Egyptian Informatics Journal 24 (2023) 107–115

Contents lists available at ScienceDirect

Egyptian Informatics Journal

journal homepage: www.sciencedirect.com

HIV/AIDS predictive model using random forest based on socio-demographical, biological and behavioral data

Sehar Un Nisa [a], Azhar Mahmood [b,*], Farhan Sabir Ujager [c], Mehwish Malik [d]

[a] Department of Computer Science, Faculty of Computing and Engineering Sciences, Shaheed Zulfikar Ali Bhutto Institute of Science and Technology (SZABIST), Islamabad, Pakistan
[b] Faculty of Computing, Capital University of Science & Technology (CUST), Department of Computer Science, Islamabad, Pakistan
[c] Faculty of Engineering & IT, Northern University, Nowshera, Pakistan
[d] School of Electrical Engineering & Computer Science, National University of Science and Technology, Islamabad, Pakistan

ARTICLE INFO

Article history:
Received 21 June 2021
Revised 30 November 2022
Accepted 8 December 2022
Available online 16 December 2022

Keywords:
HIV/AIDS
Injecting Drugs Demographic
Sexual behavior
Machine learning

ABSTRACT

Since the beginning of the AIDS epidemic, the HIV virus has infected millions of people, and the count is on the rise with every passing day. This epidemic has affected not only the children and adults but also infants born to HIV positive mothers. Unfortunately, there is no cure for HIV/AIDS; however, early and accurate prediction methods are required for early treatment and would be helpful to decrease the spread of the disease. Previous studies have investigated the early prediction of HIV infection, mainly utilizing HIV positive patients' medical records. However, besides medical records, other characteristics, such as drug injection, sexual behavior, and other behavioral and biological factors, are also essential to consider while predicting the future acquisition of HIV. In this study, we developed and validated a prediction model for the probability of future HIV acquisition in high-risk groups. We developed prediction models using electronic medical records of patients from Nai Zindagi Trust (NZ), Pakistan. We analyzed the data for socio-demographic, behavioral (drug injection, sexual behavior), and biological features to predict the patients who are at high risk of acquiring HIV. State of the art data mining and machine learning techniques have been applied along with feature selection techniques for HIV risk prediction. The proposed model predicts HIV status with 82% accuracy having an improvement of 10–15% over dominant classifiers, i.e., SVM, Neural Network, J48, and PART. The new risk scores of HIV acquisition could assist health providers in counseling high-risk groups and in targeting intensified testing, treatment, and prevention to people at the greatest risk for HIV infection.

1. Introduction

HIV/AIDS emerged in the early years of the 1980 s. Ever since it has affected millions of people. The cause of AIDS is HIV is a lent virus, which slowly attacks the immune system of the human body, affects all systems of the body, and eventually resulting in death [1]. The most advanced stage of HIV is Acquired Immune Deficiency Syndrome (AIDS) [2]. Progression from HIV to AIDS can take many years, depending upon an individual's health condition. Unfortunately, there is no treatment available to cure this fatal disease until now [2]. The leading cause of HIV/AIDS is exchanging different body fluids between infected and non-infected people. These modes can be blood, semen, breast milk, or any sexual organ discharge. Certain behavioral risk factors could cause a greater risk of infection when contacting an HIV patient. These risk factors include unprotected sex, sharing contaminated syringes or needles, use of unsterilized equipment for medical procedures, tissue transplantation, and blood transfusion [1]. Besides these factors, specific groups of people such as sex workers, bisexuals, men who have sex with men (MSM), Female Sex workers (FSW), Male sex workers (MSW), transgender, gay, and people who inject drugs (PWID) [3,4] are always at high risk to acquire HIV. It is essential to control the spread of this fatal disease by targeting high-risk groups for testing, treatment, and prevention. PWID and MSM have overlapping risk behaviors, as both groups

* Corresponding author.
E-mail address: azhar.mahmood@cust.edu.pk (A. Mahmood).

Peer review under responsibility of Faculty of Computers and Information, Cairo University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.eij.2022.12.005
1110-8665/

are involved in sexual contact and needle sharing [5]. Such kinds of bridging phenomena have been observed in Central Asia, Russia, and Europe, where the HIV cases are more among injection drug users (IDUs) than over MSM, and subsequently to the general population. Pakistan is among four countries in Asia, along with Indonesia, Malaysia, and the Philippines, where HIV cases are increasing every passing year [6–8], and resultants death tolls have been gradually rising every year since 1986–87. In the last two decades, PWID and MSM have established overlapping risk behavior in Pakistan [9]. These communities are categorized as high-risk core groups. Core groups comprise a high prevalence of HIV. According to the 2017 fact sheets of UNAIDS, in Pakistan, 150,000 people are infected with HIV, out of 20,000 people are newly infected, and 6200 have lost their lives [10].

Since HIV slowly attacks the human body and expresses its symptoms at various stages; therefore, it is possible to detect the virus at earlier stages, which could help a patient to live a normal and healthy life for years. Several studies have been conducted, such as Integrated Biological and Behavioral Surveillance (IBBS) [11] and COHORT [12] for behavioral risk assessment in core risk groups. Incidence and Prevalence are the two foremost measures for estimating the risk rate of HIV/AIDS [13]. HIV prevalence is the ratio of the population living with HIV infection, while incidence is the yearly or the quarterly number of new infections that arise in a population. The ratio of HIV incidence to the size of the uninfected population is the HIV incidence rate, which is the risk of being infected per unit time. Several approaches are currently used to obtain national statistics on HIV prevalence, including national surveys [11,14], sentinel surveillance data collected from patients under treatment at public health facilities [15], such as pregnant women receiving care at public antenatal clinics; specialized surveys of selected high-risk subpopulations [11] (such as sex workers); and back-calculation methods that use data from national HIV/AIDS surveillance registries. The most challenging fact about collecting HIV/AIDS is the social stigma associated with HIV.

Knowledge discovery techniques have proved their impact on the medical and healthcare field [16]. In healthcare and prevention, predictive data mining is a relatively novel area of research, which blends sophisticated computing and representational techniques with expert's insight to develop tools for improving the healthcare domain especially in early prediction of the disease. This automation can optimize and support the administration of healthcare services provision and clinical care research. Furthermore, data mining models can analyze various parameters such as historical and geographical factors to help discover fundamental factors related to the problem from collected knowledge in a more meaningful manner [17].

Various research studies have been conducted on the classification of HIV status of an individual and individual's survival of HIV/AIDS by utilizing data mining and machine learning techniques. A novel predictive model based on Multilayer Perceptron (MLP) neural network is proposed in [18] that claim that behavioral, socio-demography information is sufficient besides clinical result for the prediction of individual HIV status. Another predictive model is proposed to predict the HIV status of an individual based on the MLP and radial basis function (RBF) methods. Cross-Industry Standard Process for Data Mining (CRISP-DM) has been proposed in [19]. The major contributions of CRISP-DM include the identification of determinant factors and parameters for optimal output from socio-demographic and sexual behavior data. CRISP-DM utilizes well-known data mining algorithms, such as J48, PART, and Naïve Bayes, for experimentation. A Naïve Bayes based model is proposed in [20] for predicting the durability of pediatric HIV/AIDS patients and getting HIV treatment. Another predictive model on individual's survival has been proposed in [21], which utilizes CART (Classification and Regression Trees). These existing models and techniques based on machine learning and data mining techniques are essential, however, HIV/AIDS prediction and patterns extraction is still a major research challenge, as various demographic and associated parameters vary with a country and its population's behaviors.

In this study, a predictive HIV-model for the prediction of future acquisition of HIV infection for high-risk population groups in Pakistan is proposed. According to a study conducted in [5], PWID's are the highest risk population in Pakistan. Furthermore, there is very likely that PWID - would share contaminated needles, equipment, containers and would contact their spouses and bridging population (FSWs, MSMs, and transgender), which might put a larger segment of the people at risk. Considering these facts, it is essential to develop early diagnostic tools that should identify high-risk people at the early stage of their disease. These tools will help health care providers in counseling high-risk groups and in targeting intensified testing, treatment, and prevention of the spread to other people.

Previous studies have investigated the early prediction of HIV infection, mainly utilizing HIV positive patients' medical records [18,19,22,23]. However, besides medical records, other characteristics, such as drug injection, sexual behavior, and other behavioral and biological factors, are also essential to consider while predicting the future acquisition of HIV. In this study, we developed prediction models using electronic medical records of patients from Nai Zindagi Trust (NZ), Pakistan. Compared to the previously presented work, which also utilizes a dataset provided by Nai Zindagi (NZ) Trust [24] which displays the properties of density and diversity only, furthermore, this dataset depicts extensive data abundance aiding in discriminative model learning of Pakistani population.

In this study, we used additional characteristics of the high-risk group (such as information of PWIDs along with the focused demography, sexual behavior, injecting, and biological features) along with existing literature features have been analyzed. Furthermore, we adopted more rigorous data processing and feature extraction techniques to predict the HIV risk scores accurately. The proposed model contributes to the existing research area by extending the standard models developed for HIV/AIDS prediction.

2. Literature review

Machine Learning classifiers are most widely used for medical applications to predict various diseases [22,23]. This section emphasizes the feasibility of applying predictive techniques of machine learning for HIV/AIDS prediction and survival analysis for several regions such as Ethiopia [25], Nigeria [18,19], Portuguese [26], Malaysia [21], South Africa [27], and China [28]. Furthermore, this section also identifies the major risk factors that might spread HIV with respect to regions. Mainly this section is classified into two categories: classify the HIV status of individuals and their survival.

2.1. Classification of HIV status for individual

This sub-section highlights the efficacy of the machine learning techniques along with socio-demographic, sexual behavior, and antenatal factors to predict the HIV/AIDS status of individuals using predictive models.

In [18], a predictive model based on Multilayer Perceptron (MLP) neural network is proposed. The genetic algorithm was used for classification and achieved an accuracy of 84.24 % [18]. However, in this work, social parameters such as the financial status of individuals and inter-mobility are ignored, which could improve

prediction accuracy. These limitations are taken into consideration by another Neural Network (NN) based prediction model [27]. This NN-based model predicts individual HIV status with inverse configuration. The dataset utilized in this study consists of the attributes from socio-demographic and antenatal factors, which helps in achieving an accuracy of 88 %. However, this model was also tested and validated on a very small dataset; thus, the accuracy attained by the model cannot be generalized for larger datasets. Another model on individuals' HIV status prediction proposed in [29] but instead of inverse NN, MLP and RBF methods are used with an accuracy of 61–62 %. However, the authors ignored the importance of demographical attributes, which resulted in low accuracy. In [30], authors have focused on Ethiopia for HIV status prediction of an individual as it's among the most affected sub-Saharan country. This article evaluated the prediction of individuals' HIV status of Addis Ababa by following the Cross-Industry Standard Process for Data Mining (CRISP-DM). One of the major contributions this study offers is that it determines the vital factors and parameters for optimal output of socio-demographic and sexual behavior data. Dataset used in this work contains attributes related to the socio-demography, behavioral and clinical result that makes this dataset a good candidate for model generalization. Well known classifier such as J48, PART, and Naïve Bayes are adopted for experimentation and achieved accuracy of 93.95 %. However, this work only considered the major population and minority class ignored because of unavailability of their record set.

Another study utilized decision trees and Random forest trees to develop a prediction model [28]. The main objective of this study is to use these methods for the identification of HIV infection in individuals and compare four methods used for predicting HIV. Authors of have targeted the high-risk groups of the population and used demographic, sexual behavior, and serological test result to evaluate their model. Random forest, KNNs, Decision Tree, along with SVM and achieved accuracy in the range of 79–98 % on three datasets, i.e., IDU, MSM, and FSW. Datasets used in this study obtained from the China CDC information system from 2009 to 2015. As this study considers only one state, so jurisdiction bias exists. A model is proposed to predict FSW's HIV status taking aid from behavioral and socio-demographic properties endorsing the importance of HIV prediction [31]. The proposed approach is adapted from CRISP-DM applied to the dataset of IBBS FSW (Ghana) 2015. Simulations performed on J48, Random Forest, Naïve Bayes, NNs, and Logistic Regression. Results proved that the proposed model extracts useful information for FSW HIV prediction. This study strongly emphasized that behavioral, as well as socio demographical attributes, hold sufficient weightage to predict target disease to improve the prediction accuracy. Table 1 shows the critical review of various approaches proposed for the prediction of the HIV status of individuals.

2.2. Classification of individual survival

This section presents the literature concerning the survival prediction of HIV/AIDS patients. A predictive model proposed in [20], based on Naïve Bayes, is used to predict the durability of pediatric HIV/AIDS patients (on treatment). This model claims to achieve accuracy from 60 to 100 %. In another study based on Classification and Regression Tree (CART) achieved an accuracy of 60–93 %. This study considers Nigeria as the case; Nigerians are suffering largely from HIV/AIDS and possible treatment known as anti-retroviral (ARV) drugs along with Highly Active drugs (HAART). C4.5 Decision Tree algorithm is a famous classification algorithm utilized in [32] for another predictive model to extract useful rules to define the association between attributes and target classes (survival of HIV/AIDS patients). The dataset consists of 216 HIV positive patients; training and validation were performed by using 10-fold-cross- validation approach with 99 % accuracy. Algorithm recognized infection instigated through pathogens, nutritional condition, and cluster of differentiation 4, (CD4) count as the most important indicators for target disease survival. The resultant model was similarly validated using a live historical dataset and got promising results. A novel approach of the fuzzy technique of regression has utilized in the model to predict AIDS survival using adaptive fuzzy in [33]. The model is based on a few variables, such as CD8, CD4, and the count of viral load. Model competency and performance comparison presented with fuzzy logic based neural network models. It has been observed that both models' accuracies range between 60 % and 100 %. This also highlighted the importance of viral load and T cell tests in HIV prediction. Table 2 shows the critical evaluation of various approaches adopted for the prediction of individual survival.

3. Predictive model for HIV/AIDS

This section discusses the proposed HIV- model, as shown in Fig. 1, for the prediction of HIV/AIDS status of the patients. The process starts with the extraction of raw data; dataset processing, which entails pre-processing steps such as attribute selection, transformation, reduction etc. These steps enable us to convert the data into a refined format that can be utilized later by the prediction classifier.

3.1. Data set description

The dataset used in this study is taken from the electronic medical record of HIV positive patients who are receiving services from Nai Zindagi (NZ) Trust. Two major data sources are explored for collecting required data i.e., NZ's management information system (MIS) and customized biometric devices provided by NZ.

The target population selected for experimentation purpose is PWIDs from 2012 to 2019, with 57 attributes, including demographic, biological, and behavioral behaviors. Demographic attributes include age, education, and marital status. Behavioral attributes are further divided into a few categories, such as injecting behavior, which includes syringes/needle usage on daily basis. Sexual behavior attributes include sexual relationships and sexual practices (condom usage). The predictive class has two groups: Negative and Positive. Negative status indicates that a person is not exposed to HIV yet, and positive status indicates that a person is reactive and has HIV.

3.2. Preprocessing and feature selection

Following techniques used for preprocessing of data.

3.2.1. Missing values

While analyzing the dataset records, it has been observed that there are missing values for a few features. In the HIV-model data imputation strategy to substitute, missing properties have been utilized. Since most of the attributes have over 50 % values, the modal value approach is used. This strategy allows the substitution of a missing value with a modal value after determining the attribute type. For instance, Professional Skill is an attribute consisting of 46,888 records containing 18,861 Yes, 27,827 No values. In this attribute, 200 records were missing, so a modal value of NA is used for substituting these empty cells. As for the numeric data, the statistical measure of mean is used to eliminate missing values.

3.2.2. Noise Removal/Outlier removal

Outliers are removed by extracting distinct values to determine if any column contains any abnormal value. By extracting distinct

Table 1
Classify HIV status of individual.

Ref	Participants	Objective	Data mining approaches	Dataset size	Data source	***Tools***	***Accuracy (Percentage)***
[30]	Mixed	HIV/AIDS prediction in healthcare (Individual)	J48, PAR, NB	78,324	HCT local data	WEKA, MS Excel, Java	J48 – 93.95 %,NB – 80.77 %,PART – 94.01 %
[25]	Men/ Women	HIV testing prediction	J48, NB, NN, LR	30,625	Ethiopia Demographic and Health Survey (2011)	SPSS, MS Excel, WEKA	RT – 96 %,J48 – 79 %,NN-78 %,LR – 74 %
[15]	Pregnant women	Classify HIV status of individual	MLP and RBF	12,945	South African seroprevalance survey (2001)	Net lab	MLP – 84.24 %
[26]	PWID, Hetero-Homosexual, Others	Factors identification in reporting delays of HIV/AIDS cases, Prediction of misreported cases	MLP, NB, SVM, KNN	45,000	Portuguese epidemic data HIV/AIDS	R language, Rapid miner	MLP – 62.98 %,NB- 52.90 %,SVM – 62 %
[27]	Spouses	Prediction of HIV status	GA, Inverse NN	12,945	South African seroprevalance survey 2001	Netlab toolbox MatLab	Inverse NN – 88 %
[29]	Spouses	Classify HIV status of individual	MLP, RBF	3381	South African seroprevalance survey 2001	Netlab toolbox MatLab	NN – 62 %
[34]	Pregnant women	Extract rules from HIV data, -predicted HIV status	FNN	16,744	South African Survey	NA	NA
[31]	FSW	predicts the HIV status of FSW	RT, J48, NB, LR, NN	3092	Ghana's 2015 FSW IBBS data	WEKA, MS Excel	RT- 98.9 %, J48 – 93.18 %,NB – 89.97 %
[28]	PWID, MSM, FSW	Identification of HIV status and compare predictive performance	RF, KNN, SVM,DT	21,731	Urumqi CDC Survey questionnaire	R Language	RF (MSM-94.48 %, FSW-97.51 %,IDU-94.63 %),KNN (MSM-91.52 %, FSW-96.30 %,IDU-90.82 %),SVM (MSM-94.01 %, FSW-98.03 %,IDU-91.35 %),DT (MSM-79.17 %, FSW-87.02 %,IDU-74.38 %)

Table 2
Prediction of individual survival.

Ref	Participants	Objectives	Method	Dataset Size	Data Source	Tools	Acc %
[20]	Paediatric	Survival of paediatric (age < 15)	NB	137	Local hospital dataset - Nigeria	WEKA, Cox Regression	NB – 60 %
[21]	N.A	Survival years of HIV/AIDS patients	CART, ROC	998	Local Hospital Malaysia	STATA 7	CART – 93 %
[32]	Paediatric	Survival of paediatric	C4.5 DT	216	Local Hospital	WEKA	C4.5–99.07 %
[20]	Male/ Female	Survival years of AIDS patients	AFRT, FuReA	997	Malaysia Hospital	MatLab	F NN – 60–100 %
[35]	Paediatric	Survival years of HIV/AIDS patients	SMO,SVM	215	Local Hospital	WEKA	SVM – 97.7 %

values, the attribute 'Donated-Sell-Blood' contained an abnormal value "2" which hold no significant meanings as the expected response can either be 'Yes' or 'No', so extraction of distinct values highlighted these values and simply replaced by adding to the response with maximum probability, which in this case is 'No'. After removing noise, the boxplot approach was applied to a few numeric attributes to highlight the outlier from the numeric data. For example, the attribute 'Education Years' contains the maximum value of 25, which surely is an outlier as it is not possible.

3.2.3. Label encoding

Label encoding is required to change all labels to numerical data, which is understandable for the training model.

3.2.4. Scaling and normalization

To normalize the dataset, unsupervised filtering technique is preferred, where the translation range of output is set to 0 and the scaling of output range is 1.

3.2.5. Data transformation and reduction

Few attributes are discretized to minimize the distinct attribute values to acquire (pattern) knowledge and style of the dataset appropriately for data mining techniques. In the case of values between huge ranges, it's considered appropriate to transform the values into a lower scale. The attribute named "HowDoYouMostlyUseThisDrug" has 102 distinct values. After transformation, these 102 distinct values, converted to seven (7) values, which are No Answer, Smoking, Sniffing, Other, Oral, Inhaling and Inject.

3.3. Features selection

The first approach used for feature selection is the domain expert's opinion. Domain experts identified potential independent/dependent 57 attributes and 33 attributes are selected, sorted and presented into three sets named as socio-demographical, behavioral, and biological.

Automated machine learning-based attribute selection technique Select-K-Best used where features are extracted on encoded and normalized data. Features selected from normalized data are more valuable along the expert opinion are higher in priority. Another feature selection technique, Extra Tree Classifier, is also used for attribute selection.

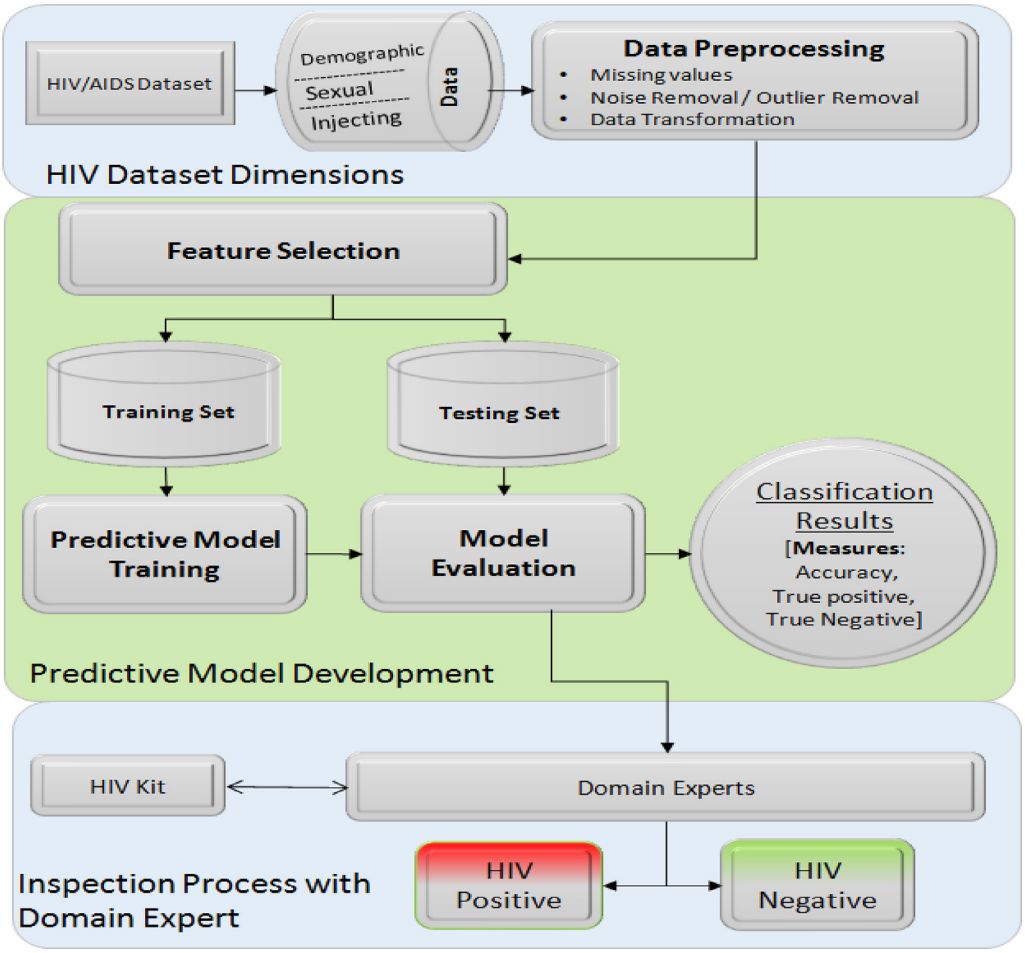


Fig. 1. Predictive Model for HIV\AIDS.

Table 3
Characteristics of Demographic Features.

States	Age	Gender
Live On Streets	Jailed For Drug Usage	Donated Sell Blood
Education Years	Source Of Income	Cities

Table 4
Characteristics of Injecting Behavior Features.

Syringe used by	SyringeOut-All	Syringe acquired from
Ever Injected	Injected by Street Doctor	SyringeOUT_Category
SyringeIN All	Exchange Rate	SyringeOUT_AVG
Daily Expense on Drug	How Use Drug	Syringe Out All

3.4. Target class attribute

This study focuses on the prediction of the status of HIV among males, females, and transgender of varying age groups. Therefore, the target attribute chosen in this study is Predictive status, which is either Positive or Negative.

3.5. Model training and classification

Several classification methods are test and train to predict HIV/AIDS status. Selected attributes received from the previous step set as input to the classification models. The classifier will return the HIV status of each individual. Classifiers utilized for this study are Random Forest (RF), PART Rule induction, and SVM.

4. Experiments and analysis

4.1. Data collection methods

With the consent of NZ Trust [24], the dataset is used to conduct research for the prediction of HIV\AIDS. The data collection mechanism used in NZ is Management Information System (MIS), which is used by 30 + CoPCs sites across Pakistan. These sites have their outreach worker who accessed PWID and provides services for prevention and treatment. Information is recorded on manual forms and later enter into MIS by Data Entry Operators via customized devices, these devices stores data in a real-time system.

4.2. Study population

Initially, 56,942 instances are selected, where 9832 instances had no class label hence used for testing the model. The final dataset consists of 47,110 numbers of instances from the year 2012 to 2019 with participant type of PWID, including dates of HIV diagnosis. Table 3 & 4 shows the characteristics of Demographic and Injecting Behavior features whereas Table 5, shows the summary of the features with distinct attributes and their number of occurrences.

4.3. Results and discussion

Analysis of the proposed model performed on the features selection, classification of HIV status, and effect of class balancing

Table 5
Characteristics of Sexual Behavioral Features with labeled class.

Sexual relation with whom	Condom consumed	Last intercourse
Relation With Any One	Used Condom	Treated For STI
Condom_AVG	Condom_Category	Class Label

of data on the classification process. Model validation is performed by three ways; firstly, all the attributes are exposed to domain experts to verify the status of HIV of an entity prior to train the model. Expert opinion is important for attribute selection. Secondly, HIV biological kit test is utilized in practice, which ensures HIV status of a person via blood samples. Thirdly, using the dataset from 2012 till June 2019 has been divided into three sets, including the training set (containing 21,496 instances), testing set (containing 10,203 instances), and validation set (containing 15,411 instances) (Table 6).

In Fig. 2, features extracted from Select-K-Best and Extra Tree classifiers are presented. The graph shows that by using Select-K-Best there are five high-scored attributes i.e., SyringAcquiredFrom, Age, FrequencyOfDailyInjecting, MaritalStatus, SexualRelationWithAnyOne. The five most scored attributes extracted from Extra-Tree classifiers are MaritalStatus, SyringAcquiredFrom, Age, JailedForDrugUsage, FrequencyOfDailyInjecting. According to the domain experts, these attributes must be of high priority as we are dealing with IDUs. Not only attributes related to their drug usage or injected behavior are important, but also their MaritalStatus and Age has significance.

As the dataset is biased to negative cases, SMOTE technique has been utilized to balance dataset and avoid.

Table 6
Dataset distribution for training, testing and validation.

	Year	Total cases		Positive cases		Negative cases	
Validation	2019	9989	15,411	1786	4494	8203	10,917
Set	2018	5422		2708		2714	
Training &	2017	5025	10,203	2912	5874	2113	4329
Testing	2016	5178		2962		2216	
	2015	7662	21,496	3078	10,244	4584	11,252
	2014	7513		3968		3545	
	2013	3866		1979		1887	
	2012	2455		1219		1236	
Grand Total		47,110		20,612		26,498	

biased prediction. Several algorithms were exploited to predict Individual's HIV status, however few of their results are depicted in Fig. 3, such as MLP, Naïve Bayes, KNN, J48, PART, Logistic Regression, SVM and RF.

Fig. 3 shows the different accuracies of the aforementioned classifiers. The y-axis shows the accuracy percentages, and the x-axis shows features efficacy in the HIV status prediction. Features have presented to the proposed model as aggregated features, independent features, and a combination of features (such as demographical and injecting behavior). Furthermore, along with the x-axis well-known classifiers applied to predict the HIV status based on the presented features for better analysis.

It has been observed that MLP achieves the second highest accuracy among all the classifiers tested on the provided dataset. Logistic regression measures the association between independent variables (set of attributes such as Age, Gender, etc.) and dependent variable (class label which Predict HIV Status), via logistic function for estimation of probabilities. Logistic Classifier has achieved an accuracy of 74.2. Naïve Bayes, achieves an accuracy of 71.5 %, which is lower than MLP and Logistic regression. Although Naïve Bayes is simple to implement; however, it tends to suffer from loss of accuracy, also known as zero frequency. J48 or C4.5 algorithm is based on decision trees and does not require normalization; however, rules pruning is heuristic-based. It has achieved an accuracy of 73.9 % against all attributes; however, it shows better accuracy in demography combined with injecting and sexual behavior, respectively. SVM (Support Vector Machine) is a famous binary classifier, and the problem under consideration is binary in nature. SVM has achieved an accuracy of 73.5 % for all features; however, as the dimensions in the given problem increase, SVM does not perform well. PART is a classifier that is based on rule induction and achieves a prediction accuracy of 72.1 % across all features. Another classifier called instance-based learning using k parameter (IBK) has achieved an accuracy of 70.7 %.

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

The proposed HIV model is based on Random Forest and shows better accuracy of 76.5 % compared to all well-known classifiers in this domain. HIV model emphasizes on the importance of selected features in a more manageable manner as it consists of multiple decision trees. It has been observed that the HIV model maintains

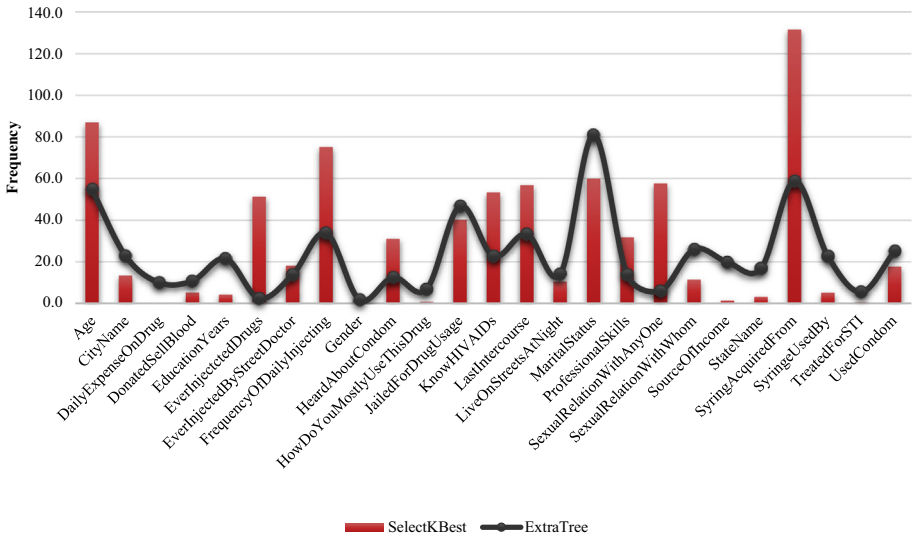
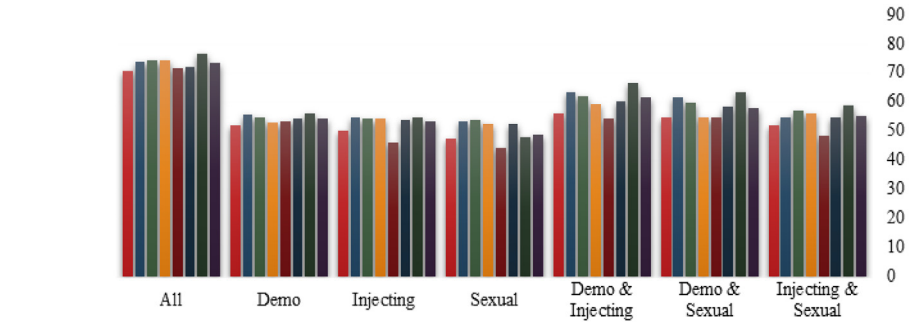


Fig. 2. Feature comparison of K-Best and Extra Tree Classifiers.



	All	Demo	Injecting	Sexual	Demo & Injecting	Demo & Sexual	Injecting & Sexual
IBK (KNN)	70.7	52	50	48	56	55	52
J48	73.9	56	55	53	63	61	55
Logistic	74.2	55	54	54	62	60	57
MLP	74.4	53	54	52	59	55	56
NaiveBayes	71.5	54	46	44	54	55	48
PART	72.1	54	54	53	60	58	55
HIV-Model	76.5	56	55	48	67	63	59
SVM	73.5	54	53	49	61	58	55

Fig. 3. Comparison of classifiers with respect to dimensions.

better accuracy in combined features analysis; accuracies achieved in independent features are reasonable; however, the result observed over Sexual feature (48 %) is lower than various other classifiers used for accuracy performance. The objective of this comparison is to discuss a better prediction model and the importance of features to predict HIV status.

As discussed earlier, the dataset is imbalanced; class balancing is performed by utilizing the SMOTE technique to avoid biased classification results. After class balance, the total numbers of instances are 19905. The proposed model correctly classified 16,323 (82.36 %), while incorrectly classified cases were 3544 (17.74 %). This 17.74 % of cases were sent to domain experts to verify the status as positive or negative. It has been observed that class balancing has further improved prediction accuracy. Furthermore, Table 7 explains the model-building time of each classifier; it has been observed that HIV model took less time than MLP and IBK, however greater than the rest of the classification model.

Evaluation is the most crucial step while dealing with classifiers and also serve as the key to creating actual progress, since it provides us a rich view concerning the strength of every model discretely, and empower us to make comparison among various models on the bases of widely used performance matrices. It also helps in determining whether the model is working well or not. The model's performance is analyzed using a series of performance measures. For the evaluation HIV model in this study, a standard metric of performance called confusion matrix is used. The confusion matrix gives information such as specificity (True Negative Percentage), Sensitivity (True Positive Rate), accuracy, F-measure, Recall, Precision, and Receiver Operating Characteristics Curve. Every measure stated here is used where required in performance analysis. True Positive Rate (TPR), is an important evaluation metric of the classification when it comes to the healthcare applications. TPR of a classification technique explains that the actual positive will be positive. The significance increases as a patient with positive HIV should not or with minimum probability should be classified as negative.

Cross-validation of 10-folds approach is adopted to train and validate RF, J48, PART and SVM. Below equation shows formulas for computation of True Positive, which denotes positive rows of the right classification, True Negative denotes the negative rows of right classification, False Positive denotes the positive rows of the wrong classification, and False Negative denotes the negative rows of the wrong classification. The proposed technique HIV-Model (Random Forest), shows promising results in TPR, as it has the highest TPR value when compared with the existing techniques. HIV-Model (Random Forest) TPR results suggest that it is classifying actual positive will be positive with higher confidence. Also, as far as accuracy is concerned, it means correctly predicted sample count. Other than evaluation using these performance metrics, performance is also computed to measure efficiency. It defines the execution time a classifier takes for training and generating prediction results to necessitate required resources to compute.

For inspection, the validation set is utilized and balanced using SMOTE to avoid biased classification results. After class balance, the total numbers of instances are 19905, as shown in Fig. 4. HIV model correctly classified 15,748 (82.36 %), while incorrectly classified cases were 4157 (17.64 %). These 17.64 % of cases sent to domain experts to verify the status as positive or negative and to review details of these incorrectly classified cases and make decisions for their HIV status. A visualization tool is developed during

Table 7
Classifiers results before and after class balancing.

NZ - Before class balancing	Accuracy (%)	Time(s)	TP Rate	FP Rate
HIV-Model(RandomForest)	76.5	0.56	0.77	0.23
MultilayerPerceptron	74.4	1637.65	0.74	0.26
Logistic	74.2	0.1	0.74	0.26
J48	73.9	0.02	0.74	0.26
SMO	73.5	0.04	0.73	0.27
PART	72.1	0.9	0.72	0.28
NaiveBayes	71.5	0.28	0.71	0.29
IBK (KNN)	70.7	17.08	0.71	0.29
BayesNet	69.6	0.11	0.70	0.30
RandomTree	68.6	0.01	0.69	0.31
NZ - After class balancing	Accuracy (%)	Time(s)	TP Rate	FP Rate
HIV-Model(RandomForest)	82.3619	13.64	0.82	0.18
J48	78.3779	0.02	0.78	0.22
PART	77.9785	0.48	0.78	0.22
BayesNet	77.7123	0.04	0.78	0.22
IBK (KNN)	74.5174	31.49	0.75	0.25
RandomTree	74.1561	0.01	0.74	0.26
MultilayerPerceptron	73.8361	3273.25	0.74	0.26
SMO	73.2148	0.08	0.73	0.27
Logistic	72.7679	0.04	0.73	0.27
NaiveBayes	66.1215	0.13	0.66	0.34

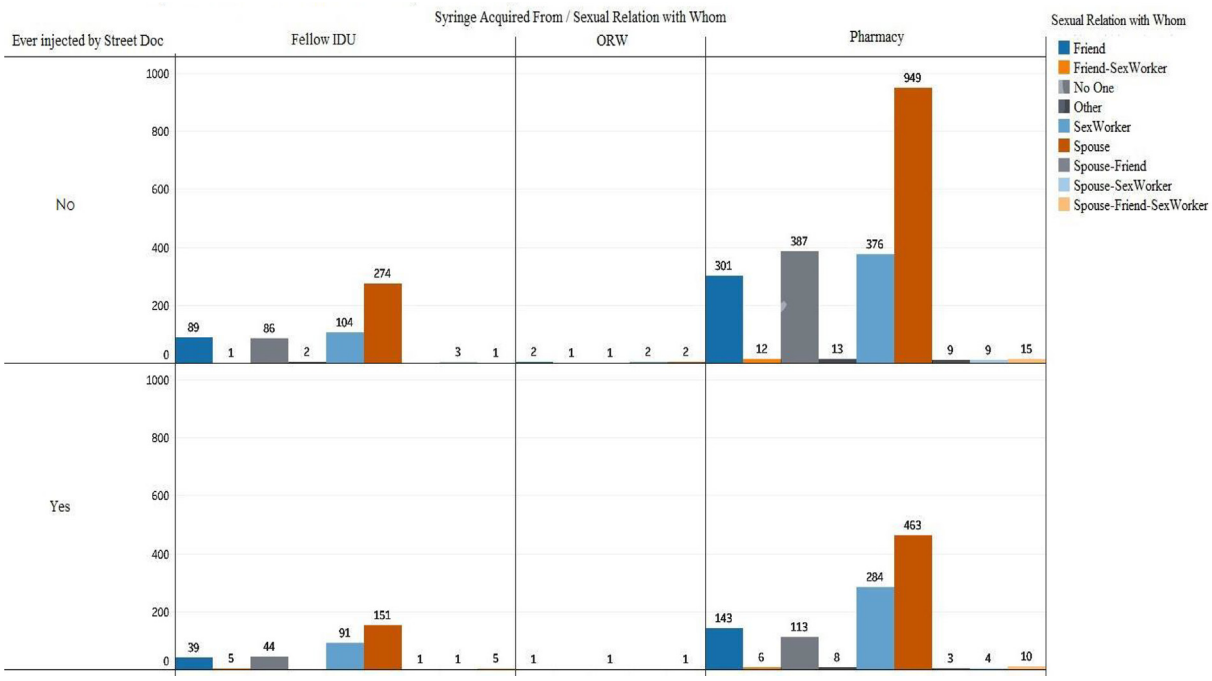


Fig. 4. Inspection on validation set.

this research, which makes the decision-making process convenient for the experts.

5. Conclusion

HIV/AIDS is a problem that keeps on progressing regardless of geography, community, age, and gender. Prediction of HIV/AIDS is complicated and requires careful handling to generate better results. Patients' life span can be extended by predicting the disease in earlier stages and helping them to live a normal and healthy life. Another concern associated is with the collection of data because of social stigma, people are reluctant to provide inaccurate details about their sexual and social behaviors. Measures have been taken for the accountability of disease burden, many cohort studies are conducting, and surveillance data has been analyzed for trends in any community, estimation for future trends, better implementation of programs, and policymaking. HIV shows its symptoms at various stages; thus, it is possible to predict the probability of infection that a patient has.

Previous research has raised concerns about the validation and small sample size of the utilized datasets. The Nai Zindagi (NZ) organization provides an optimum size dataset of HIV/AIDS patients. NZ claims that 95 % of the information of participants is accurate as NZ is tracking their participants since 2012 by an outreach program. Another limitation of the previous studies is the optimum attributes (features) selection, which requires smart techniques along with health expert's opinion. In this study, domain experts are involved explicitly in the attribute selection and reduction phase to include only those attributes that are worthy of being included for the prediction of HIV/AIDS. Another contribution in this study to introduce additional features such as information of PWIDs along with the focused demography, sexual behavior, injecting, and other biological features for comprehensive analysis. Information about PWIDs is vital in the prediction of the HIV/AIDS status of a patient as PWIDs have bridging potential to spread HIV to the general population.

For smart and accurate HIV/AIDS patient's status prediction, this study utilizes and analyzes the state-of-the-art machine learning techniques for the preprocessing and classification such as Naïve Bayes, NN, SVM, Random Forest, and ensembles. These techniques are powerful and effective as compared to traditional techniques and can predict far more precision and accuracy than previous techniques. These approaches also have their limitations where validation of results is challenging in addition to the challenge of data gathering. Hence, the aim of this study, to conduct a comparative analysis of studies focusing on HIV/AIDS prediction using knowledge discovery approaches. Mainly proposed an HIV-model to predict HIV status to scale up the acquired knowledge extracted of HIV status with 82 % accuracy using selected features. We can apply results in Pakistan but not in other regions due to social, geographical, and health policy factors.

Conflict of interest

The authors declare no conflict of interest.

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.

Acknowledgement

We are very thankful for the Nai Zindagi, Pakistan (https://www.naizindagi.org/) to provide HIV\AIDS data and also the possible assistance in this research study.

References

[1] (15th June). *What Are HIV and AIDS?* Available: https://www.hiv.gov/hiv-basics/overview/about-hiv-and-aids/what-are-hiv-and-aids.
[2] WHO. (16th June). *HIV/AIDS*. Available: https://www.who.int/en/news-room/fact-sheets/detail/hiv-aids.
[3] C. f. D. C. a. Prevention. *CDCP overview* Available: https://www.cdc.gov/hiv/default.html.
[4] UNAIDS. *Global HIV & AIDS statistics — 2020 fact sheet*. Available: https://www.unaids.org/en/resources/fact-sheet.

[5] Khanani MR, Somani M, Rehmani SS, Veras NM, Salemi M, Ali SH. The spread of HIV in Pakistan: bridging of the epidemic between populations. PLoS One 2011;6:e22449.
[6] M. D. a. Health). *Partners*. Available: https://english.mainline.nl/page/partners.
[7] UNAIDS, "UNAIDS DATA 2018," UNADIS2018.
[8] T. Brown and W. Peerapatanapokin, "HIV/AIDS in Asia: we need to keep the focus on key population groups," 2019.
[9] Awan S, Zia N, Sharif F, Shah SS, Jamil B. Types and risk factors of violence experienced by people living with HIV, Pakistan: A cross-sectional study. East Mediterr Health J 2020;26.
[10] UNAIDS. (2017). *Country Fact Sheet- Pakistan*. Available: https://www.unodc.org/unodc/en/hiv-aids/new/drug-use_and_HIV.html.
[11] (2017). *Integrated Biological and Behavioral Surveillance in Pakistan 2016-17. National AIDS Control Program. (2017)*. Available: Available: https://www.aidsdatahub.org/integrated-biological-and-behavioral-surveillance-pakistan-2016-17-national-aids-control-program.
[12] Ansari JA, Salman M, Safdar RM, Ikram N, Mahmood T, Zaheer HA, et al. HIV/AIDS outbreak investigation in Jalalpur Jattan (JPJ), Gujrat, Pakistan. J Epidemiol Global Health 2013;3:261–8.
[13] Brookmeyer R. Measuring the HIV/AIDS epidemic: Approaches and challenges. Epidemiol Rev 2010;32:26–37.
[14] T. Reza, D. Y. Melesse, L. A. Shafer, M. Salim, A. Altaf, A. Sonia, G. C. Jayaraman, F. Emmanuel, L. H. Thompson, and J. F. Blanchard, "Patterns and trends in Pakistan's heterogeneous HIV epidemic," *Sexually transmitted infections*, vol. 89, pp. ii4-ii10, 2013.
[15] A. B. Kharsany, J. A. Frohlich, N. Yende-Zuma, G. Mahlase, N. Samsunder, R. C. Dellar, M. Zuma-Mkhonza, S. S. A. Karim, and Q. A. Karim, "Trends in HIV prevalence in pregnant women in rural South Africa," *Journal of acquired immune deficiency syndromes (1999)*, vol. 70, p. 289, 2015.
[16] Ujager FS, Mahmood A. A context-aware accurate wellness determination (CAAWD) model for elderly people using lazy associative classification. Sensors 2019;19:1613.
[17] Chen S, Bergman D, Miller K, Kavanagh A, Frownfelter J, Showalter J. Using applied machine learning to predict healthcare utilization based on socioeconomic determinants of care. Am J Manag Care 2020;26:26–31.
[18] B. Leke-Betechuoh, T. Marwala, T. Tim, and M. Lagazio, "Prediction of HIV status from demographic data using neural networks," in *2006 IEEE International Conference on Systems, Man and Cybernetics*, 2006, pp. 2339-2344.
[19] G. S. Lakshmi and P. I. Devi, "Prediction of anti-retro viral for HIV and STD patients using data mining technique," in *2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS)*, 2017, pp. 1-6.
[20] I. P. Adebayo, A. Olutola, and T. Aladekomo, "The Prediction of Paediatric HIV/AIDS Patient Survival: A Data Mining Approach," *Asian Journal of Computer and Information Systems*, vol. 4, 2016.
[21] Kareem SA, Raviraja S, Awadh NA, Kamaruzaman A, Kajindran A. Classification and regression tree in prediction of survival of aids patients. Malays J Comput Sci 2017;23:153–65.
[22] S. G. Jacob and R. G. Ramani, "Data mining in clinical data sets: a review," *IJAIS-ISSN: 2249-0868 Foundation of Computer Science FCS, New York, USA*, vol. 4, 2012.
[23] Bisaso KR, Anguzu GT, Karungi SA, Kiragga A, Castelnuovo B. A survey of machine learning applications in HIV clinical research and care. Comput Biol Med 2017;91:366–71.
[24] N. Z. TRUST. Available: https://www.naizindagi.org.
[25] Hailu TG. Comparing data mining techniques in HIV testing prediction. Intell Inf Manag 2015;7:153.
[26] Oliveira A, Faria BM, Gaio AR, Reis LP. Data mining in HIV-AIDS surveillance system. J Med Syst 2017;41:51.
[27] Betechuoh BL, Marwala T, Tettey T. Using inverse neural networks for HIV adaptive control. Int J Comput Intell Res 2007;3:11–5.
[28] Tang D, Zhang M, Xu J, Zhang X, Yang F, Li H, Feng L, Wang K, Zheng Y. Application of data mining technology on surveillance report data of HIV/AIDS High-Risk Group in Urumqi from to 2015. Complexity 2009;2018:2018.
[29] T. N. H. Tim, "Predicting HIV status using neural networks and demographic factors," 2006.
[30] Zewdu T, Beshah T. Prediction of HIV Status in Addis Ababa using Data Mining Technology. HiLCoE J Comput Sci Technol 1998;2:65–71.
[31] D. A. Annang, "Performance Comparison of Data Mining Techniques for Predicting Hiv Status Among Female Sex Workers in Ghana," University Of Ghana, 2018.
[32] Aladekomo TA. Survival model for pediatric HIV/AIDS patient using C4. 5 decision tree algorithm. Int J Child Health Human Develop 2017;10:143.
[33] Dom RM, Kareem SA, Abidin B, Kamaruzaman A, Kajindran A. The prediction of AIDS survival: a data mining approach. In: In *ProceedIngs of the 2nd WSEAS International Conference on Multivariate Analysis and its Application In Science and EngIneerIng*. p. 48–53.
[34] Tettey T, Nelwamondo FV, Marwala T. HIV data analysis via rule extraction using rough sets. In: in *2007 11th International Conference on Intelligent Engineering Systems*. p. 105–10.
[35] C. Olayemi Olufunke, O. Olasehinde Olayemi, and O. Agbelusi, "Predictive Model of Pediatric HIV/AIDS Survival in Nigeria using Support Vector Machine," *Communications*, vol. 5, pp. 29-36.