Predicting CD4 count changes among patients on antiretroviral treatment: Application of data mining techniques

Mihiretu Kebede, Desalegn Tegabu Zegeye, Berihun Megabiaw Zeleke

DOI
10.1016/j.cmpb.2017.09.017

Published in
Computer Methods and Programs in Biomedicine

Document version
Accepted manuscript

Online publication date
21 September 2017

Corresponding author
Mihiretu Kebede

Citation
Kebede M, Zegeye DT, Zeleke BM. Predicting CD4 count changes among patients on antiretroviral treatment: Application of data mining techniques. Comput Methods Programs Biomed. 2017;152:149-57.

Predicting CD4 count changes among patients on Antiretroviral Treatment: Application of data mining techniques

Mihiretu Kebede , Desalegn Tigabu Zegeye ,
Berihun Megabiaw Zeleke

PII: S0169-2607(17)30127-X
DOI: 10.1016/j.cmpb.2017.09.017
Reference: COMM 4503

To appear in: *Computer Methods and Programs in Biomedicine*

Received date: 6 February 2017
Revised date: 6 September 2017
Accepted date: 14 September 2017

Please cite this article as: Mihiretu Kebede , Desalegn Tigabu Zegeye , Berihun Megabiaw Zeleke , Predicting CD4 count changes among patients on Antiretroviral Treatment: Application of data mining techniques, *Computer Methods and Programs in Biomedicine* (2017), doi: 10.1016/j.cmpb.2017.09.017

Highlights

- Predicted CD4 count changes using J48, neural network and random forest data mining algorithms
- Compared the predictive performance of three data mining algorithms
- Identified variables important for prediction of CD4 count changes

Predicting CD4 count changes among patients on Antiretroviral Treatment: Application of data mining techniques

Mihiretu Kebede[1, 2*], Desalegn Tigabu Zegeye[3], Berihun Megabiaw Zeleke[4,5]

[1]Leibniz Institute for Prevention Research and Epidemiology – BIPS, Achterstraße 30, Bremen, Germany

[2]University of Gondar, Institute of Public Health, Department of Health Informatics, Gondar, Ethiopia

[3]Health Policy and Planning Directorate, Ethiopian Federal Ministry of Health

[4]Department of Epidemiology and Preventive Medicine, School of Public Health and Preventive Medicine, Monash University, Melbourne, Australia

[5]Department of Epidemiology and Biostatistics, Institute of Public Health, University of Gondar, Gondar, Ethiopia

***Corresponding Author**

Mihiretu Kebede
Achterstraße 30
28359, Bremen
Germany

Email: mihiretaabush@gmail.com

Abstract

Background and Objectives: To monitor the progress of therapy and disease progression, periodic CD4 counts are required throughout the course of HIV/AIDS care and support. The demand for CD4 count measurement is increasing as ART programs expand over the last decade. This study aimed to predict CD4 count changes and to identify the predictors of CD4 count changes among patients on ART.

Methods: A cross-sectional study was conducted at the University of Gondar Hospital from 3,104 adult patients on ART with CD4 counts measured at least twice (baseline and most recent). Data were retrieved from the HIV care clinic electronic database and patients' charts. Descriptive data were analysed by SPSS version 20. Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology was followed to undertake the study. WEKA version 3.8 was used to conduct a predictive data mining. Before building the predictive data mining models, information gain values and correlation-based Feature Selection methods were used for attribute selection. Variables were ranked according to their relevance based on their information gain values. J48, Neural Network, and Random Forest algorithms were experimented to assess model accuracies.

Result: The median duration of ART was 191.5 weeks. The mean CD4 count change was 243 (SD 191.14) cells per microliter. Overall, 2427 (78.2%) patients had their CD4 counts increased by at least 100 cells per microliter, while 4% had a decline from the baseline CD4 value. Baseline variables including age, educational status, CD8 count, ART regimen, and haemoglobin levels predicted CD4 count changes with predictive accuracies of J48, Neural Network, and Random Forest being 87.1%, 83.5%, and 99.8%, respectively. Random Forest algorithm had a superior performance accuracy level than both J48 and Artificial Neural Network. The precision, sensitivity and recall values of Random Forest were also more than 99%.

Conclusions: Nearly accurate prediction results were obtained using Random Forest algorithm. This algorithm could be used in a low-resource setting to build a web-based prediction model for CD4 count changes.

Key words- CD4 count change; Antiretroviral Treatment; computational methods; Random Forest; Neural Network; J48, Decision tree

Introduction

Fighting the global burden of HIV/AIDS requires a concerted diagnosis and treatment support that includes global expansion of Antiretroviral Treatment(ART) programs, access to CD4 count and viral load monitoring [1,2]. Throughout the comprehensive course of HIV/AIDS care and support programs, CD4 counts are used to decide HIV treatment initiation, to monitor disease progression and to assess treatment response in both adults and children [3]. CD4 count is mainly helpful to assess immunological response particularly for patients having limited access to viral load monitoring which is helpful to monitor viral suppression [4,5]. After tested for HIV, positive results must have their CD4 counted for assessing the eligibility to start ART [2,5,6].

Once a patient started ART, taking the treatment throughout life is essential. However, due to the associated side effects, long-term continuation and optimal adherence to the treatment remained a key challenge. Therefore, the WHO guidelines recommend the time for ART initiation to be based on CD4 count and WHO clinical staging. In western countries, plasma viral load (viral load rising above 10,000 copies/μl) and CD4 counts are main parameters to monitor disease progression and response to treatment [4,7]. A wider access to monitoring tools particularly CD4 count and viral load testing for patients on ART are important [2,3,5]. However, in low income settings, the limited access to advanced laboratory investigations made the decision making for ART initiation and follow up of HIV/AIDS patients to mainly depend on clinical staging [3,7]. A study from South Africa, however, revealed that treatment shall be initiated independent of CD4 count or viral load results [8].

Data mining refers to the non-trivial extraction of hidden, previously unknown and potentially useful information from data in databases. [9]. Unlike traditional statistical techniques, data mining is mainly about learning from data rather than turning data into information [10]. The subtasks of data mining use different mathematical data mining algorithms or models. For instance, for the task of classification and prediction like decision tree algorithms, Neural Network, and regression analysis can be used [9,11]. It has been applied in variety of fields such as business management, banking, and administration [12]. Relative to this, literature on application of data mining in medicine is scant. A study from India applied Multi-Layer Feed Back Ward Neural Network back propagation algorithm to predict how long a patients can extend their life with specific ART regimens [13]. A study report from South Africa indicated artificial Neural Network

algorithm almost accurately predicted HIV status from demographic and socio-economic variables of the sero-prevalence survey information such as age, education, location, race, parity and gravidity [14]. Fuzzy regression technique (FuReA) and fuzzy Neural Network models length of survival of AIDS patients based on their CD4, CD8 and viral load counts with 60% to 100% ranges of accuracy levels based on selected dependent variables [15]. A study from Malaysia reported, Classification and Regression Tree (CART) predicted survival of HIV/AIDS patients with an accuracy of 60-93% [16]. CART also predicted Global HIV/AIDS prevalence patterns with 95% level of accuracy [17].

Other attempts to apply data mining algorithms in HIV/AIDS medicine shows, ANN, support vector machine (SVM) and Random Forest predicted viroligic response to combination HIV therapy with comparable level of accuracy to a committee of ANN models [18].Providing CD4 count service has been challenging in developing countries particularly regarding technology selection, developing laboratory infrastructure, human resource, cost-effectiveness, instrument cost and maintenance, and ensuring testing access. A recent survey has reported lack of machine installations, frequent breakdowns, lack of timely and proper maintenance and lack of reagents as major challenges of CD4 count service provision [19]. Available big medical data can be used to develop predictive data mining models in many contexts such as forecasting infant mortality [20], HIV/AIDS disease progression and CD4 counts [15,21,22]. Therefore, this study aimed to predict the CD4 count changes and to identify potential predictor variables using the data mining approaches.

Methods

Study design, area and period

Cross sectional study was conducted at the University of Gondar Hospital. University of Gondar Hospital started free ART service since March 2005.

All HIV positive patients registered in the ART clinic were the source population of this study. All adult (age greater than 14) HIV/AIDS patients on ART registered in the ART database and having baseline and follow up CD4 counts were included in this study. Patients who started ART and their information is incomplete, unreadable or their manual record is lost, patients who do

not have at least one follow up CD4 count measure were excluded from the study. In addition, patients who are transferred in for ART service from elsewhere were not included in this study because their data other than their baseline and recent CD4 count, required for the purpose of this study were incomplete.

Of the 6,444 patients ever enrolled at the ART clinic of the university hospital. Overall, 3888 (1561 males, 2327 females) were on ART (Moreover, 3412 (87.8%) were taking first line ART regimens. We were not able to retrieve the record of 784 (20%) of patients because of difficulty to acquire their record, have no current and baseline CD4 count values and patients who were transferred from other ART centers with incomplete ART. data. Hence, the final analysis was based on 3104 (79.8%) patients.

The ART clinic is linked with laboratory department which has CD4 counting machines (1 CELL-DYN and 1 FACS Calibur CD4 counting machines).

Data collection procedures

A spreadsheet containing de-identified patients' ART-related data was obtained from the University Hospital's ART database unit. Variables not included in the database were manually extracted from the paper based patient charts.

Data quality

The principal investigator provided training about extraction of data to data collectors and supervised the overall data collection process. While filling the manual spreadsheets, data were crosschecked with both the electronic and the paper based documents for similarity and consistency.

The outcome variable (Class Attribute) of this study was CD4 count change. CD4 count change was calculated by subtracting the most recent CD4 count from the baseline. The change in CD4 count was categorized in 4 categories i.e $\Delta CD4<0$, $0 \leq \Delta CD4 \leq 50$, $51 \leq \Delta CD4 \leq 100$, and $\Delta CD4 > 100$. The independent variables included in the study were sociodemographic variables such as age, sex, religion, marital status, educational status and occupation. ART and HIV care variables included in the study were baseline CD4 count, functional status (working, ambulatory, bedridden) at the start of ART, WHO clinical staging, baseline hemoglobin, baseline liver and renal function test results, baseline weight, pregnancy status during enrolment and

most recent CD4 count. In addition, type of baseline regimen, type of current regimen, regimen change, and number of weeks on ART were also included.

Data processing and analysis

Data retrieved from the ART database and manually collected from patient charts were appended using medical record numbers and one spreadsheet form was created. This excel data was converted to Comma Separated Version (CSV) and then to the Attribute-Relation File Format (ARFF). The ARFF format data were then imported to WEKA 3.8 for analysis.

CRISP Methodology for data mining

CRISP methodology for data mining was followed to conduct this study. CRISP was launched in late 1996 by three "veterans" of data mining market Daimler Chrysler (then Daimler-Benz), SPSS (then ISL) and NCR. In order for the data mining process to be reliable and standardized, the CRISP methodology for data mining recommends step-by-step procedures. This methodology follows a cyclic process that includes business understanding, data understanding, data preparation, model building evaluation and deployment [23].

Business Understanding

The first step in the CRISP data mining sequence is to deeply understand in the view of business perspective [23].

HIV is a retrovirus, hence a drug against this virus is known as Anti-Retroviral drug, in short ARV. Giving this drug in a correct with support for adherence and continuous monitoring is known by the name Anti-Retroviral Treatment (ART). ART has many advantages such as prolonging and improves the life of the patients, decrease morbidity, mortality and hospitalization and prevention of mother to child transmission of HIV. The most commonly used antiretroviral drugs in Ethiopia Nucleoside reverse transcriptase inhibitors (NsRTI), Nucleotide Reverse Transcriptase Inhibitor (NtRTI), Non-nucleoside reverse transcriptase inhibitors (NNRTI) and Protease Inhibitors (PI). NsRTI includes Stavudine (d4T), Lamivudine (3TC), Zidovudine (AZT), Didanosine (ddI), and Abacavir (ABC) while NtRTI are mainly Tenofovir Disoproxil Fumarate (TDF). NNRTI are Nevirapine (NVP) and Efavirenz (EFV). In addition PIs are Saquinavir (SQV/r), Ritonavir (RTV), Indinavir (IDV/r), Nelfinavir (NFV) and Lopinavir (LPV/r)[24]

Protease Inhibitors (PI): ART service is one of the comprehensive cares provided within the HIV care. The comprehensive cares include Voluntary Counseling and testing (VCT), Prevention of mother to Child Transmission (PMTCT), Tuberculosis (TB), and prevention and management of Sexually Transmitted Diseases (STDs). These services are linked to each other.

When patients are linked to ART services, they are assessed for eligibility. Eligibility for ART is mainly guided by immunological and clinical criteria. Before initiating treatment, providers make sure that patient preparations included counselling for optimum adherence to treatments are performed and major opportunistic infections are well treated. Adherence is defined as "accepting, agreeing and correctly following a prescribed treatment (participation of the patient)" [24].

Data Understanding

After understanding the overall business process going on in the ART clinic and identifying business objectives, the next step in the CRISP methodology is to understand the data. To apply this step, we started with an initial collection of ART data and followed by activities necessary to be familiar with this data. The data retrieved from the ART database did not include all relevant variables. Hence, we collected data manually from the patient folders. One excel spreadsheet was produced by conjugating the manually collected data with the electronic ART database.

Data Description

The excel spread sheet contains 3104 records and 36 variables. Variables such as name, data type and the possible values were investigated. Age was categorized according to the WHO standard for age classification to simplify the data mining task and issues related with memories that could delay the data mining process.

The main outcome variable was CD4 count change. CD4 count change was calculated by subtracting the most recent CD4 count from the baseline CD4 count. It was then categorized into four categories.

Category 1(A) is when a patient is on ART and have a CD4 count change greater than 100, Category 2(B) is when a patient is on ART and has a CD4 count change between 51 and 100, Category 3(C) is when a patient is on ART and has a CD4 count change between 1 and 50, Category 4 (D) is when a patient was on ART and has a CD4 count change less than zero.

Data Preparation for Analysis

The data identified for data mining was collected, preprocessed, assessed, consolidated, cleaned, transformed, missing values handled and changed to appropriate format to be recognized WEKA. The data from ART clinic were stored as Ms Access. This Ms Access database was imported to Ms Excel format. Using the medical record number of each patient, the data from patient charts were recorded for variables which were not available from the electronic database. Descriptive statistics was performed using SPSS version 20The completed excel data were changed into CSV comma delimited form. WEKA 3.8 data mining software was used to build the models [25,26]. It analyses data in the form of attribute relation format. Therefore the CSV comma delimited format file was transformed in to Attribute-Relation File Format (ARFF) by opening the CSV format using Microsoft note pad program, then after setting the relation name, attribute and its values as a header in the data, it was saved as ARFF file. .

Results

Socio-demographic Profile

Out of the 6,444 patients ever started ART, 3104 (48.2%) were included in this study. Their mean (SD) age was 33.5 ±8.6 years and 60.9% were females. Nearly half (49%) were married and a third (35.2%) were employed (Table 1).

Table 1: Characteristics of patients on ART at the University of Gondar Hospital

Variables		Count	%
Sex	Female	1889	60.9
	Male	1215	39.1
Age	15-19 (A)	3	0.1
	20-24 (B)	5	0.2
	25-29 (C)	13	0.4
	30-34 (D)	56	1.8
	35-39 (E)	178	5.7
	40-44 (F)	483	15.6
	45-49 (G)	1145	36.9
	50-54 (H)	520	16.8
	55-59 (I)	328	10.6
	60-64 (J)	213	6.9
	65+ (K)	160	5.2

Residence Condition	Rural	134	4.3
	Urban	2970	95.7
Residence	In Gondar	2502	80.6
	Outside Gondar	602	19.4
Educational status	No	836	26.9
	Primary	934	30.1
	Secondary	947	30.5
	Tertiary	387	12.5
Employment	Employed	596	19.2
	Farmer	134	4.3
	Not Employed	2011	64.8
	Retired	13	0.5
	Self Employed	350	11.2
Marital status	Divorced	693	22.3
	Married	1521	49.0
	Separated	87	2.8
	Single	413	13.3
	Widow/widower	390	12.6
Total		3104	100

For more than three-fourth (79.1%), ART was initiated while the patients were classified as being on a "working" functional status. Majority of them had their baseline renal and liver function tests results within normal range. Overall, 1890(61%) patients started ART at WHO clinical stage III. From the total 1889 female patients, 35(1.9%) were pregnant at enrollment (Table 2).

Table 2: ART-related characteristics of patients

Variable	Possible values	Count	%
Patient Functional status	Ambulatory	521	16.8
	Bedridden	127	4.1
	Working	2456	79.1
WHO Clinical Stage	Stage I	311	10
	Stage II	500	16.1
	Stage III	1890	60.9
	Stage IV	403	13
RFT	Elevated	91	2.9
	Normal	3013	97.1
LFT	Elevated	900	29

	N	2204	
		71	
Patient Pregnancy status during enrolment	No pregnancy or male patient	3069	98.9
	Pregnant	35	1.1
Baseline ART regimen	AZT-3TC-EFV	371	11.9
	AZT-3TC-EFVKid	1	0.03
	AZT-3TC-NVP	1151	37.1
	AZT-3TC-NVPKid	2	0.06
	OTHER	547	17.6
	d4T(30)-3TC-EFV	317	10.21
	d4T(30)-3TC-NVP	669	21.65
	d4T(40)-3TC-EFV	11	0.35
	d4T(40)-3TC-NVP	35	1.1
Current ART regimen	AZT-3TC-EFV	464	14.9
	AZT-3TC-EFVKid	1	0.03
	AZT-3TC-NVP	1143	36.85
	DROP	4	0.12
	LOST	23	0.74
	OTHER	974	31.4
	RESTART	5	0.16
	d4T(30)-3TC-EFV	155	5
	d4T(30)-3TC-NVP	335	10.8
Regimen change during course of treatment			
	Changed	1040	33.5
	Not Changed	2064	66.5
ART adherence	Good	3068	98.8
	Poor	36	1.2
Total		3104	100

*Other: Second line regimens such as ABC or TDF + 3TC + EFV, ABC or TDF + 3TC+ LPV/rc, AZT + 3TC + LPV/rc.

CD4 count changes after initiation of ART

The median baseline CD4 count for all patients was 139 cells/micro liter and the median of the most recent CD4 count was 358. CD4 count changes after initiation of ART ranged between -544 and 1675, with a mean change of 243 (SD 191.14), and a median of 208 CD4 cells over ma median of 191.49 weeks follow up.

Of the total clients on ART, 2427 (78.2%) had a CD4 count change of greater than 100 cells/micro liter. However, for 137(4.4%) patients CD4 count were less than their baseline values (Table 3).

Table 3: CD4 count changes in patients on ART at the University of Gondar Hospital

CD4 count change	Frequency	Percent
ΔCD4 > 100	2427	78.2
ΔCD4 51 -100	336	10.8
ΔCD4 1-50	204	6.6
ΔCD4 < 0	137	4.4
Total	3104	100

Attribute selection

Applying the Ranker + InfoGainAttributeEval method of attribute selection method, variables were ranked according to their importance in classifying the CD4 count change. Information Gain values were used to rank variables (Table 4).

Ranked attributes:

Table 4: Attribute rank and its information Gain score

Rank	Attribute	Information gain score
1	Patient Age	0.355846
2	Education	0.281198
3	Baseline CD4/CD8 ratio	0.259066
4	Baseline CD8 cell count	0.249143
5	Baseline ART Regimen	0.239833
6	Current ART Regimen	0.212706
7	Baseline hemoglobin level	0.197933
8	Marital Status	0.164389
9	Patient Sex	0.155603
10	Baseline CD4 count	0.151298
11	Baseline Platelet count	0.144812
12	Number of weeks on ART	0.119318
13	Regimen change	0.084873

14	Liver Function Test(LFT)	0.081058
15	Employment status	0.068596
16	WHO clinical stage	0.054502
17	Baseline Functional Status	0.026612
18	Baseline weight	0.023363
19	Residence	0.021713
20	Religion	0.005935
21	ART adherence	0.004278
22	Residence condition	0.004228
23	Renal Function Test(RFT)	0.002742
24	Patient Cotrimoxazole adherence	0.001459
25	Tuberculosis Screening status	0.001353

We also assessed other methods such as Best First +CFS Subset Eva and Random Search +CFS Subset Evaluator to predict CD4 count changes. The two methods agree that baseline hemoglobin level, CD4/CD8 ratio, liver function test, marital status, educational status, sex, age, age, WHO clinical stage, type Baseline ART Regimen, type of Current ART Regimen, Cotrimoxazole adherence and number of weeks on ART as the main subset of attributes important in predicting the CD4 count

Model Building

Different experiments were carried out to build the model. These experiments are based on two different attribute and instance combinations (input). These inputs were either:

A. Input 1: the experiment is conducted using the 27 variables (including the outcome variable). A total 3104 records (2427, 336, 204, 137 values for the CD4 count change classes of A, B, C and D respectively)
B. Input 2: using 27 variables (including the outcome variable) and SMOTE (Synthetic Minority Over Sampling Technique) balanced class.

In principle, imbalanced data causes machine learning algorithms to achieve reduced performance accuracy of classification algorithms due to the minority class or the rarely occurred classes (as happened in D or 137 in the above case). Values will be misclassified to the majority class occurrences, because the classifier algorithm will be overwhelmed with the majority class

and ignores the minority class [27]. After experimenting 23 levels of different SMOTE algorithms, the 19th level of SMOTE made the class approximately balanced, making a total number of instances of 42,284 (Summarized in table 5).

Table 5: CD4 count change class with respective to the SMOTE balanced class values

class	Class values	Percentage
A	9,708	23
B	10,752	25.4
C	13,056	30.9
D	8,768	20.7
Total	42,284	100

For each algorithm, these two inputs in 10% cross validations and also default 90/10 training/ testing test options are conducted resulting in four experiments per an algorithm. In total, 10 experiments were conducted for three different predictive models.

Test Option 1: 10% cross validation

Test Option 2: default 90/10 training/ testing split

Input 1 and test option1: J48 algorithm with input 1and 10% cross validation applied to imbalanced data (original data)

Input 1 and Test option 2: J48 algorithm with 90/10 training test applied to imbalanced data (original data)

Input 2 test option 1: J48 algorithm with input 2 and 10% cross validation, using SMOTE balanced data

Input 2 test option 2: J48 algorithm with input 2 and 90/10 training/testing split, using SMOTE balanced data

Similarly, these experiment options were conducted for Neural Network and Random Forest algorithms. The outputs from the three algorithms are summarized in table 6.

Table 6: Accuracy of J48, Neural Network and Random Forest algorithms

Experiment	Test Name	No. of attributes used	Instances used	Size of the tree	No. of leaves	Time taken to build the model(in seconds)	Accuracy achieved

Model	J48						
1	Input 1 and test option1	27	3104	304	208	0.64	79.99
2	Input 1 and Test option 2	27	3104	304	208	0.3	87.82
3	Input 2 test option 1	27	42284	1989	1308	12.77	96.60
4	Input 2 test option 2	27	42284	1989	1308	12.72	98.69
Model	**Neural Network**						
1	Input 1 and test option1	27	3,104			352.05	75.74
2	Input 1 and Test option 2	27	3,104			350.66	94.20
3	Input 2 test option 1	27	42,284			4693.67	95.47
4	Input 2 test option 2	27	42,284			5112.05	96.62
Model	**Random Forest**						
1	Input 1 and test option1	27	3,104			1.33	79.99
2	Input 2 test option 2	27	42,284			11.81	99.98

Interesting rules extracted from the decision tree

J48 output can be better described with a tree. However, the tree developed from this algorithm is large. There are interesting rules generated from the tree such as the following

1. If patient age is between 35 and 39, liver function test normal, functional status working and patient baseline regimen is AZT-3TC-EFV→the CD4 count change expected will be greater than 100.
2. If patient age is between 35 and 39, liver function test normal, functional status working, patient original regimen AZT-3TC-NVP, female, WHO stage II, unable to read and write, married and CD8 cell count is greater than or equal to 489→the CD4 count change will be between 51 and 100.
3. If patient age is between 40 and 44, male and has working functional status → the CD4 count change will be greater than 100
4. If patient age is between 40 and 44, female, unemployed, WHO clinical stage II, baseline regimen is AZT-3TC-NVP, has elevated level of liver function test results, CD4/CD8 ratio 0.096→The CD4 count change will be less than zero.
5. Patients aged 45-49 years, unable to read and write, had normal liver function tests, undergone ART regimen changes, initiated with d4T(30)-3TC-NVP regimen, had a baseline CD8 cell count of >= 1123 cells/microliter, hemoglobin value of <= 13 mg/dl, and a follow up duration of <= 201 weeks, had negative CD4 count changes.

Among the four serial experiments conducted for J48, J48 algorithm with 90/10 training/testing percentage split, the use of SMOTE balanced data showed superior accuracy 98.69% over the other J48 experiments. It correctly classified the 98.69% (41,731) of the instances from the total number of SMOTE balanced instances of 21,142. However, the time taken to build the model was higher than the other three experiments.

Similarly, the best classification accuracy in Neural Network algorithm was achieved by applying Neural Network in SMOTE balanced data with tuning the test option to default 90/10 training/testing test mode. This test classified the SMOTE balanced data correctly with accuracy of 96.62% that means 40,857 instances were correctly classified.

These experiments were repeated also by using Random Forest algorithm as displayed in table 6. Note that here, there is no need to cross validate Random Forest to get unbiased estimate of the test set error because it is estimated internally, while leaving out one third of the data (37%) in the boot strap. Therefore, there was no experiment conducted using Random Forest test with cross validation.

Comparing the three algorithms to find out which algorithm works best in classifying the data is important. Three of the algorithms achieve very good accuracies. In all cases, the SMOTE balanced data is more correctly classified than the imbalanced data. By comparing the J48, Neural Network and Random Forest; Random Forest SMOTE balanced data in 90/10 training/testing algorithm outperforms J48 and Neural Network and achieved a classification accuracy of 99.99%. Random Forest outperforms Neural Network and the J48 decision tree by achieving 3.6% and 0.94% higher classification accuracy Neural Network respectively.

Evaluation of the developed models

Therefore, the three better classifying models were reevaluated. J48 algorithm and Neural Network with 90/10 training/testing split using SMOTE balanced data and Random Forest are taken to train the data again by assigning the test options in default 90/10 testing/training split. Therefore the target data set is supplied for the models and analyzed with this test option and achieves a classification accuracy of 87.11 and 83.54 and 99.84% for J48, Neural Network and Random Forest respectively. The actual and predicted instances of the three algorithms are displayed in table 7.

Table 7: Re-evaluation of most accurate J48, Neural Network and Random Forest models by supplying the algorithms with the original data with the 3104 instances

J48		Classified as					
Actual		A	B	C	D	Total	Percentage accuracy
	A	2318	44	34	31	2427	95.5
	B	119	198	12	7	336	58.9
	C	63	20	117	4	204	57.4
	D	46	9	11	71	137	51.8%
Total		2546	271	174	113	3104	87.11
Neural Network							
	A	2310	42	45	30	2427	95.2
	B	158	145	18	15	336	43.2
	C	101	19	72	12	204	35.3
	D	54	9	8	66	137	48.2
Total		2623	215	143	123	3104	83.54
Random Forest							
	A	2424	0	1	2	2427	99.88
	B	1	335	0	0	336	99.70
	C	0	0	204	0	204	100
	D	1	0	0	136	137	99.27
Total		2426	271	174	113	3104	99.84

The Random Forest of 10 trees was constructed using five random features. The out of bag error is 0.0759 and the time taken to build the model is 12.36 seconds. Note that the out of bag error comes from the out of bag data. When the training set used to build the tree by using sampling with replacement, one third of the cases are left out of sample(out of bag data). This data was then used to get a running unbiased estimate of the prediction error while trees are added to the forest.

By comparing the classification performance accuracies of the three algorithms, Random Forest model applied to SMOTE balanced data with 90/10 split, works well when supplied and re-evaluated with the original 3104 instances data. It produced more than 99% classification accuracy.

The Random Forest algorithm with accuracy of 99.84% scores an AUC value of value of 1 which means it perfectly classifies the data. J48 with accuracy of 87.11% showed an AUC value of 0.938. Neural Network with the 83.54 accuracy has an AUC value of 0.812. So that the AUC also revealed that Random Forest works best for classifying the CD4 count changes. The percent agreement or the kappa statistic is also excellent (99.57%) as it achieves almost perfect percentage agreement between the CD4 count change predicted by the model and the observed CD4 count change. The detailed classification accuracy is summarized in table 8

Table 8: Classification accuracy of the best working Random Forest model

Class	True positive rate	False positive rate	Precision	Recall	ROC Area
A	0.999	0.003	0.999	0.999	1
B	0.997	0	1	0.997	1
C	1	0	0.995	1	1
D	0.993	0.001	0.986	0.993	1
Weighted average	0.998	0.002	0.998	0.998	1

The AUC for Random Forest classification for each class outcome was 1.0(Table 7). This implies that Random Forest achieves perfect performance in classifying the CD4 count changes accordingly. Note that a random guess produces a ROC area value of 0.5.

Discussion

This study aimed at predicting CD4 count changes using data mining techniques. The results demonstrate that nearly accurate prediction levels were achieved using Artificial Neural Network and J48 decision tree algorithms. In comparison, Random Forest algorithm outperforms both Neural Network and J48 decision tree algorithms with a classification accuracy close to perfection, 99.84%. However, classification algorithms with high accuracy do not necessarily

equate to better performance on target datasets. Other evaluating mechanisms such as Area under the ROC (Receiver Operability Characteristics) curve are considered to be more robust measures. For highly imbalanced data, the Area under ROC (AUC) is a better performance metric. The AUC is the probability that a classifier ranks a randomly chosen positive instance higher than a negative instance [28]. Graphically, ROC curve is true positive rate (Sensitivity) plotted in a function of the false positive rate (100-Specificity) [29,30]. Hence, we evaluated the area under the ROC curve of the Random Forest algorithm which showed a value close to 1, reflecting a perfect performance. Using other evaluation metrics such as precision, true positivity rate, false positivity rate and kappa also demonstrated the supreme performance of the Random Forest over J48 and the Artificial Neural Network prediction algorithms. Comparable level of AUC was reported from a recent study by Revell and colleagues [31]. Hence, Random Forest can be used to develop cost effective web-based prediction models to help providers forecast patients' future likelihood of immunologic reconstitution without conducting expensive CD4 counts. In developing countries, the main challenges in CD4 counting are lack of machine installations, frequent machine breakdowns, lack of timely and proper maintenance, and lack of reagents [19]. Two recent reviews identified the main classification and regression models that were applied in predicting HIV drug resistance and response to therapy [32,33]. Considering the availability of HIV/AIDS-related big datasets, the use of data driven predictive models are becoming widely explored. Hence, the use of these models could be an alternative approach to overcome CD4 count challenges in resource-limited settings.

This study identified the most important attributes for classifying CD4 count changes. These variables are age, educational status of the patient, baseline hemoglobin level, baseline CD4/CD8 ratio, Liver Function Test(LFT), marital status, sex, WHO clinical stage, baseline ART regimen type, current ART regimen, patient`s level of adherence to Cotrimoxazole, and ART treatment duration. Variables such as patient's tuberculosis status during last visit and pregnancy status during enrollment have had minimal information gain values. The information gain measures the expected reduction in entropy due to splitting on a certain attribute. As an example, the total entropy is reduced by 0.281, when the classification on decision tree is split by educational status. That means, having the information about educational status of a patient reduces the entropy by 0.281 and goes to the next splitting attribute in classifying the CD4 count changes.

Hence, this information gain values show how much the splitting attribute contributed in the classification of CD4 count changes. A study that used decision tree and Random forest algorithm revealed that total Lymphocyte Count, hemoglobin and total platelet counts have significant prognostic value for monitoring HIV/AIDS clinical progression [34].

Comparable predictive performance accuracies of data mining algorithms were reported by previous studies [21,22,28,35-38]. In 2011, a study from India applied Random Forest algorithm to predict anti-tubercular molecules using machine learning on high-throughput biological screening datasets and reported that Random Forest outperformed other classification algorithms such as J48 and SVM classification [28]. Random Forest algorithm was applied on HIV flow cytometry data to predict CD4 immune reconstitution outcome and identify key determinants using variable importance scores [21]. Wang and colleagues compared Random Forest and Support vector machine algorithms with a committee of artificial Neural Network on their performance on accurately predicting the virologic response of patients to antiretroviral treatment. Comparably accurate performance was shown by Random Forest and Support vector Machine [38]. Random Forest has identified that CD4 cell count cut of value of 400 cells/microliter was power classifier in differentiating immune-discordant and immune-concordant participants [35].

Compared to Neural Network and J48 decision tree algorithms, Random Forest predicted CD4 count changes among patients on ART with superior level of performance accuracy. This result is consistent with the findings from the UK and Romania which showed Random Forest algorithm has achieved an overall accuracy of more than 77% and 79% respectively [22,36]. In addition, Random Forest models were trained to predict response to antiretroviral treatment in other low-resource settings such as South Africa, Romania and India. The models identified available alternative antiretroviral treatment combinations and predicted the virologic failure with a prediction accuracy of 94, 99, and 93 in South Africa, India and Romania respectively [37]. A study from the UK[REF] reported that Random Forest and SVM models can produce predictions of virological response to HIV treatment comparable in accuracy to a committee of artificial Neural Network models. Consistent with our findings, Artificial Neural Network models were significantly inferior to Random Forest and SVM [18]. Another study from UK also

revealed a committee of 10 Random Forest algorithms used to predict virologic response to therapy showed an overall accuracy of 63% [31].

Unpublished study at Felege Hiwot Referral Hospital to predict adherence status of ART clients reported that comparable performance accuracy (93.14%) was demonstrated using J48 Decision tree algorithm [39]. But this study applies an extra algorithm; Random Forest, which performed excellent in predicting the CD4 count changes better than decision tree and Neural Network.

This study has notable strengths. Firstly, the extraction of data from both the electronic database and patient clinical record. Secondly, three powerful algorithms were employed for building the prediction models.

However, the study has a number of limitations. First, the data set has low number of instances and attributes. In principle, machine learning algorithms perform better if employed with large number of dataset and large number of variables. A review on the use of computational models for the prediction of response to Antiretroviral therapy revealed machine learning algorithms tend to infer knowledge by considering heterogeneous information. For instance coupling viral genotype with phenotypic traits such as laboratory markers (eg. CD4 count, CD4/CD8, HIV RNA), therapeutic history, epidemiological and demographic factors provide a more complex and powerful prediction model [32]. Therefore, one of the limitations of this study is it did not include important variables that are known to be potentially associated with CD4 count such as viral load, types of opportunistic infections and nutritional status of the patient. We recognize the most recent Ethiopian HIV/AIDS treatment policy is moving towards implementation of test and treat strategy as well as the expansion of access to viral load monitoring. Therefore, future studies need to consider including these variables if more helpful classification result is to be achieved. However, due to resource constraints, viral load of patients is unavailable in many contexts of HIV care in developing countries. Hence, HIV care of these countries remains to be dependent on periodic CD4 counting and WHO clinical staging. Future studies also need to test more algorithms, preferably algorithms which are proved to learn medical data well might be beneficial to be tested. Secondly, this study only predicts CD4 count change in a range. Algorithms that are able to predict the absolute count such as fuzzy linear regression should be tested. In addition, lack of similar studies in the study area limit the effort to compare the results of our study with others.

Conclusion

For predicting CD4 count changes, a nearly accurate prediction model was developed using Random Forest algorithm. Baseline patient characteristics such as CD4 count, CD8 cell count, type of ART regimen, hemoglobin level, and baseline platelet count predict CD4 count changes. Likewise current ART regimen type, duration of ART, being on same ART regimen throughout, liver function test, and WHO clinical stage, as well as demographic factors such as sex, age and educational status were identified as predictors of CD4 count change. The results of this study can be used to develop cost effective web-based prediction models which help providers to picturize patients' future health status and to make recommendations for better prognosis, management and resource allocation.

Abbreviations

ABC	Abacavir
ddI	Didanosine
d4T	Stavudin
EFV	Efavirenz
FTC	Emotricitabine
LPV/r	Lopinavir/ritonavir
NFV	Nelfinavir
NNRTI	Non-nucleoside reverse transcriptase inhibitor
NRTI	Nucleoside Analogue Reverse Transcriptase Inhibitor
NVP	Nevirapine
SQV/r	Saquinavir/ritonavir
TDF	Tenofvir
ZDV	Zidovudine
3TC	Lamivudine
WHO	World health Organization

Declarations

Statement of ethical approval

Ethical approval was obtained from the Institute of Public Health of the College of Medicine and Health Science, University of Gondar. Support letter was obtained from University of Gondar Hospital. De-identified information were received from the ART database.

Consent for publication

Not applicable

Availability of data and material

The data used for this study can be obtained from the corresponding author with justifiable request.

Funding

No funding received for this study

Competing interests

The authors declare no competing interests.

Authors' contributions

MK initiated the idea, oversaw the data collection, analyze the data and wrote the manuscript

BMZ and DZ involved in conception, design, data interpretation and review of subsequent drafts

All authors read and approved the final manuscript

Acknowledgements

The authors would like to express their gratitude to ART clinic database administrators of the University of Gondar Hospital.

References

1. Trevor P, Emily W, Richard F, et al. Challenges in Implementing CD4 Testing in Resource-Limited Settings: Cytometry Part B (Clinical Cytometry). Wiley InterScience Clinical Cytometry Society 2008;**74B**((Suppl. 1)):S123–S30.
2. Organization WH. Guideline on when to start antiretroviral therapy and on pre-exposure prophylaxis for HIV. Switzerland WHO, 2015.
3. WHO. Antiretroviral therapy for HIV infection in adults and adolescents: recommendations for a public health approach. Geneva, Switzerland, 2006, revision.
4. Ford N, Meintjes G, Pozniak A, et al. The future role of CD4 cell count for monitoring antiretroviral therapy. Lancet Infect Dis 2015;**15**(2):241-7.
5. Organization WH. Consolidated guidelines on the use of antiretroviral drugs for treating and preventing HIV infection Switzerland: WHO, June 2013.

6. Kilmarx PH, Simbi R. Progress and Challenges in Scaling Up Laboratory Monitoring of HIV Treatment. PLoS Med 2016;**13**(8):e1002089.
7. Jaffar S, Grosskurth H, Amuron B, et al. Use of WHO clinical stage for assessing patient eligibility to antiretroviral therapy in a routine health service setting in Jinja. Uganda AIDS Research and Therapy 2008;**5**(4).
8. Johnstone S, Hargrove J, Williams B. Antiretroviral therapy initiated soon after HIV diagnosis as standard care: potential to save lives? . HIV/AIDS - Research and Palliative Care 2011;**3**:9-17.
9. Jiawei H, Micheline K. *Data Mining: Concepts and Techniques*. 2n dedition ed. United States of America: Elsevier, 2006.
10. Mdzingwa N. Data Mining with Oracle 10g using Clustering and Classification Algorithms COMPUTER SCIENCE HONOURS PROJECT 2005.
11. Witten I, Eibe F. *Data Mining Practical Machine Learning Tools and Techniques*. second ed. AMSTERDAM, BOSTON, HEIDELBERG, LONDON, NEW YORK, OXFORD, PARIS , SAN DIEGO, SAN FRANCISCO , SINGAPORE , SYDNEY , TOKYO: Elsevier, 2005.
12. Ullah I. Data Mining Algorithms And Medical Sciences. International Journal of Computer Science & Information Technology (IJCSIT) December 2010;**Vol 2**(No 6).
13. LILLY M, BALASUBRAMANIE P. multi layer feed backward neural network model for medical decision support: implementation of back propagation algorithm in hiv/aids regimens. International Journal of Reviews in Computing 2009(E-ISSN: 2076-3336).
14. Leke B, Marwala T, Tettey T. Using Inverse Neural Networks for HIV Adaptive Control. International Journal of Computational Intelligence Research 2007;**3**(1):11-15.
15. The Prediction of AIDS Survival: A Data Mining Approach. 2nd WSEAS International Conference on Multivariate Analysis and its Application in Science and Engineering; 2010; Malaysia.
16. Kareem S, Awadh N, Kamaruzaman A, et al. Classification and regression tree in prediction of survival of aids patients. Journal of Computer Science 2010 **23** (3):153-65.
17. Madigan E, Curet O, Zrinyi M. Workforce analysis using data mining and linear regression to understand HIV/AIDS prevalence patterns. Human Resources for Health 2008;**6**(2).
18. Wang D, Larder B, Revell A, et al. A comparison of three computational modelling methods for the prediction of virological response to combination HIV therapy. US National Library of Medicine National Institutes of Health 2009;**Vol. 47(1)**.
19. Habiyambere V, Ford N, Low-Beer D, et al. Availability and Use of HIV Monitoring and Early Infant Diagnosis Technologies in WHO Member States in 2011-2013: Analysis of Annual Surveys at the Facility Level. PLoS Med 2016;**13**(8):e1002088.
20. Tesfaye B, Atique S, Elias N, et al. Determinants and development of a web-based child mortality prediction model in resource-limited settings: A data mining approach. Computer Methods and Programs in Biomedicine 2017;**140**:45-51.
21. Eliot M, Azzoni L, Firnhaber C, et al. Tree-Based Methods for Discovery of Association between Flow Cytometry Data and Clinical Endpoints. Adv Bioinformatics 2009;**2009**:235320.
22. Revell AD, Ene L, Duiculescu D, et al. The use of computational models to predict response to HIV therapy for clinical cases in Romania. Germs 2012;**2**(1):6-11.
23. Chapman P, Clinton J, Kerber R, et al. CRISP-DM 1.0: Step-by-step data mining guide: NCR Systems Engineering Copenhagen (USA and Denmark), DaimlerChrysler AG (Germany), SPSS Inc. (USA) and OHRA Verzekeringen en Bank Groep B.V (The Netherlands), August 2000.
24. Federal HIV/AIDS Prevention and Control Office Ministry of Health E. *National Comprehensive HIV Care/Antiretroviral Therapy (ART) Training Course: Participant Manual for Training of Trainers on Basic Chronic HIV Care, Antiretroviral Therapy and Prevention*, April, 2008.
25. Wenjia DW. Tutorial for Weka: a data mining tool. CMP: Data Mining and Statistics within the Health Services 2010.

26. WEKA 3 6 4 software [program]. New Zealand: University of Waikato, 2010.
27. Learning Classifiers from Imbalanced, Only Positive and Unlabeled Data Sets. 14th ACM SIGKDD International Conference on Knowledge discovery and data mining 2008.
28. Vinita P, Jinuraj R, Abdul J, et al. Predictive models for anti-tubercular molecules using machine learning on high-throughput biological screening datasets. Biomed Central 2011;**4**(504).
29. Nuts and Bolts of Data Mining: Classifiers & ROC Curves. Salford Analytics and Data Mining Conference 2012, insight for data enthusiasts, May 24-25; 2012 February 2012; San Diego.
30. Tom F. ROC Graphs: Notes and Practical Considerations for Data Mining Researchers. Intelligent Enterprise Technologies Laboratory HP Laboratories Palo Alto 2003:1-27.
31. Revell A, Khabo P, Ledwaba L, et al. Computational models as predictors of HIV treatment outcomes for the Phidisa cohort in South Africa. Southern African Journal of HIV Medicine 2016;**17**:1-7.
32. Prosperi MC, De Luca A. Computational models for prediction of response to antiretroviral therapies. AIDS Rev 2012;**14**(2):145-53.
33. Isis B. Machine Learning for Prediction of HIV Drug Resistance: A Review. Current Bioinformatics 2015;**10**(5):579-85.
34. Rafatpanah H, Essmailian L, Hedayati-Moghaddam MR, et al. Evaluation of Non-Viral Surrogate Markers as Predictive Indicators for Monitoring Progression of Human Immunodeficiency Virus Infection: An Eight-Year Analysis in a Regional Center. Jpn J Infect Dis 2016;**69**(1):39-44.
35. Perez-Santiago J, Ouchi D, Urrea V, et al. Antiretroviral therapy suppressed participants with low CD4+ T-cell counts segregate according to opposite immunological phenotypes. AIDS 2016;**30**(15):2275-87.
36. Revell AD, Wang D, Boyd MA, et al. The development of an expert system to predict virological response to HIV therapy as part of an online treatment support tool. AIDS 2011;**25**(15):1855-63.
37. Revell AD, Wang D, Wood R, et al. Computational models can predict response to HIV therapy without a genotype and may reduce treatment failure in different resource-limited settings. J Antimicrob Chemother 2013;**68**(6):1406-14.
38. Wang D, Larder B, Revell A, et al. A comparison of three computational modelling methods for the prediction of virological response to combination HIV therapy. Artif Intell Med 2009;**47**(1):63-74.
39. Asresu T., Zegeye DT., Meshesha M. . Knowledge discovery for Antiretroviral Adherence Prediction. University of Gondar 2011.