522 (−14.72%)	0.6542 ± 0.0212 (−10.48%)	0.6420 ± 0.0214 (−5.69%)	0.6349 ± 0.0212 (−12.74%)	0.6426 ± 0.0271 (−11.91%)
$C_3 = \{17, 18, \ldots, 51\}$	GRM A (6)	0.9222 ± 0.0060 (–)	0.9276 ± 0.0049 (–)	0.9116 ± 0.0072 (–)	0.9234 ± 0.0059 (–)	0.9287 ± 0.0058 (–)
	GRM_1 (23)	0.9390 ± 0.0078 (1.82%)	0.9428 ± 0.0045 (1.64%)	0.9155 ± 0.0057 (0.43%)	0.9377 ± 0.0044 (1.55%)	0.9413 ± 0.0060 (1.35%)
	GRM_2 (17)	0.7444 ± 0.0151 (−19.28%)	0.7564 ± 0.0130 (−18.46%)	0.7345 ± 0.0151 (−19.43%)	0.7521 ± 0.0068 (−18.55%)	0.7545 ± 0.0086 (−18.76%)
$C_4 = \{52, 53, \ldots, 70\}$	GRM A (10)	0.8419 ± 0.0117 (–)	0.8391 ± 0.0124 (–)	0.8235 ± 0.0108 (–)	0.8475 ± 0.0087 (–)	0.8460 ± 0.0077 (–)
	GRM_1 (23)	0.8678 ± 0.0151 (3.08%)	0.8656 ± 0.0126 (3.16%)	0.8367 ± 0.0175 (1.60%)	0.8677 ± 0.0086 (2.38%)	0.8674 ± 0.0105 (2.53%)
	GRM_2 (13)	0.6630 ± 0.0134 (−21.25%)	0.6683 ± 0.0129 (−20.36%)	0.6554 ± 0.0136 (−20.41%)	0.7054 ± 0.0102 (−16.77%)	0.6724 ± 0.0122 (−20.52%)

recognition than other attributes and cover most of the information that can be used for gender recognition.

6. Conclusion

In this paper, we proposed a method for calculating the difference of medical attributes between males and females at a specified age via the distribution of medical attributes of males and females. We find that only a few medical attributes have an obvious differences between males and females. Then, we further analyzed the differences between males and females by cluster analysis, and found some CIAs in a certain age subrange. In addition, we proposed two GRMs based on the results of data analysis to identify the gender of samples according to whether the age is known or not. Experiment results have shown that in a certain age range, such as 17 to 51 years old, the proposed GRM can reach 92.8% accuracy using only six medical attributes. In addition to evaluating GRM recognition effect, we also verified the effectiveness of GRM B, cluster analysis, and CIAs through comparative experiments.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Hydari MZ, Telang R, Marella W. Saving patient ryan - Can advanced electronic medical records make patient care safer? Manage Sci 2018;65:2041–59.

[2] Junmei Z, Xiu Y, Jian W, Zhuquan S. Artificial intelligence based data governance for Chinese electronic health record analysis. Int J Data Min Knowl Manage Process 2018;8(3):29–41.

[3] Chen L, Yang J-J, Wang Q. Privacy-preserving data publishing for free text Chinese electronic medical records. In: IEEE annual computer software and applications conference. 2012, p. 567–72.

[4] Yoon J, Jordon J, van der Schaar M. GAIN: Missing data imputation using generative adversarial nets. In: International conference on machine learning, Vol. 80. 2018, p. 5689–98.

[5] Yang Y, Wu Z, Tresp V, Fasching PA. Categorical EHR imputation with generative adversarial nets. In: IEEE international conference on healthcare informatics (ICHI). 2019.

[6] Luo Y, Cai X, Zhang Y, Xu J, Yuan X. Multivariate time series imputation with generative adversarial networks. In: International conference on neural information processing systems. 2018, p. 1603–14.

[7] Yoon J, Zame WR, Der Schaar MV. Estimating missing data in temporal data streams using multi-directional recurrent neural networks. IEEE Trans Biomed Eng 2019;66(5):1477–90.

[8] McLean C, Ransom J, Galaznik A. PCN432 evaluation of missing data imputation strategies in clinical trial and emr data using standardized data models. Value Health 2019;22:S520.

[9] Foomani FH, Anisuzzaman D, Niezgoda J, Niezgoda J, Guns W, Gopalakrishnan S, Yu Z. Synthesizing time-series wound prognosis factors from electronic medical records using generative adversarial networks. 2021, arXiv preprint arXiv:2105.01159.

[10] Li B, Lian X, Lu B. Gender classification by combining clothing, hair and facial component classifiers. Neurocomputing 2012;76(1):18–27.

[11] Hosseini S, Lee SH, Kwon HJ, Koo HI, Cho NI. Age and gender classification using wide convolutional neural network and gabor filter. In: International workshop on advanced image technology. 2018, p. 1–3.

[12] Nie Y, Liang B, Huang P, Ren W, Dai J. A study on image based gender classification using convolutional neural network. In: International conference on deep learning technologies (ICDLT). Association for Computing Machinery; 2019, p. 81–4.

[13] pal M, Bhattacharyya S, Sarkar T. Euler number based feature extraction technique for gender discrimination from offline hindi signature using SVM & BPNN classifier. In: Emerging trends in electronic devices and computational techniques. 2018, p. 1–6.

[14] Choi Y, Kim Y, Kim S, Park K, Park J. An on-device gender prediction method for mobile users using representative wordsets. Expert Syst Appl 2016;64:423–33.

[15] Dongxu LI, Yongjun LI, Wenli JI. Gender identification via reposting behaviors in social media. IEEE Access 2017;PP(99):1.

[16] Kaur B, Singh D, Roy PP. Age and gender classification using brain-computer interface. Neural Comput Appl 2019;31(10):5887–900.

[17] Serkan O, Turan M, Zülal O. Estimation of gender by using decision tree, a machine learning algorithm, with patellar measurements obtained from MDCT images. Med Rec 2021;3(1):1–9.

[18] Nitta Y, Shinomiya Y, Park K, Yoshida S. A 3D-CNN classifier for gender discrimination from diffusion tensor imaging of human brain. In: International conference on soft computing and intelligent systems and 21st international symposium on advanced intelligent systems (SCIS-ISIS). 2021, p. 1–4.

[19] Zheng J, Hui Y, Fu Q. Investigation and analysis on reference value range of peripheral blood routine in preschool children from Shanghai. Int J Lab Med 2014;(16):2194–6.

[20] Min-Jie WU, Liang-Feng HU, Miao Y, Shen GJ, Hospital SP. Influence of age and gender on the red cell parameters in healthy people. Chin J Health Lab Technol 2016;26(11):1619–21.
[21] Clifford SM, Bunker AM, Jacobsen JR, Roberts WL. Age and gender specific pediatric reference intervals for aldolase, amylase, ceruloplasmin, creatine kinase, pancreatic amylase, prealbumin, and uric acid. Clin Chim Acta 2011;412(9):788–90.
[22] Casella G, Berger RL. Statistical inference, Vol. 2. Pacific Grove, CA: Duxbury; 2002.
[23] Hosmer Jr DW, Lemeshow S, Sturdivant RX. Applied logistic regression, Vol. 34. 2013, p. 358–9, no. 3.
[24] Breiman L. Random forests. Mach Learn 2001;45(1):5–32.
[25] Fukunaga K. Inroduction to statistical pattern recognition, Vol. 22. 1990, p. 70.
[26] Domingos P, Pazzani M. On the optimality of the simple Bayesian classifier under zero-oneloss. Mach Learn - ML 1997;29:103–30.
[27] Friedman JH. Greedy function approximation: A gradient boosting machine. Ann Statist 2001;29(5):1189–232.