Hindawi
Complexity
Volume 2018, Article ID 9193248, 17 pages
https://doi.org/10.1155/2018/9193248

Research Article

Application of Data Mining Technology on Surveillance Report Data of HIV/AIDS High-Risk Group in Urumqi from 2009 to 2015

Dandan Tang[1], **Man Zhang**[2], **Jiabo Xu**[3], **Xueliang Zhang**[4], **Fang Yang**[5], **Huling Li**[1], **Li Feng**[1], **Kai Wang**[4], **and Yujian Zheng**[1]

[1]*College of Public Health, Xinjiang Medical University, Urumqi 830011, China*
[2]*Department of Information Engineering, Xinjiang Institute of Engineering, Urumqi, 830000, China*
[3]*College of Medical Engineering and Technology, Xinjiang Medical University, Urumqi 830011, China*
[4]*Department of Medical Engineering, The Affiliated Tumor Hospital, Xinjiang Medical University, Urumqi 830011, China*
[5]*Department of AIDS/STD Control and Prevention, Urumqi Center for Disease Control and Prevention, Urumqi, Xinjiang 830026, China*

Correspondence should be addressed to Kai Wang; wangkaimath@sina.com and Yujian Zheng; 147854307@qq.com

Received 29 May 2018; Accepted 17 September 2018; Published 10 December 2018

Guest Editor: Panayiotis Vlamos

Objective. Urumqi is one of the key areas of HIV/AIDS infection in Xinjiang and in China. The AIDS epidemic is spreading from high-risk groups to the general population, and the situation is still very serious. The goal of this study was to use four data mining algorithms to establish the identification model of HIV infection and compare their predictive performance. *Method*. The data from the sentinel monitoring data of the three groups of high-risk groups (injecting drug users (IDU), men who have sex with men (MSM), and female sex workers (FSW)) in Urumqi from 2009 to 2015 included demographic characteristics, sex behavior, and serological detection results. Then we used *age*, *marital status*, *education level*, and other variables as input variables and whether to infect HIV as output variables to establish four prediction models for the three datasets. We also used confusion matrix, accuracy, sensitivity, specificity, precision, recall, and the area under the receiver operating characteristic (ROC) curve (AUC) to evaluate classification performance and analyzed the importance of predictive variables. *Results*. The final experimental results show that random forests algorithm obtains the best results, the diagnostic accuracy for random forests on MSM dataset is 94.4821%, 97.5136% on FSW dataset, and 94.6375% on IDU dataset. The k-nearest neighbors algorithm came out second, with 91.5258% diagnostic accuracy on MSM dataset, 96.3083% diagnostic accuracy on FSW dataset, and 90.8287% diagnostic accuracy on IDU dataset, followed by support vector machine (94.0182%, 98.0369%, and 91.3571%). The decision tree algorithm was the poorest among the four algorithms, with 79.1761% diagnostic accuracy on MSM dataset, 87.0283% diagnostic accuracy on FSW dataset, and 74.3879% accuracy on IDU. *Conclusions*. Data mining technology, as a new method of assisting disease screening and diagnosis, can help medical personnel to screen and diagnose AIDS rapidly from a large number of information.

1. Introduction

Acquired immunodeficiency syndrome (AIDS) is a malignant infectious disease with a very high fatality rate caused by human immunodeficiency virus (HIV) [1]. It alters the immune system making people much more vulnerable to infections and diseases [2]. Up to now, the HIV/AIDS epidemic has been one of the most important and crucial public health problems facing both developed and developing nations. Since the first case of HIV infection of China discovered in 1985, the number of the infected patients has been increasing year by year. The spread trend of AIDS in China has not been fundamentally controlled; AIDS prevention and control situation in Xinjiang is even severer. Xinjiang Uygur Autonomous Region is one of the provinces hardest hit by AIDS in China. The first HIV/AIDS case in Xinjiang was reported in 1995. At the end of 2011, the cumulative total of HIV/AIDS cases reported in Xinjiang accounted for

7.7% of all cumulative total of HIV/AIDS cases in the country, ranking the fifth position in China [3]. The total number of HIV/AIDS reported cases from 2004 to 2015 had been accumulated to 14,696, and it accounted for 5.56% of the total number of AIDS patients reported in China. There were also 3830 people died of HIV, which took up 4.56% of the total death cases induced by AIDS. The reported AIDS cases increased from 20 to 1868 with the average annual growth rate of 28.74, and the reported deaths increased from 5 to 680 with the average annual growth rate of 28.74 in the past decades, which were higher than that of the national average annual growth level [4]. Urumqi, the capital of Xinjiang Uygur Autonomous Region, is one of the main districts of AIDS infection in Xinjiang, and its AIDS epidemic has been consistently high. The largest group of HIV infection is injecting drug users in Urumqi. But in the late 2011, the proportion of the sexual route of transmission of infection is more than the intravenous drug users sharing syringes; the infection became the first way. More and more sexual partners, men and men crowd into the spread of AIDS high-risk groups [5, 6]. The situation of stemming the spread of HIV in persons at high risk of exposure and blocking the AIDS epidemic moving from high-risk groups to the general population proliferation is still very flinty. Therefore, HIV infection continues to be a major global public health issue.

Data mining is a newly developing technology based on machine study in artificial intelligence and database, and it can be classified into two categories: unsupervised learning and supervised learning [7]. Data mining is the process of selecting, exploring, and modeling large amounts of data, which aims at discovering unknown patterns or relationships and infer prediction rules from the data [8]. In the recent years, great advancement has been achieved in the medical research of data mining. Studies have applied data mining to analyze volumes of data, explore unknown factors of disease, develop predictive models, and produce meaningful reports in different medical research fields [9–11]. In the new period, the study of prevention, diagnosis, and treatment of HIV disease entered a new phase. A lot of domestic and foreign researchers have done on using the data mining technology to discover the relationship of the AIDS patient's potential factors and the result of treatment based on HIV surveillance data or comprehensive clinical data [12]. Oliveira et al. built multilayer artificial neural networks (MLP), naive Bayesian classifiers (NB), support vector machines (SVM), and the k-nearest neighbor algorithm (KNN) in order to identify the main factors influencing reporting delays of HIV-AIDS cases within the Portuguese surveillance system. The results of this study strongly suggested that MLP provided the best results, with a higher classification accuracy (approximately 63%), precision (approximately 76%), and recall (approximately 60%) [13]. Wang et al. had developed three computational modeling methods to predict virological response to therapy from HIV genotype and other clinical information. The comparison results showed that an artificial neural network (ANN) models were significantly inferior to random forests (RF) and support vector machines (SVM) [14]. Hai-Lei, et al. constructed a 133 HIV carriers forecasting model based on support vector machines (SVM), and the HIV carriers were found in the port of a province in China during the period of 2004–2009. The overall accuracy rate of forecasting model was 90.60%, and its sensibility and specificity were 90.29% and 90.90%, respectively [15]. Hailu compared the prediction of the different data mining technologies, which were used to develop the HIV testing prediction model. Four popular data mining algorithms (decision tree, naive Bayes, neural network, and logistic regression) were used to build the model that predicted whether an individual was being tested for HIV. The final experimentation results indicated that the decision tree (random tree algorithm) performed the best with an accuracy of 96% [16].

However, in previous studies, few researches considered the use of data mining methods to construct predictive mathematical models of AIDS high-risk group based on several potential risk factors for surveillance report data. This paper aims at using data mining technology to identify the main factors influencing on the status of AIDS high-risk group infection (including injecting drug user (IDU), female sex worker (FSW), and men who have sex with men (MSM)) on surveillance report data in Urumqi and compare the prediction power of the different forecast models based on data mining technology. In order to accomplish this objective, several data mining classification models were considered, namely, random forests (RF), support vector machine (SVM), k-nearest neighbors (KNN), and decision tree (DT), using a 10-fold cross-validation technique. The classification performance was evaluated in terms of a confusion matrix, accuracy, sensitivity, specificity, precision, recall, and AUC values of the receiver operating characteristic (ROC) curves.

2. Materials and Methods

2.1. Study Population. The target populations that met the inclusion criteria in this paper were selected from the data between 2009 and 2015 that the sentinel surveillance of CDC at all levels in Urumqi was reported to China CDC Information System. There are three populations at higher risk of HIV exposure that were considered, including FSW which was defined as women who engaged in commercial sex trade during the investigation; IDU was defined as who takes oral, inhaling, or injecting heroin, cocaine, opium, morphine, marijuana, k-powder, methamphetamine, ecstasy, leprosy, etc.; and MSM was defined as people who have had intercourse or oral sex in the past years.

2.2. Data Source. The data applied in this paper consisted of three datasets from the higher risk of HIV/AIDS exposure populations collected between 2009 and 2015 by the Urumqi CDC. The three datasets are FSW dataset that included 9090 FSWs and 53 attributes, MSM dataset that included 5304 MSM and 57 attributes, and IDU dataset that included 7337 IDUs and 56 attributes. The collected data had three core survey questionnaires: FSW questionnaire, MSM questionnaire, and IDU questionnaire. The survey items included demographic characteristics (age-at-birth, gender,

marital status, nation, place of household registration and educational level, etc.), serological detection results (antibody detection of HIV, syphilis, and HCV), high-risk behaviors factors (drug abuse behavior and sexual behavior), and AIDS prevention strategies and measures (the awareness of AIDS/HIV prevention knowledge, the conditions of prevention, and intervention service and situation of test-accepting).

2.3. Data Preprocessing. Data preprocessing plays an important role in the data mining tasks. Data preprocessing contains many kinds of methods for different preprocessing purposes, including data cleaning, data transformation, and data reduction [17]. In this study, we have selected some appropriate methods to optimize the original dataset. First, the attributes unrelated to the data mining goal were removed in advance, such as questionnaire ID, investigation date, and area codes. And the attributes with a large number of missing values were also removed. Second, the data grouping technique was used to simplify the data mining task. In the multiple distinct values of some attributes, such as age, a numerical variable was discretize into different category groups based on WHO standard for age classification. Ethnicity, originally with 56 distinct values, were converted into three distinct categories according to the constituent ratio of different nationalities as Hans, Uygurs, and others. In addition, simple statistical computations were performed with the R language and software environment, version 3.4.3, to analyze the distribution of the attributes. The dependent variable (T03C) was a binary outcome variable of people who has been tested for HIV with two categories: 0 and 1, where 0 means the HIV test results were negative and 1 means the HIV test results were positive. The results of the attributes description are presented in Tables 1, 2, and 3.

Table 1 shows a total of 5304 MSM respondents tested for HIV. Among them, 377 (7.11%) were detected as HIV positive and 4927 (92.9%) were detected as HIV negative. Table 2 shows a total of 9090 FSW respondents who had received a HIV test; 9041 (99.5%) were HIV-positive, while only 49 (0.5%) were HIV negative. Table 3 shows 7337 IDU respondents who had accepted a HIV test; the HIV negative and positive were 6087 (83%) and 1250 (17%), respectively. These results indicate that there is a need of balancing these two classes of the three datasets. In this article, we employed the Synthetic Minority Over-sampling Technique (SMOTE) [18] to dispose unbalanced samples. In SMOTE algorithm, majority class samples use the undersampling method and minority class samples use the oversampling technique. It potentially performs better than simple oversampling and it is widely used [19, 20].

2.4. Attribute Selection. In a data mining task, the selection of the input attributes is usually a highly important step to improve the classification ability of the models, to reduce the classier complexity, to save the computational time, and to simplify the obtained results. Filtering and wrapper are two main different approaches to select a subset of attributes from all of the attributes used in machine learning. Filtering is to make an independent assessment based on the data general characteristics. Wrapper is to select a feature subset using the evaluation function based on a machine learning algorithm [21]. In this paper, the wrapper methods based on random forests (RF) was used to select the attributes as the inputs of the classification model. RF algorithm is an ensemble learning method based on the aggregation of a large number of decision trees and has proved to be very powerful in many different applications [22–24]. A feature selection based on the random forest classifier has been found to provide multivariate feature importance scores, which are relatively easy to obtain and have been successfully applied to high dimensional data [25, 26]. The quantification procedures of the variable importance scores can be described as follows: computing the variable importance score and permuting score, then selecting the features that have more contribution to classification model, and building models through the feature evaluation criteria of random forest algorithm. The Gini importance considers conditional higher-order interactions among the variables and might be a preferable ranking criterion than a univariate measure [27, 28] and is the feature importance evaluation criteria of random forest algorithm which was used in this study.

2.5. Classification Models

2.5.1. Random Forests (RF). The first algorithm for random decision forests was created by Ho (1995) [29], and its extension version was developed by Breiman [30]. The RF is an ensemble learning method based on decision tree and has been successfully used in several types of classification and regression, especially for accurate identification of disease diagnosis problems [31–33]. RF builds a large number of decision trees using a bootstrap sample with replacement from the training set and predicts the class of each tree according to the test set, and the final RF prediction class is presented based on the majority of the votes [34]. It has been shown to give excellent performance on numerical and categorical data.

2.5.2. Support Vector Machine (SVM). Support vector machine, a novel type of learning machine derived from statistical learning theory, constructs a hyperplane or set of hyperplanes in high- or infinite-dimensional space, which can be used for classification, regression, or other tasks like outliers detection, function estimation, and high-dimensional pattern recognition problems [35–38]. The SVM mainly deals with the problems of binary classification. In addition to performing linear classification, SVM can efficiently perform nonlinear classification through kernel techniques [39] implicitly mapping their inputs into high-dimensional feature spaces. SVM categorization model can be constructed in two ways, as follows: (1) converting the input space into higher dimensional feature space by a nonlinear mapping function. (2) Building the separating hyperplane based on maximum distance from the closest points of the training set [40].

Table 1: Details of the attributes of the MSM dataset.

Variables	Description	Category	Total number	Percentage (%)
A01B	Monitoring sites	Sayibak District	2981	56.2
		Xinshi District	624	11.8
		Shuimogou District	361	6.8
		Tianshan District	1338	25.2
A06	Sample source	Bar/dancehall/ tearooms/club	783	14.8
		Bath/sauna/ pedicure/massage	669	12.6
		Park/public toilet/ grassland	188	3.5
		Network recruiting	3605	68.0
		Others	59	1.1
B01	Age	1(15–17)	27	0.5
		2(18–28)	2704	51.0
		3(29–40)	2140	40.3
		4(41–48)	362	6.8
		5(49–55)	59	1.1
		6(56–65)	8	0.2
		7(>66)	4	0.1
B02	Marital status	Unmarried	4377	82.5
		Married	602	11.3
		Cohabitation	76	1.4
		Divorced or widowed	249	4.7
B03	The location of household register	Xinjiang Uygur Autonomous Region	4620	87.1
		Others	684	12.9
B04	Nation	Hans	4546	85.7
		Uygurs	333	6.3
		Others	425	8.0
B05	Inhabit time	<3 months	123	2.3
		3–6 months	59	1.1
		7–12 months	96	1.8
		1–2 years	332	6.3
		>2 years	4694	88.5
B06	Educational level	Illiteracy	8	0.2
		Primary school	43	0.8
		Junior middle school	346	6.5
		High school or technical school	1081	20.4
		College or above	3826	72.1
C08	Knowledge and awareness of HIV	No	186	3.5
		Yes	5118	96.5
D01	Have you ever had anal sex with a person of the same sex in the last six months	No	356	6.7
		Yes	4948	93.3
D03	Did you use a condom for sex with the same sex last time	No	1169	22.0
		Yes	4135	78.0

Table 1: Continued.

Variables	Description	Category	Total number	Percentage (%)
E01	Have you had any commercial sex with people of the same sex last 6 months	No	5024	94.7
		Yes	280	5.3
F01	Did you have sex with the opposite sex last 6 months	No	4801	90.5
		Yes	503	9.5
G01	Did you take drugs	No	5270	99.4
		Yes	34	0.6
H01	Have you ever been diagnosed with an STD in the last year	No	5168	97.4
		Yes	136	2.6
I01	Have you ever received a condom promotion or HIV/AIDS counselling and testing to prevent HIV/AIDS	No	633	11.9
		Yes	4671	88.1
I02	Have you ever received a community medication to maintain or providing or exchanging cleaning needles to prevent HIV/AIDS	No	5268	99.3
		Yes	36	0.7
I03	Have you ever received a companion education to prevent HIV/AIDS	No	1709	32.2
		Yes	3595	67.8
J01	Has HIV been tested in the last year	No	1726	32.5
		Yes	3578	67.5
T04C	Syphilis test results	No	4979	93.9
		Yes	325	6.1
T05C	Hepatitis test results	Yes	39	0.7
		No	5265	99.3
T03C	HIV test results	No	4927	92.9
		Yes	377	7.1

2.5.3. K-Nearest Neighbors (KNN). The k-nearest neighbors algorithm (KNN) is the simplest but more powerful nonparametric classification method of all data mining methods, since it is a type of instance-based or lazy learning algorithm [41]. KNN classifier has been widely used in many fields, such as text classification, pattern recognition, and disease detection and diagnosis, based on the advantages such as simplicity, high efficiency, and easy to implement [42, 43]. KNN arithmetic idea mainly considers three points: the value of k, distance measurement, and decision rules of classification. The k, as a user-defined constant, will directly affect the KNN classification performance. And the distance metric measures commonly use Euclidean distance, Manhattan distance, and Minkowski distance. The decision rules of classification depend on the majority voting.

2.5.4. Decision Trees (DT). A decision tree is a kind of commonly used data mining method with many advantages such as easy to understand, readable, and quick classification [44]. A decision tree is the organization of the nodes that make decisions like a tree, which consists of decision nodes, branches, and leaf nodes. Each decision node represents a data category or attributes to be classified, and each leaf node represents a result [45]. The whole decision-making process starts from the root decision node, and from top to bottom, it is determined until the classification results are given. There are three commonly used typical decision tree algorithms in data mining at present, such as ID3 algorithm, C4.5 algorithm, and CART algorithm [46].

2.6. Performance Evaluation. In this paper, a confusion matrix and some indicators including accuracy, sensitivity, specificity, precision, recall, and the receiver operating characteristic (ROC) curve were used to appraise the performance of the four classification models. A 10-fold cross-validation was applied to RF, SVM, KNN, and DT validation. A confusion matrix consists of the parts shown in Table 4. In Table 4, TP (true positive) is the positive records of the correct classification, TN (true negative) is the negative records of the correct classification, FP (false positive) is the positive records of the incorrect classification, and FN (false negative) is the negative records of the incorrect classification.

Several important measures, such as accuracy, sensitivity, specificity, precision, and recall, can be calculated by using the confusion matrix. The accuracy is the number of samples correctly classified. The sensitivity is a description of measuring the proportion of correctly classified positive samples. The specificity is a description of measuring the proportion of correctly classified negative samples. The precision is a description of the number of positive samples to the proportion of all predicted positive samples. The recall is a description of the ratio of positive samples

Table 2: Details of the attributes of the FSW dataset.

Variables	Description	Category	Total number	Percentage (%)
A01B	Monitoring sites	Sayibak District	2557	28.1
		Xinshi District	1099	12.1
		Economic Development District	522	5.7
		Shuimogou District	2653	29.2
		Tianshan District	2259	24.9
A06	Sample source	Sauna/bath center	778	8.6
		Nightclub	3157	34.7
		Karaoke hall/ ballroom/bar	2388	26.3
		Guesthouse/hotel	551	6.1
		Foot washing room/ hair salon	1551	17.1
		Roadside shop/little dine	656	7.2
		Street	9	0.1
B01B	Age	1(15–17)	109	1.2
		2(18–28)	6479	71.3
		3(29–40)	2028	22.3
		4(41–48)	394	4.3
		5(49–55)	66	0.7
		6(56–65)	7	0.1
B02	Marital status	Unmarried	5288	58.2
		Married	2359	26.0
		Cohabitation	1051	11.6
		Divorced or widowed	392	4.3
B03	The location of household register	Xinjiang Uygur Autonomous Region	4947	54.4
		Others	4143	45.6
B04	Nation	Hans	7405	81.5
		Uygurs	785	8.6
		Others	900	9.9
B05	Educational level	Illiteracy	118	1.3
		Primary school	949	10.4
		Junior middle school	3738	41.1
		High school or technical school	3383	37.2
		College or above	902	9.9
B06	How long were you working here this time	>=1 year	3180	35.0
		6–12 months	1930	21.1
		1–6 months	2773	30.5
		<1 months	1207	13.3
C08	Knowledge and awareness of HIV	No	401	4.4
		Yes	8689	95.6
D01	Did you use condoms with your guests the last time	No	932	10.3
		Yes	8158	89.7
D02	How often did you use condoms when you have sex with a guest last month	Never used	190	2.1

Table 2: Continued.

Variables	Description	Category	Total number	Percentage (%)
		Sometimes used	2160	23.8
		Every time used	6740	74.1
E01	Did you take drugs	No	9029	99.3
		Yes	61	0.7
F01	Have you ever been diagnosed with an STD in the last year	No	9060	99.7
		Yes	30	0.3
G01	Have you ever received a condom promotion or HIV/AIDS counselling and testing to prevent HIV/AIDS	No	504	5.5
		Yes	8586	94.5
G02	Have you ever received a community medication to maintain or providing or exchanging cleaning needles to prevent HIV/AIDS	No	8950	98.5
		Yes	140	1.5
G03	Have you ever received a companion education to prevent HIV/AIDS	No	2485	27.0
		Yes	6632	73.0
H01	Has HIV been tested in the last year	No	4429	48.7
		Yes	4661	51.3
T04C	Syphilis test results	No	8904	98.0
		Yes	186	2.0
T05C	Hepatitis test results	No	8986	98.9
		Yes	104	1.1
T03C	HIV test results	No	9041	99.5
		Yes	49	0.5

to the total number of positive samples. The accuracy, sensitivity, specificity, precision, and recall are defined as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \times 100\%, \quad (1)$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100\%, \quad (2)$$

$$\text{Specificity} = \frac{\text{TP}}{\text{TN} + \text{FP}} \times 100\%, \quad (3)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100\%, \quad (4)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\%. \quad (5)$$

The ROC curve is originally derived from statistical decision theory, which can comprehensively describe the classification performance of the classifiers with different discriminant thresholds [47]. The vertical axis of the ROC curve is TP rate, and the horizontal axis is FP rate. However, in a practical application, the AUC (the area under the ROC curve) is often used to evaluate the performance of the classifier.

3. Experimental Results

R is an open source programming language and software environment for statistical computing and graphics. Based on the R language environment, the implementation of each algorithm in this experiment is carried out. Here, we used SMOTE (DMwR), randomForest (randomForest), ksvm (kernlab), kknn (kknn), and rpart (rpart) packages. All experiments were validated with a 10-fold cross-validation technique in order to present a more stable accuracy rate after applying the four classification models. Some evaluation indexes were used to compare the classification performance of four data mining algorithms.

Table 5 shows the three original datasets and the three artificial datasets obtained using SMOTE algorithm. It is evident that the original datasets are biased; the imbalance rate of each original datasets is 13.0689, 184.5102, and 4.8696, respectively. In order to achieve the data balance to avoid the result bias, we used SMOTE algorithm combining the oversampling the minority class and undersampling the majority class techniques. We apply the function SMOTE in the DMwR package in R software. The three main parameters of function SMOTE are perc.over, perc.under, and k. The parameter perc.over and perc.under control the amount of oversampling of the minority classes and undersampling of the majority classes, respectively. The parameter k controls the way of the new examples created. For the parameters in the SMOTE algorithm, the value of k was set to 5. For the initial dataset of MSM with 377 minority samples and 4927 majority samples, we set the parameters perc.over = 1200 and perc.under = 110, respectively. Firstly, the number of minority samples was increased; a total of $1200 \times 377/100$ new minority samples were generated. The original minority samples and the new minority samples consisted of the new dataset. Secondly, sampling the majority sample, we obtain a new sample of the majority, which is $(110/100) \times 1200 \times 377/100$. We put the new sample of the majority into the new dataset which was created

Table 3: Details of the attributes of the IDU dataset.

Variables	Description	Category	Total number	Percentage (%)
A01B	Monitoring sites	Sayibak District	2147	32.9
		Xinshi District	892	12.2
		Shuimogou District	1802	24.6
		Tianshan District	1922	26.2
		Toutun River District	56	0.8
		Urumqi County	248	3.4
A06	Sample source	Compulsory detoxification setting	1617	22.0
		Community	5063	69.0
		Methadone clinic (urine test positive)	657	9.0
B02	Age	1(15–17)	49	0.7
		2(18–28)	1719	23.4
		3(29–40)	3493	47.6
		4(41–48)	1721	23.5
		5(49–55)	305	4.2
		6(56–65)	43	0.6
		7(>66)	5	0.1
B01	Gender	Male	6549	89.3
		Female	788	10.7
B03	Marital status	Unmarried	2586	35.2
		Married	3241	44.2
		Cohabitation	225	3.1
		Divorced or widowed	1285	17.5
B04	The location of household register	Xinjiang Uygur Autonomous Region	6762	92.2
		Others	575	7.8
B05	Nation	Hans	2452	33.4
		Uygurs	3880	52.9
		Others	1005	13.7
B06	Educational level	Illiteracy	377	5.1
		Primary school	1561	21.3
		Junior middle school	3231	44.0
		High school or technical school	1673	22.8
		College or above	495	6.7
C08	Knowledge and awareness of HIV	No	139	1.9
		Yes	7198	98.1
D01	How many drugs did you use at present	1 kind	6618	90.2
		2 kinds	646	8.8
		3 kinds	57	0.8
		4 kinds	13	0.2
		5 kinds	2	0.0
		6 kinds	1	0.0
D02	Did you take drugs	Sometimes used	2160	23.8
		Every time used	6740	74.1
		No	1812	24.7

Table 3: Continued.

Variables	Description	Category	Total number	Percentage (%)
E01	Have you ever had sex last month	Yes	5525	75.3
		No	4562	62.2
F01	Have you ever had sex with a commercial partner in the last year	Yes	2775	37.8
		No	6516	88.8
G01	Have you ever received a condom promotion or HIV/AIDS counselling and testing to prevent HIV/AIDS	Yes	821	11.2
		No	1453	19.8
G02	Have you ever received a community medication to maintain or providing or exchanging cleaning needles to prevent HIV/AIDS	Yes	5884	80.2
		No	3056	41.7
G03	Have you ever received a companion education to prevent HIV/AIDS	Yes	4281	58.3
		No	3353	45.7
H01	Has HIV been tested in the last year	Yes	3984	54.3
		No	3080	42.0
T04C	Syphilis test results	Yes	4257	58.0
		No	7093	96.7
T05C	Hepatitis test results	Yes	244	3.3
		No	3389	46.2
T03C	HIV test results	Yes	3948	53.8
		No	6087	83.0
		Yes	1250	17.0

Table 4: Confusion matrix for the two-class problem.

	Predicted negative	Predicted positive
Actual negative	TN	FP
Actual positive	FN	TP

Table 5: Description of original data and balanced data.

Dataset	Minority class	Majority class	Samples in total	Imbalance rate
MSM (original)	377	4927	5304	13.0689
MSM (SMOTE)	4901	4976	9877	1.0153
FSW (original)	49	9041	9090	184.5102
FSW (SMOTE)	9849	9898	19,747	1.0049
IDU (original)	1250	6087	7337	4.8696
IDU (SMOTE)	6250	6300	12,550	1.008

above. Eventually, in this new dataset, both the minority sample and the majority sample were $(1 + 1200/100) \times 377$ and $(110/100) \times 1200 \times 377/100$, respectively. For the initial dataset of FSW with 49 minority samples and 9041 majority samples, we set the parameters perc.over = 20,000 and perc.under = 101. The oversampling and undersampling algorithms also were utilized in the MSM dataset. The result demonstrated the new dataset with minority samples $(1 + 20{,}000/100) \times 49$ and majority samples $(101/100) \times 20{,}000/49/100$. For the initial dataset of IDU with 1250 minority samples and 6087 majority samples, setting the parameters perc.over = 400 and perc.under = 216, the minority sample and the majority sample were $1 + 400/100 \times 1250$ and $216/100 \times 400 \times 1250/100$, respectively.

Figures 1, 2, and 3 describe the importance of the sorted variables of the three datasets (MSM dataset, FSW dataset, and IDU dataset) according to the Gini index criterion from RF. From Figure 1, for the MSM dataset, the most important variables are B01, B06, A01B, A06, and B05. The least important variables are I02, G01, H01, and D01. From Figure 2, for the FSW dataset, the most important variables are B01B, T05C, A06, B05, and B06. The least important variables are F01, G02, C08, E01, and D01. From Figure 3, for the IDU dataset, the most important variables are B02, A01, T05C, B06, and B05. The least important variables are C08, B04, T04C, F01, and D01. Finally, applying the rank + MeanDecreaseGini method of attribute selection method, variables were ranked based on their importance in classifying the HIV patients. We also asked the CDC doctors about the importance of lower-ranking attributes, combining the two methods agree that B01, B06, A01B, A06, B05, B04, B02, D03, I03, J01, I01, B03, F01, T04C, and E01 as the main subset of attributes important in predicting the HIV patients from MSM population, B01B, T05C, A06, B05, B06, B04, B02, A01B, D02, H01, T04C, G03, B03, and G01 as the main subset of attributes important in predicting the HIV patients from female sex workers population, and B02, A01, T05C, B06, B05, B03, A06, D02, H01, G02, G03, E01, G01, and B01 as the main subset of attributes important in predicting the HIV patients from drug users population. The detailed descriptions of the selected attributes were shown in Tables 6, 7, and 8.

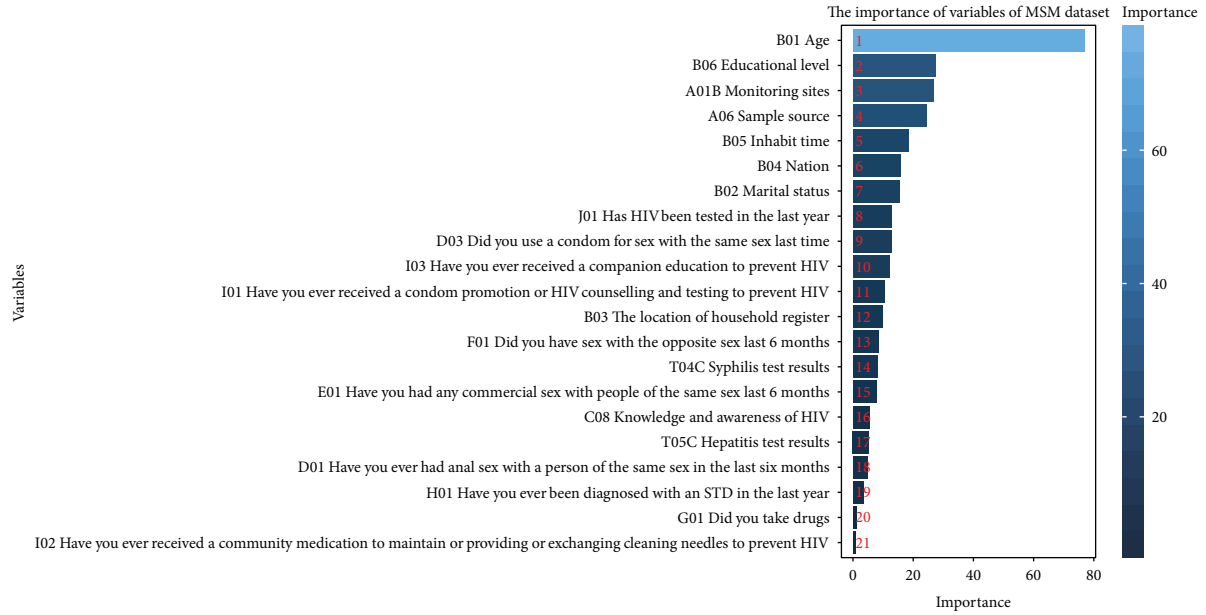

Figure 1: The importance of variables of MSM dataset.

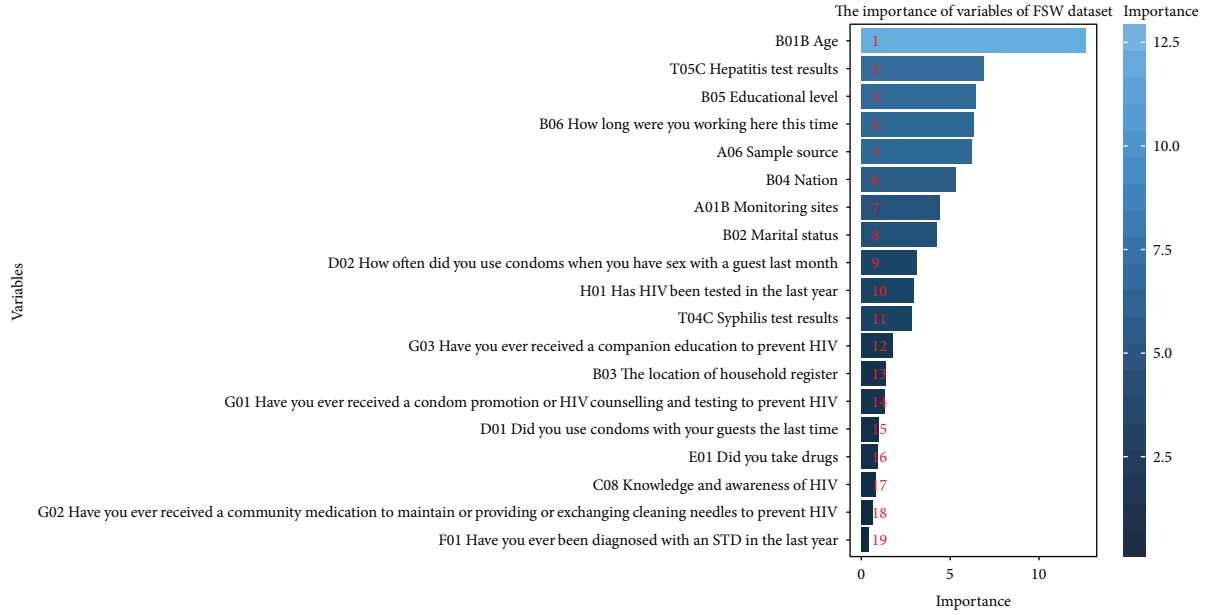

Figure 2: The importance of variables of FSW dataset.

Figures 4, 5, and 6 show the ROC curve obtained for the three datasets with the four classifiers. The AUC scores for RF, SVM, KNN, and DT on MSM dataset are 0.9802, 0.9401, 0.9747, and 0.7917; 0.9981, 0.9803, 0.9967, and 0.8702 on FSW dataset; and 0.9874, 0.9135, 0.9802, and 0.7438 on IDU dataset. It is obvious that RF performed significantly better than the other three classifiers. The AUC scores achieved for MSM dataset, FSW dataset, and IDU datasets are 0.9802, 0.9981, and 0.9874, respectively. The maximum value of the AUC (0.9981) was obtained for the FSW dataset with RF algorithm. Moreover, the value of AUC of DT algorithm with IDU dataset is 0.7438 which is the minimum of all AUC scores.

Figures 7, 8, and 9 depict the classification performance when the four classifiers are applied on MSM dataset, FSW dataset, and IDU dataset, respectively. The accuracy, precision, and recall for RF, SVM, KNN, and DT on the three datasets were compared. For the MSM dataset (Figure 7), the SVM model achieved a classification accuracy of 87.8404%, with a precision of 89.5130% and a recall of 85.5132%. The KNN model had a classification accuracy of 91.5258%, with a precision of 89.5130% and a recall of 85.5132%. For the

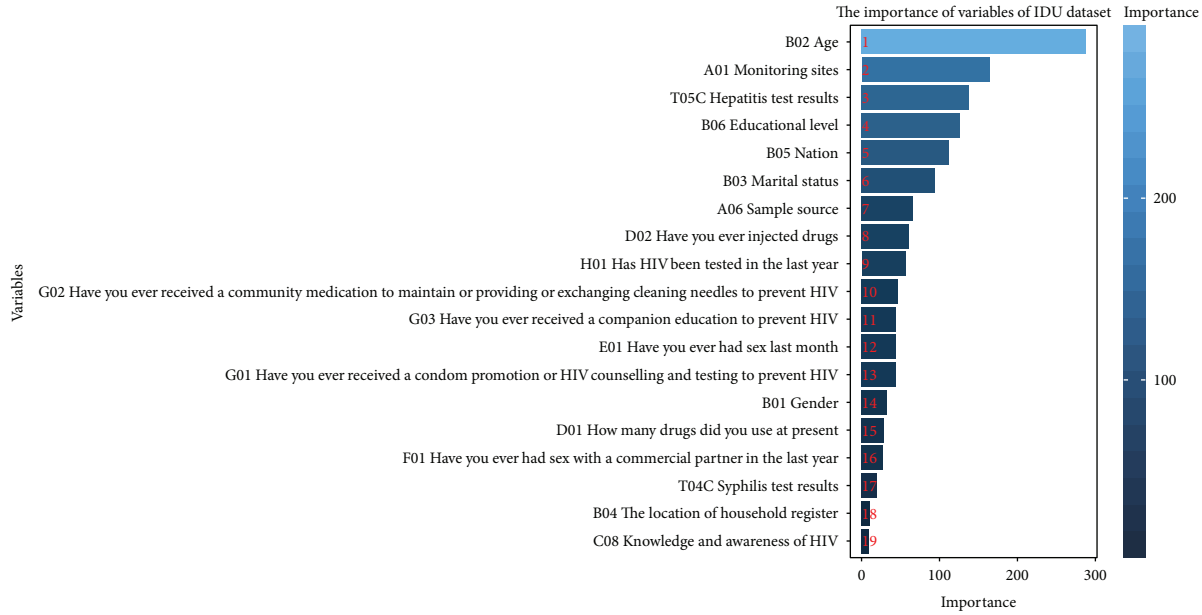

Figure 3: The importance of variables of IDU dataset.

Table 6: Selection attributes used in models of MSM dataset.

Rank	Attribute	MeanDecreaseGini
1	B01: age	76.8033
2	B06: educational level	27.1032
3	A01B: monitoring sites	26.0119
4	A06: sample source	23.9942
5	B05: inhabit time	18.3735
6	B04: nation	16.2218
7	B02: marital status	14.9883
8	D03: did you use a condom for sex with the same sex last time	12.7123
9	I03: have you ever received a companion education to prevent HIV/AIDS	12.2440
10	J01: has HIV been tested in the last year	12.1464
11	I01: have you ever received a condom promotion or HIV/AIDS counselling and testing to prevent HIV/AIDS	10.1819
12	B03: the location of household register	9.7513
13	F01: did you have sex with the opposite sex last 6 months	8.5185
14	T04C: syphilis test results	8.2889
15	E01: have you had any commercial sex with people of the same sex last 6 months	7.6851

decision tree, the accuracy, precision, and recall were 76.7440%, 77.6199%, and 74.6582%, respectively. The random forest algorithm performed best among the four evaluated models with an accuracy of 94.4821%, a precision of 98.5511%, and a recall of 90.2061%.

For the FSW dataset (Figure 8), the final experimental results demonstrated that the random forest algorithm showed the best with an accuracy of 97.5136%, and the precision and recall were 97.4638% and 91.6160%, respectively. The KNN model came out to be the second with a classification accuracy of 96.3083%, and the precision and recall were 97.4210% and 95.1163%, respectively, followed by SVM model with a classification accuracy of 93.3560%, the precision and recall equal to 94.1554% and 92.4155%, respectively. The decision tree has also performed the least classification accuracy of 85.0408%, and the precision and recall were 86.9467% and 82.3739%, respectively.

For the IDU dataset (Figure 9), the RF classifier showed the best predictive performances; the accuracy, precision, and recall gave 94.6375%, 97.4638%, and 91.6160%, respectively. In the SVM model, they were 83.4821%, 84.8141%, and 81.4080%, respectively. As shown in the confusion matrix in Table 10, the KNN learning algorithm scored an accuracy of 90.8287%; the precision and recall were 94.7831%, 86.3360%, respectively. Using the decision tree had a lower overall performance, with an accuracy of

Table 7: Selection attributes used in models of FSW dataset.

Rank	Variables	MeanDecreaseGini
1	B01B: age	12.6253
2	T05C: hepatitis test results	6.7033
3	A06: sample source	6.6001
4	B05: educational level	6.3421
5	B06: how long were you working here this time	6.1513
6	B04: nation	5.2128
7	B02: marital status	4.6192
8	A01B: monitoring sites	4.4660
9	D02: how often did you use condoms when you have sex with a guest last month	2.9029
10	H01: has HIV been tested in the last year	2.8776
11	T04C: syphilis test results	2.8470
12	G03: have you ever received a companion education to prevent HIV/AIDS?	1.6805
13	B03: the location of household register	1.4158
14	G01: have you ever received a condom promotion or HIV/AIDS counselling and testing to prevent HIV/AIDS	1.2143

Table 8: Selection attributes used in models of IDU dataset.

Rank	Variables	MeanDecreaseGini
1	B02: age	292.3608
2	A01: monitoring sites	166.8695
3	T05C: hepatitis test results	142.0867
4	B06: educational level	125.3663
5	B05: nation	112.2430
6	B03: marital status	92.2254
7	A06: sample source	63.6016
8	D02: have you ever injected drugs	58.3517
9	H01: has HIV been tested in the last year	55.1894
10	G02: have you ever received a community medication to maintain or providing or exchanging cleaning needles to prevent HIV/AIDS	45.0500
11	G03: have you ever received a companion education to prevent HIV/AIDS	44.9624
12	E01: have you ever had sex last month	43.5729
13	G01: have you ever received a condom promotion or HIV/AIDS counselling and testing to prevent HIV/AIDS	42.9014
14	B01: gender	33.0323

71.2271%, precision and recall were 69.8690% and 74.2400%, respectively.

The other performance metrics confusion matrixes, such as sensitivity and specificity, were also employed to measure the performance of different classifiers for the three datasets. As a whole, the RF classifier has the best performance as compared to the other three methods and has obtained higher accuracies 94.4821%, 97.5136%, and 94.6375% on MSM dataset, FSW dataset, and IDU dataset, respectively. The decision tree has also achieved the least classification accuracy 76.7440%, 85.0408%, and 71.2271% on MSM dataset, FSW dataset, and IDU dataset, respectively. The detailed classification outcomes of each model for the three datasets are shown in Tables 9, 10, and 11.

4. Discussion

The AIDS epidemic in Urumqi is still very serious. The increasing number of high-risk groups, such as prostitutes, male sex workers, and floating population, has exacerbated the difficulty of AIDS prevention and treatment. Data mining has been widely used in the field of diagnosis, evaluation, and other medical fields [48]. This study aimed at using four mature data mining algorithms (random forests, support vector machine, k-nearest neighbors, and decision tree) to build identification models for AIDS patients based on the sentinel monitoring data of HIV high-risk populations (MSM, FSWs, and IDUs) in Urumqi and compared the prediction power of the different models. However, considering

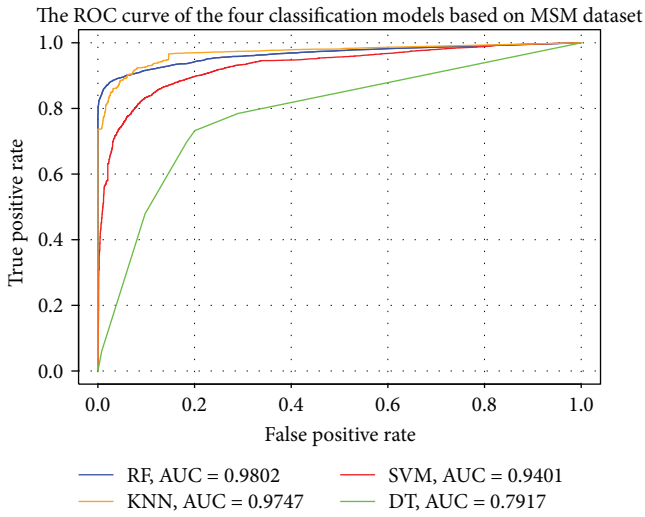

Figure 4: ROC curve of different classifiers for MSM dataset.

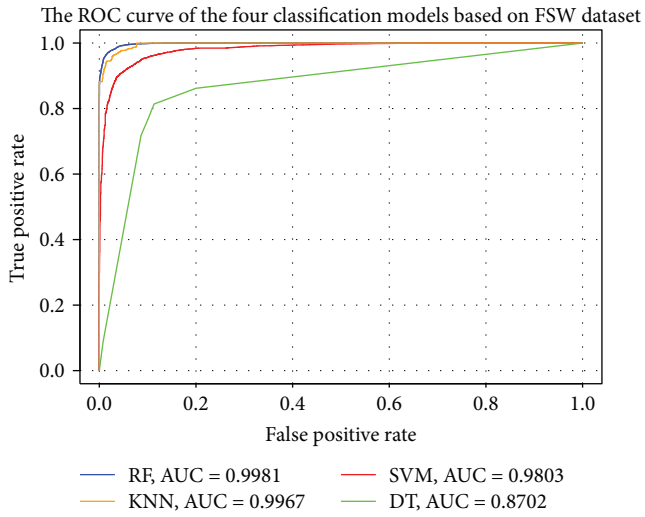

Figure 5: ROC curve of different classifiers for FSW dataset.

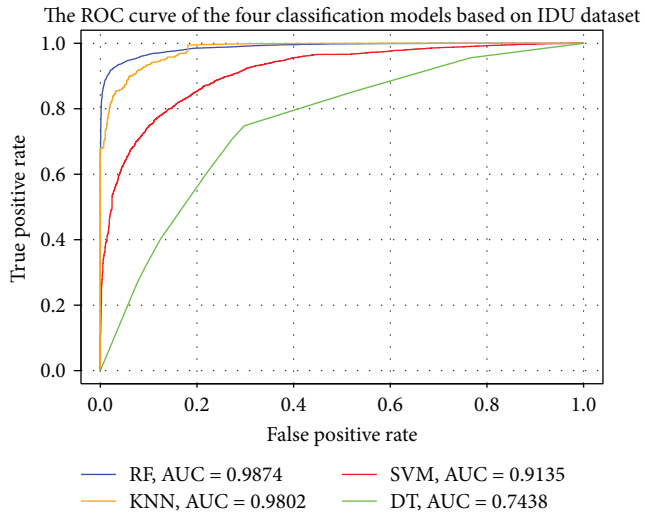

Figure 6: ROC curve of different classifiers for IDU dataset.

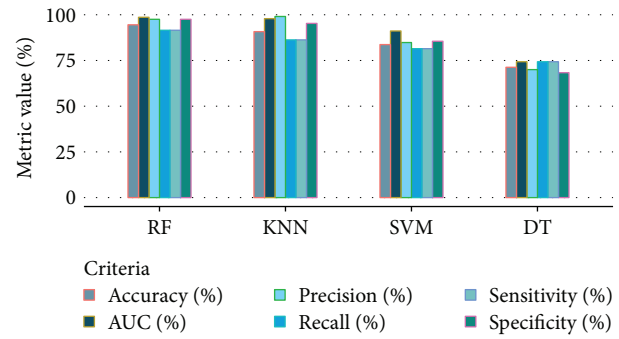

Figure 7: Performance of different classification models for MSM dataset.

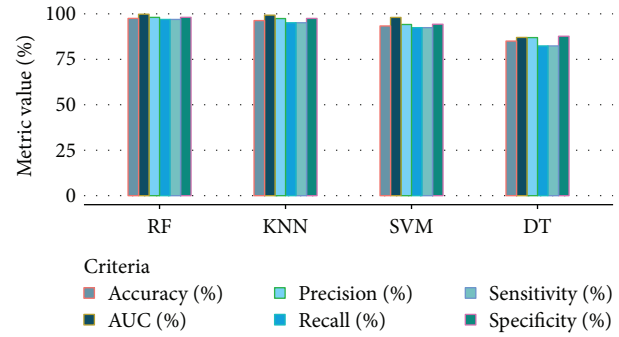

Figure 8: Performance of different classification models for FSW dataset.

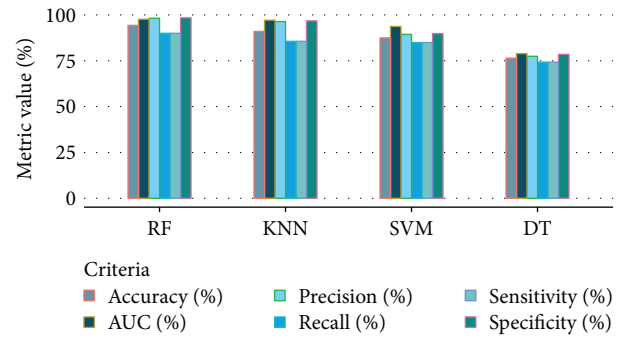

Figure 9: Performance of different classification models for IDU dataset.

that the major defect in the model build process is class imbalances, the SMOTE method has been used to simulate the data balance and overcome the problem of overfitting according to the previous research [49].

For all datasets, the final experimental results showed that RF algorithm obtains the best results; the diagnostic accuracy for RF on MSM dataset are 94.4821%, 97.5136% on FSW dataset, and 94.6375% on IDU dataset. The KNN algorithm came out second, with 91.5258% diagnostic accuracy on MSM dataset, 96.3083% diagnostic accuracy on FSW dataset, and 90.8287% diagnostic accuracy on IDU dataset, followed by SVM (94.0182%, 98.0369%, and

Table 9: Performance measures of the classifiers for MSM dataset.

Testing criteria	RF		SVM		KNN		DT	
Confusion matrix	4911	65	4485	491	4828	148	3921	1055
	480	4421	710	4191	689	4212	1242	3659
Accuracy (%)	94.4821		87.8404		91.5258		76.7440	
Sensitivity (%)	90.2061		85.5132		85.9416		74.6582	
Specificity (%)	98.6937		90.1326		97.0257		78.7982	
Precision (%)	98.5511		89.5130		96.6055		77.6199	
Recall (%)	90.2061		85.5132		85.9416		74.6582	
AUC (%)	98.0217		94.0182		97.4709		79.1761	

Table 10: Performance measures of the classifiers for IDU dataset.

Testing criteria	RF		SVM		KNN		DT	
Confusion matrix	6151	149	5389	911	6003	297	4299	2001
	524	5726	1162	5088	854	5396	1610	4640
Accuracy (%)	94.6375		83.4821		90.8287		71.2271	
Sensitivity (%)	91.6160		81.4080		86.3360		74.2400	
Specificity (%)	97.6349		85.5397		95.2857		68.2381	
Precision (%)	97.4638		84.8141		94.7831		69.8690	
Recall (%)	91.6160		81.4080		86.3360		74.2400	
AUC (%)	98.7495		91.3571		98.0208		74.3879	

Table 11: Performance measures of the classifiers for FSW dataset.

Testing criteria	RF		SVM		KNN		DT	
Confusion matrix	9709	189	9333	565	9650	248	8680	1218
	302	9547	747	9102	481	9368	1736	8113
Accuracy (%)	97.5136		93.3560		96.3083		85.0408	
Sensitivity (%)	96.9337		92.4155		95.1163		82.3739	
Specificity (%)	98.0905		94.2918		97.4944		87.6945	
Precision (%)	98.0588		94.1554		97.4210		86.9467	
Recall (%)	96.9337		92.4155		95.1163		82.3739	
AUC (%)	99.8114		98.0369		99.6712		87.0283	

91.3571%). The DT algorithm was the poorest of the four algorithms, with 79.1761% diagnostic accuracy on MSM dataset, 87.0283% diagnostic accuracy on FSW dataset, and 74.3879% accuracy on IDU. These results suggested that the four established data mining models can predict whether a person is infected with HIV. But compared with SVM, decision tree, and KNN, random forest model through a large number of random sample method balance the sampling error; the effect of classifying the results produces a large number of different test data. A comprehensive assessment is just a single test sample for fitting the results of the other three models more reliably [50].

This study based on the importance score of independent variables for random forest model identified the most important influencing factor for the HIV infection in the three high dangerous populations in Urumqi. For the MSM dataset, these variables are age, educational level, monitoring sites, sample source, inhabit time, nation, marital status, etc. Variables such as age show that the MSM population in Urumqi is mainly the young and middle-aged active population aged from 18 to 40 years old, accounting for 91.3%, which is similar to the monitoring results in Chengdu [51] and show that sexually active people are still the focus of AIDS prevention and treatment. The majority (82.5%) of the participants had never been married. More than half (56.2%) came from the Sayibak District, 68% of the participants were recruited through the network, and 72.1% had some college or higher education. Therefore, based on the epidemic characteristics of MSM population in Urumqi, personal characteristics and social factors should be taken into account comprehensively when education intervention measures are carried out for this population. For the FSW dataset, the results showed that

most of the female sex workers (FSWs) in Urumqi were young women under 30 years old, 58.2% were unmarried, 65% of female sex workers (FSWs) worked in a local workspace for less than a year, and more than half were primary school and junior middle school and had come mainly from nightclub, karaoke, ballroom, and bar. Therefore, we should focus on the actual epidemic characteristics of FSWs to take corresponding measures to publicize education and intervene. For the IDU dataset, the age of the 7337 participants ranged from l1 to 71 years, with more than half (94.5%) of them aged 18–48 years. Among them, 2586 (35.2%) were single, with 2147 (32.9%) participants coming from Sayibak District, and 5169(66.4%) participants were junior high school and below. Among the participants, 89.3% were male and 69% were from the community. These results can provide evidence for the prevention of HIV infection among drug users through the promotion of education, especially for adolescents, low cultural level population, floating population, drug abuse, sexual disorder, etc.

As we have shown above, data mining models can accurately identify diseases based on certain important attributes. These predictive models are valuable tools in the medical field. However, there are areas of concern in the development of predictive models: (1) the model should include all clinically relevant data, (2) the model should be tested on an independent sample, and (3) the model must make sense to the medical personnel who are supposed to make use of it. It has been shown that not all predictive models constructed using data mining techniques satisfy all of these requirements [52].

There are some limitations to this article. First, all individuals are recruited in Urumqi, which was limited by geographical and population characteristics. Therefore, the information bias may exist during the experiment process. If the study population could be expanded to more than one province or to the whole country, the model recognition effect would be better. Second, in the epidemiological investigation of HIV-infected persons, due to subjective, objective, and other reasons, respondents may provide unreal information, which leads to a certain influence on the analysis results. In the future, more feature selection methods, class imbalance processing methods, and data mining algorithms are expected to be tested.

5. Conclusion

In general, four prediction models were established and compared for predicting whether a person is infected with HIV. The results showed that the random forest model performed the best in classification accuracy. This study can provide some effective ways for medical staffs to quickly screen and diagnose AIDS from a large amount of information.

Data Availability

The (CSV) data used to support the findings of this study are restricted in order to protect patient privacy. Data are available from Kai Wang (wangkaimath@sina.com) for researchers who meet the criteria for access to confidential data.

Conflicts of Interest

The authors declare no conflict of interest.

Authors' Contributions

Dandan Tang, Kai Wang, and Yujian Zheng designed the project; Man Zhang, Jiabo Xu, and Xueliang Zhang participated in the data collection; Dandan Tang, Li Feng, and Huling Li performed the analysis of the data; Dandan Tang and Fang Yang wrote the manuscript. All authors contributed to the interpretation of the results, revised the manuscript critically, and approved the final version of the manuscript.

Acknowledgments

This project was supported by the National Natural Science Foundation of China (11461073, 11301451).

References

[1] O. Singh and E. C. Y. Su, "Prediction of HIV-1 protease cleavage site using a combination of sequence, structural, and physicochemical features," *BMC Bioinformatics*, vol. 17, Supplement 17, pp. 478–289, 2016.

[2] M. A. Nowak and A. J. Mcmichael, "How HIV defeats the immune system," *Scientific American*, vol. 273, no. 2, pp. 58–65, 1995.

[3] N. I. Ming-Jian, J. Chen, Y. Zhang et al., "Analysis of epidemic status of HIV/AIDS in Xinjiang," *Bulletin of Disease Control and Prevention*, vol. 27, no. 2, pp. 1–3, 2012.

[4] Q. Zheng, J. Wang, Y. Dong et al., "Analysis of monitoring data of AIDS in Xinjiang from 2004 to 2015," *Bulletin of Disease Control & Prevention*, vol. 32, no. 1, pp. 34–48, 2017.

[5] M. A. Ling, "HIV/AIDS epidemic in Urumqi from 1995 to 2011," *Modern Preventive Medicine*, vol. 109, pp. 2727–2729, 2013.

[6] M. A. Ling and Y. X. Wang, "Characteristics of man who have sex with men HIV/AIDS cases reported through internet based direct reporting system in Urumqi," *World Latest Medicine Information*, vol. 16, no. 52, pp. 1-2, 2016.

[7] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, Google Ebook, 2011.

[8] D. A. Adeniyi, Z. Wei, and Y. Yongquan, "Automated web usage data mining and recommendation system using k-nearest neighbor (knn) classification method," *Applied Computing and Informatics*, vol. 12, no. 1, pp. 90–108, 2016.

[9] H. B. Burke, P. H. Goodman, D. B. Rosen et al., "Artificial neural networks improve the accuracy of cancer survival prediction," *Cancer*, vol. 79, pp. 857–862, 1997.

[10] C. D. Chang, C. C. Wang, and B. C. Jiang, "Using data mining techniques for multi-diseases prediction modeling of hypertension and hyperlipidemia by common risk factors," *Expert Systems with Applications*, vol. 38, no. 5, pp. 5507–5513, 2011.

[11] X. H. Meng, Y. X. Huang, D. P. Rao, Q. Zhang, and Q. Liu, "Comparison of three data mining models for predicting diabetes or prediabetes by risk factors," *The Kaohsiung Journal of Medical Sciences*, vol. 29, no. 2, pp. 93–99, 2013.

[12] L. Wang, "Application of data mining technology in diagnosis and treatment of AIDS," *Journal of Mathematical Medicine*, vol. 26, no. 1, pp. 97–99, 2013.

[13] A. Oliveira, B. M. Faria, A. R. Gaio, and L. P. Reis, "Data mining in HIV-AIDS surveillance system," *Journal of Medical Systems*, vol. 41, no. 4, p. 51, 2017.

[14] D. Wang, B. Larder, A. Revell et al., "A comparison of three computational modelling methods for the prediction of virological response to combination HIV therapy," *Artificial Intelligence in Medicine*, vol. 47, no. 1, pp. 63–74, 2009.

[15] W. U. Hai-Lei, J. S. Qian, and C. Zhang, "A HIV carrier forecasting model for quarantine based on support vector machines," *Practical Preventive Medicine*, vol. 17, no. 11, pp. 2152–2155, 2010.

[16] T. G. Hailu, "Comparing data mining techniques in HIV testing prediction," *Intelligent Information Management*, vol. 07, no. 3, pp. 153–180, 2015.

[17] A. Famili, W. Shen, R. Weber, and E. Simoudis, "Data preprocessing and intelligent data analysis," *Intelligent Data Analysis*, vol. 1, no. 1-4, pp. 3–23, 1997.

[18] R. Blagus and L. Lusa, "SMOTE for high-dimensional class-imbalanced data," *BMC Bioinformatics*, vol. 14, no. 1, pp. 106–116, 2013.

[19] L. Zhang, C. Zhang, R. Gao, R. Yang, and Q. Song, "Using the SMOTE technique and hybrid features to predict the types of ion channel-targeted conotoxins," *Journal of Theoretical Biology*, vol. 403, pp. 75–84, 2016.

[20] E. M. Karabulut and T. Ibrikci, "Effective automated prediction of vertebral column pathologies based on logistic model tree with smote preprocessing," *Journal of Medical Systems*, vol. 38, no. 5, p. 50, 2014.

[21] H. Liu, X. Shi, D. Guo, Z. Zhao, and Yimin, "Feature selection combined with neural network structure optimization for HIV-1 protease cleavage site prediction," *BioMed Research International*, vol. 2015, Article ID 263586, 11 pages, 2015.

[22] J. R. Bienkowska, G. S. Dalgin, F. Batliwalla et al., "Convergent random forest predictor: methodology for predicting drug response from genome-scale data applied to anti-TNF response," *Genomics*, vol. 94, no. 6, pp. 423–432, 2009.

[23] X. Chen and H. Ishwaran, "Random forests for genomic data analysis," *Genomics*, vol. 99, no. 6, pp. 323–329, 2012.

[24] M. Kotti, L. D. Duffell, A. A. Faisal, and A. H. Mcgregor, "Detecting knee osteoarthritis and its discriminating parameters using random forests," *Medical Engineering & Physics*, vol. 43, pp. 19–29, 2017.

[25] A. Hapfelmeier and K. Ulm, "A new variable selection approach using random forests," *Computational Statistics & Data Analysis*, vol. 60, pp. 50–69, 2013.

[26] M. Sandri and P. Zuccolotto, "Variable selection using random forests," in *Data Analysis, Classification and the Forward Search*, pp. 263–270, 2006.

[27] B. H. Menze, B. M. Kelm, R. Masuch et al., "A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data," *BMC Bioinformatics*, vol. 10, no. 1, pp. 213–216, 2009.

[28] A.-L. Boulesteix, A. Bender, J. Lorenzo Bermejo, and C. Strobl, "Random forest Gini importance favours SNPs with large minor allele frequency: impact, sources and recommendations," *Briefings in Bioinformatics*, vol. 13, no. 3, pp. 292–304, 2012.

[29] T. K. Ho, "Random decision forests," in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, p. 278, Montreal, Quebec, Canada, August 1995.

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

[31] M. Dauwan, J. J. van der Zande, E. van Dellen et al., "Random forest to differentiate dementia with Lewy bodies from Alzheimer's disease," *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, vol. 4, pp. 99–106, 2016.

[32] A. T. Azar, H. I. Elshazly, A. E. Hassanien, and A. M. Elkorany, "A random forest classifier for lymph diseases," *Computer Methods and Programs in Biomedicine*, vol. 113, no. 2, pp. 465–473, 2014.

[33] T. Shaikhina, D. Lowe, S. Daga, D. Briggs, R. Higgins, and N. Khovanova, "Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation," *Biomedical Signal Processing and Control*, vol. 37, pp. 1025–1042, 2017.

[34] S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C. Westland, "Data mining for credit card fraud: a comparative study," *Decision Support Systems*, vol. 50, no. 3, pp. 602–613, 2011.

[35] Y. Tian, X. Ju, Z. Qi, and Y. Shi, "Efficient sparse least squares support vector machines for pattern classification," *Computers & Mathematics with Applications*, vol. 66, no. 10, pp. 1935–1947, 2013.

[36] C. S. Lo and C. M. Wang, "Support vector machine for breast MR image classification," *Computers & Mathematics with Applications*, vol. 64, no. 5, pp. 1153–1162, 2012.

[37] Y. Guo, L. Yu, Z. Wen, and M. Li, "Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences," *Nucleic Acids Research*, vol. 36, no. 9, pp. 3025–3030, 2008.

[38] H. Yang, L. Chan, and I. King, "Support vector machine regression for volatile stock market prediction," in *Intelligent Data Engineering and Automated Learning — IDEAL 2002*, vol. 2412, pp. 391–396, 2002.

[39] C. K. I. Williams, "Learning with kernels: support vector machines, regularization, optimization, and beyond," *Publications of the American Statistical Association*, vol. 98, pp. 489–489, 2002.

[40] V. P. Gladis Pushpa Rathi, "A novel approach for feature extraction and selection on MRI images for brain tumor classification," *International Conference on Computer Science, Engineering and Applications.*, vol. 10, no. 5, pp. 225–234, 2012.

[41] M. Akhil Jabbar, B. L. Deekshatulu, and P. Chandra, "Classification of heart disease using k-nearest neighbor and genetic algorithm," *Procedia Technology*, vol. 10, pp. 85–94, 2013.

[42] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[43] E. A. Aydın and M. K. Keleş, "Breast cancer detection using k-nearest neighbors data mining method obtained from the bow-tie antenna dataset," *International Journal of RF and Microwave Computer-Aided Engineering*, vol. 27, no. 6, 2017.

[44] F. I. Alam, F. K. Bappee, M. R. Rabbani, and M. M. Islam, "An optimized formulation of decision tree classifier," *Communications in Computer and Information Science*, vol. 361, pp. 105–118, 2013.

[45] J. R. Neto, Z. M. de Souza, S. R. de Medeiros Oliveira et al., "Use of the decision tree technique to estimate sugarcane productivity under edaphoclimatic conditions," *Sugar Tech.*, vol. 19, no. 6, pp. 662–668, 2017.

[46] K. Boonchuay, K. Sinapiromsaran, and C. Lursinsap, "Decision tree induction based on minority entropy for the class imbalance problem," *Pattern Analysis and Applications*, vol. 20, no. 3, pp. 769–782, 2017.

[47] T. Fawcett, "An introduction to roc analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[48] Y. U. Chang-Chun, H. E. Jia, S. C. Fan et al., "Application of data mining in medical field," *Academic Journal of Second Military Medical University*, vol. 24, pp. 1250–1252, 2003.

[49] D. M. Herrera-Ibatá, A. Pazos, R. A. Orbegozo-Medina, F. J. Romero-Durán, and H. González-Díaz, "Mapping chemical structure-activity information of HAART-drug cocktails over complex networks of AIDS epidemiology and socioeconomic data of U.S. counties," *Bio Systems*, vol. 132-133, pp. 20–34, 2015.

[50] T. A. Almeida, R. M. Silva, and A. Yamakami, "Machine learning methods for spamdexing detection," *International Journal of Information Security Science*, vol. 2, pp. 1–22, 2016.

[51] Y. Feng, Z. Wu, R. Detels et al., "HIV/STD prevalence among men who have sex with men in Chengdu, China and associated risk factors for HIV Infection," *Journal of Acquired Immune Deficiency Syndromes*, vol. 53, Supplement 1, pp. S74–S80, 2010.

[52] D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability: a comparison of three data mining methods," *Artificial Intelligence in Medicine*, vol. 34, no. 2, pp. 113–127, 2005.