JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125 VOL 10, ISSUE 04, 2023

Predictive Analytics for Retention in Care and Antiretroviral Therapy Adherence using Supervised Learning – A Case Study of County Health Facilities in Kenya

Geoffrey Sagwe Ombui
Department of Computer Science,
University of Nairobi, Kenya
sagweombui@students.uonbi.ac.ke

Andrew M. Kahonge
Department of Computer Science
University of Nairobi, Kenya
Andrew.mwaura@uonbi.ac.ke

Teresa K. Mwendwa
Department of Human Anatomy and Medical Physiology,
University of Nairobi, Kenya
tmwendwa@uonbi.ac.ke

ABSTRACT

In healthcare organizations, a great problem is faced by healthcare providers to know the ART adherence and status of HIV/AIDS patients. In this research, a predictive model using supervised learning is developed to let clinicians and healthcare providers know the ART adherence of PLHIV using features of the patients' treatment history. The methodology used was CRISP-DM data mining process. The research used easily measurable baseline demographic and clinical variables such as body weight, ART regimen, patients enrolled in care, and phenotype. Data preprocessing and transformation was done to ensure the dataset collected was clean. The dataset was split into training and test set i.e., 80% for training and 20% for testing.
The baseline results from the benchmark and performance evaluation showed that random forest model performed the best with accuracy of 81% and AUC of 79.3% compared to other binary algorithms and classification error rate of 0.333. The model that performed poorly was Naïve Bayes with an accuracy score of 20.0%. The model that performed poorly was Naïve Bayes with an accuracy score of 20.0%. The researcher retrospectively followed 21551 records of patients who were seeking care at county health facilities.

General Terms

Knowledge Discovery in Databases (KDD), Generalized Estimating Equations (GEE), Case Adherence Index (CAI), Treatment Change Episode (TCE).

Keywords

Area Under Curve (AUC), Cross-Industry Process for Data Mining(CRISP-DM), Supervised Learning, Viral Suppression, Antiretroviral Therapy (ART), Retention in Care, Demographic.

1. INTRODUCTION

The global scale-up of life-saving antiretroviral treatment (ART) is a critical turning point in the clinical management of HIV and reduction in AIDS-related deaths. Ability of countries to provide and sustain effective long-term HIV care with ART and prevention is critical which requires effective patient monitoring system integrated with care, prevention and treatment at the health facility. According to [1] observational analyses and mathematical models suggests that an intensive global investment to expand ART coverage could alter the epidemic trajectory and improve longevity, health, and economic productivity. Realizing this potential requires diagnosing HIV, initiating ART, and suppressing viral replication in most HIV-positive persons, a progression referred to as the HIV care cascade. In Kenya, HIV programme provides comprehensive treatment and care interventions for improved health outcomes among people living with HIV.
Data shows significant improvement in viral suppression rates and reduction in mortality rates over the years.
Estimate models demonstrate the scale up of ART treatment-initiated scale since 2004, has averted over 733,600 AIDS related deaths by the end of 2019. With adoption of the universal test and start policy, retention on ART at 12 months has declined from 92.4% in 2013 to 83% in 2019 [2]. Despite this progress, there remains a substantial deficit in overall life expectancy among PLHIV, with their survival between 5 and 10 years less than among uninfected people. These excess deaths among PLHIV are occurring due to care and treatment and ART interruptions [3]. The low self-efficacy is complicated by the fact that some patients are hesitant to tell their providers that they are not taking ART, for example in cases when providers have prescribed ART and assume patients have initiated it [4]. Poor linkage to care and treatment where clients who are referred, do not get to the facilities is a major bottleneck to HIV care and treatment in Kenya [5]. A 2018 Kenya Population based HIV impact Assessment (KenPHIA) report describe 79.5% of PLHIV were aware of their HIV-positive status prior to survey. This therefore means that almost one in four PLHIV in Kenya were not on ART. We conducted a prospective cohort study where we trace all participants who started and switched the ART in a clinical setting. Our objectives is to quantify the proportion of individuals retained in clinic versus retained in care at 12 months and to determine risk factors for attrition from HIV care at 12 months using machine learning models.

2. OBJECTIVES

To develop a machine learning model to predict retention in care and ART adherence for community-based testing treatment program in County HIV clinics.

3. PREVIOUS EFFORTS IN ANTIRETROVIRAL THERAPY ADHERENCE AND VIRAL SUPPRESSION

In a clinical setting, assessment of retention and viral suppression levels after HIV treatment and ART adherence status in daily routine program settings represent the backbone of data-driven public health efforts to bring the epidemic under control. According to [9], the study from Pakistan on the antiretroviral therapy was influenced by persons who were

infected with HIV. A nonlinear control algorithm approach was used to control the readiness of the HIV patients, their system was aligned with more backstepping controls to check and improve the efficiency of the drug to the infected persons on their T cells. A study from sub-Saharan African countries states that they follow up adherence to treat the HIV patients.

They distinguished adherence and nonadherence patients' facing more problems like failure of immune systems and fatigues. The results show around 72.9% of adherence patients are saved and reduced transmission [10]. Sub-Saharan Project 2030 results through simulation model suggest that antiretroviral drugs reduce the rate of HIV victims, especially in case of adults [11]. In the study conducted in Zambia nearly 1.9 million people are affected by HIV. The researchers proposed ART to help the HIV infected patients of long-term survival and they suggested the government to allocate the budget to sustain this technique in forthcoming days. A collaborative study from Malawi, South Africa, and Zambia which addressed the mortality rates increasing in these countries. A study [12] from Southeast Coast of Africa shows tremendous difficulties to treat the HIV victims to follow up on their health. The patient's data was monitored for mitigating these risks, with different levels of prevention measures taken to reduce those problems. Fidelity of victims faced numerous problems like increasing number of HIV patients and even pediatric cases were increased. To eradicate the above stated risk, they launched different control mechanisms such as SMS reminders and mobile health applications that are used to notify the HIV+ mothers to prevent themselves. With these mechanisms, the number of cases were reduced and many children were saved from this pandemic.

In a study conducted in the USA, HIV-1 infection dynamic model was designed and evaluated with their data to ease the improvement in PLHIV [13]. Bootstrapping was used in order to correlate the different parameters, among those parameters, confidence is considered more crucial to interpret their clinical proofs, though this efficacy of the immune system was much improved. The researchers proposed CART (Combination Antiretroviral Therapy) as a mixed ART drug, which was given to the HIV + patients through blood plasma technique.

A study done by [14], presents the viability usage of data mining with ART towards the HIV positive victims shows the greater performances with prediction rate of 80.5%, while designed ART predictive model achieved their results of 66% with different hospital data in Amhara Region, Ethiopia. Knowledge Discovery in Databases (KDD) model was used to maintain the HIV patients in details. Zambian Population-based HIV Impact Assessment (ZAMPHIA) suggests viral suppression in nearly 90% of people self-reporting ART use; patients lost to follow-up from treatment programs and who have not been on treatment for some time may not be captured in the denominator, thus potentially overestimating suppression [15]. Researchers examined retention and viral suppression in a large public health program across 4 provinces in Zambia, a country with an estimated 1,200,000 adults living with HIV. They used a sampling-based approach in which they selected facilities from 4 provinces (with probability proportional to facility size) and then intensively tracked a random sample of persons inversely proportional to facility size lost to follow-up in each of these selected sites. In a sample of 1 of the 4 provinces (Lusaka), researchers assessed data on plasma HIV RNA suppression levels among a sample of both in-care and lost-to-follow-up patients. The predictive approach yielded both a representative estimate of overall retention and viral suppression in a large region of Zambia and site-level estimates of retention with enough precision to assess site-to-site variation [16].

4. PROBLEM STATEMENT

Given the high patient volume and burden of HIV infection in this setting, the risk of patients facing barriers to linkage to care, retention in care, and medication adherence is high [6]. In ART Centers, manually ordering ART and ARV regimen by itself has a problem and is more sensitive to error. The disordering of the regimen without considering the aforementioned parameters could cause great side effects on patient side.

Healthcare providers, in an effort to achieve viral suppression, often fail to recognize the immediate needs that PLHIV face when enduring debilitating side effects.

This could contribute to disengagement with HIV services, particularly among PLHIV who had not experienced the transformational effects of ART [7]. Reaching a high level of viral suppression at the population level is a good proxy for both transmission risk and ART effectiveness which requires good outcomes at every step along the care continuum [8]. To monitor the quality of ART coverage and to prepare for implementing new ART guidelines, it is essential to assess the numbers of people needing ART. The aim of the study is to extend past research on persons living with HIV/AIDS on clinic setting and explore the feasibility of recruiting PLHIV starting their ART and switching first-line and second-line ART from HIV hospital clinics, and gain insights to refine recruitment procedures for future research, describe the clinic and peer-recruited cohorts and compare them with respect to background and health characteristics, including reasons for delaying/declining or discontinuing ART and explore whether those who have taken ART in the past differ from those who have never taken ART on background, health and other facilities.

5. METHODOLOGY

The research took a quantitative research design with emphasis in the objective measurements, statistical and numerical analysis of data. It articulated the required data, the methods used to gather it, and how the research questions will be answered by the data collected. Quantitative research was suitable during the collection and preparation of aggregated cohort dataset, which is in numeric form, to be used to develop the model. The methodology and the entire process of this study was guided by Cross Industry Standard Process for Data Mining (CRISP-DM). The research design served to outline the work done and described how the results was accomplished to identified research objectives.

CONCEPTUAL DESIGN

The conceptual design was guided by available Data Mining Technique and genetic algorithms, naive Bayes, nearest neighbor method, and decision trees. Dataset attributes were extracted, including the phenotype, ART regimen, patients enrolled in care, ART status, first-line and second-line ART to transform or consolidate data into forms suitable for mining strategies like attribute construction, aggregation, and normalization.

In this study, data preparation and preprocessing constructing a dataset from one or more data sources was used for exploration and modeling. The following diagram shows the conceptual model.

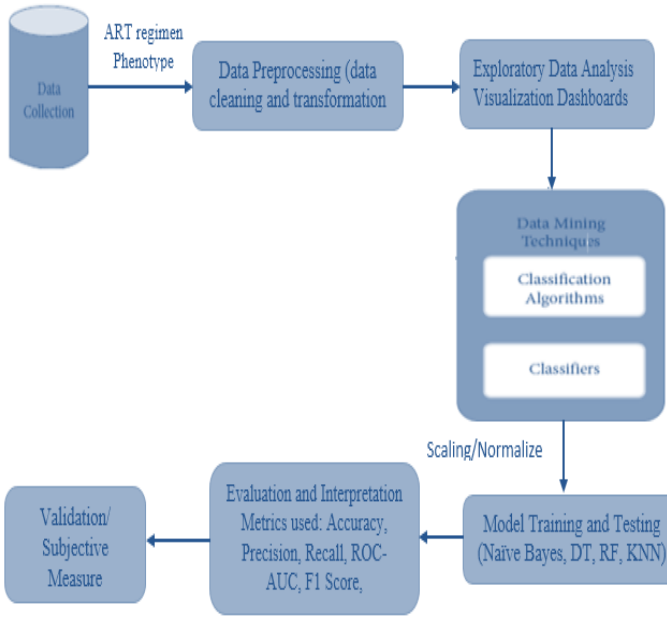

Figure 1: The Conceptual model

6. Data Collection and Study Population

The research retrospectively evaluated patients who were diagnosed with HIV and retained in care for ART at various County health facilities between Jan 2018 to Dec 2018, Jan 2019 to Dec 2019 and Jan 2020 to Dec 2020.

The significant data was collected from patient repository data, which was in ART Center, facility sampling unit MOH; that is, 21551 health facilities records were collected for conducting the experiment.

The required parameters like patients enrolled in care, patients currently on ART, those starting ART, First-line and second-line ART, patient weight, patients Max Z-Scores and outlier weight and patient ART regimen are collected from the repository. The structural cohort datasets was collected for subjective measures to evaluate the predicted model (extensively extracted new knowledge). Participants were followed from the time of a positive HIV test for up to 12 months. The researcher conducted a retrospective descriptive analysis of program data on treatment outcomes of PLHIV who are engaged in active patient ART. The data was cleaned and prepared for modeling. We did data analysis just to get familiar with the data and features of the dataset. More patients were hospitalized and retained for ART treatment in 2018 as compared to 2019 and 2020 respectively.

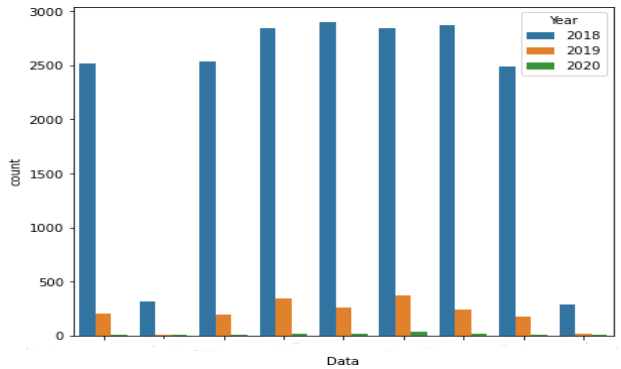

Figure 2: Study Sample of ART cohort data from 2018 to 2020

List of Top Counties with Highest ART Coverage. Checking frequency Distribution of values in workclass variable.

Counties	ART Coverage and Adherence
Nairobi County	1607
Homa Bay County	1380
Kakamega County	1367
Bungoma County	1216
Siaya County	1029
Kisii County	969
Kisumu County	969
Nakuru County	775

Table1: Counties with Highest ART Coverage

7. Data Preparation and Cleaning

The data preparation tasks included data cleaning, integration and transformation, we performed the below steps:

Data cleaning: The data collected was classified into HIV cohort 2018, 2019 and 2020 clinical variables for antiretroviral treatment. A python function was used to create the labels for the dataset including, correction and elimination of anomalies or outliers in the values of records and attributes. The dataset containing missing columns was dropped and filled with missing data (NaN) values with median for continuous variable and 0/1 for discrete variables. The remaining final data had 21551 entries which were sufficient for training. For Data selection; The categorical columns were converted to one-hot encoding, label encoding based on the categories available, Integration of data was reformatted by converting string values that stored numbers to numeric values to perform mathematical operations.

We cleaned the datasets collected, removing duplicates, white spaces and highlighting errors, filling in missing data, a process called imputation. The process improved the quality of the dataset and saved training time. We filled the missing data with the **median** values for the continuous variable and 0/1 for discrete variables and replaced missing values with "?" character with "NaN", this is because numerical python library ignores the nan values while performing mathematical imputations.

I. Data Transformation

We transform data into a format that machine learning algorithm can understand and model to discover information that will help select the features. We normalized our dataset using Scikit learn's preprocessing class called StandardScaler. The goal of **normalization** is to change the values of numeric columns in the dataset to a common scale, without distorting differences in the ranges of values. We encoded all the categorical data that was available and added dummy variables for the categorical data to make the data more readable and understandable. We normalized the data using the StandardScaler () class by Scikit learn library between a scale of 0 and 1. This is to ensure that outliers do not skew the model.

II. Feature Engineering

Feature design was guided by prior literature and domain expertise of retention and viral suppression. The dataset was collected and aggregated from different intervals of time. Factors previously shown to be associated with retention in HIV care such as ART, Gap Weight and 12 months survival and retention on ART. For each feature, measures were aggregated

by time (e.g., ART Net cohort at 12 months, 1ST line and 2ND line retention on ART) as well as different aggregation functions (mean, min, max, standard deviation) were calculated for each feature. Categorical variables such as data and county were dummified using label encoder and one-hot encoding. Missing data was imputed with imputation method on the variable missing e.g., Orgunit name. We investigated/evaluated cardinality and number of labels within the categorical variable. We dropped categorical columns with high cardinality since they pose serious problems in the machine learning model such as space consumption and curse of dimensionality.

III. Data Analysis

Exploratory analysis of time from HIV diagnosis to ART start, factors and effects on survival were be used to identify patients currently enrolled in care and starting on ART for targeted mapping outreach.

For qualitative and descriptive statistics, number and percentages were used to provide description for categorical variables while median was used for continuous variables like time. The study presented a classical exploratory data analysis approach to provide the needed intuition about data. Descriptive statistical techniques were used in interpreting and selecting qualitative features relevant to the research to have global view of the dataset and extract essential features. The research analyzed the frequency of features and correlation between the different key features of ART treatment. The longitudinal data for patients enrolled and retained in care for ART from January 2018 to December 2018 of persons living with HIV at County health facilities spread throughout the Country. Different times in the cascade of HIV care was examined including the duration from date patients were enrolled in HIV care, duration from enrollment to eligibility for ART and time from eligibility to initiation of ART.

The following predictors including information on demographics, clinical and antiretroviral treatment records including outcomes, clinical visits, laboratory outcomes and other HIV key indicators are used. For both outcomes, retention and suppression of ART therapy at 12 months, anonymized patient data who had accessed HIV care (observed in routine HIV electronic records) and, during the period 2018-2020, and had at least recorded/started their ART clinic visit.

8. RESULTS AND DISCUSSION

The problem of identifying HIV patients currently enrolled in care and starting on ART was cast as a binary classification problem using a variety of labels that are of interest to this study. A classification algorithms over a hyper-parameter grid such as (Naïve Bayes, Decision Trees, K-Nearest Neighbor and Random Forest Classifier) were used to develop the models.

The model's performance was evaluated in terms of accuracy (proportion/ratio of observations correctly classified by the algorithm among all observations in the unseen test set to the total number of input samples), sensitivity (the proportion of known positive outcomes in the unseen test set that are correctly identified as such by the algorithm), positive predictive value also known as precision (the proportion of positive outcomes predicted by the algorithm that correspond to known positive outcomes in the unseen test set) and specificity (the proportion of known negative outcomes in the unseen test set that are correctly identified as such) [17]. Research utilized the area under the curve (AUC) of a receiver operating characteristic (ROC) curve to evaluate the broad predictive classification performance of the model. A range of 0.5 indicated no predictive power while 1.0 indicates perfect predictive power. We used f1_score performance measure since it takes both recall and precision into consideration.

9. MODEL BUILDING

The analytic set was evaluated to establish the baseline prevalence of each outcome (% of original and alternative 1st line and 2nd line at 12 months survival and retention on ART). The analytic set was then randomly sampled on a 80/20 split into a training (80%) and test (20%) set. The training set for each outcome was then down-sampled to 50/50 positives (visits classified as enrolled in care and on ART Net cohort at 12 months survival) and negatives (currently on ART and starting on ART) to produce a balanced set of positives and negative examples for the classification algorithm to learn from. This step also addresses bias tendency towards predicting the majority class observed in many machine learning algorithms known as the class imbalanced problem.The classifier algorithm was then trained by input of predictor features as well as the specified target outcomes to produce an optimal configuration (predictive model) such that the predictor features correspond to the specified target outcomes as often as possible. After the model was trained, we separated the unseen test set into predictor features and outcomes. The unseen predictor features were given to the model which generated its predicted outcome for each observation (in this case, observations were each scheduled visit for the retention outcome and clinical, behavioral visit pattern outcome). The predictions were then scored for accuracy against the known outcomes in the unseen test set.

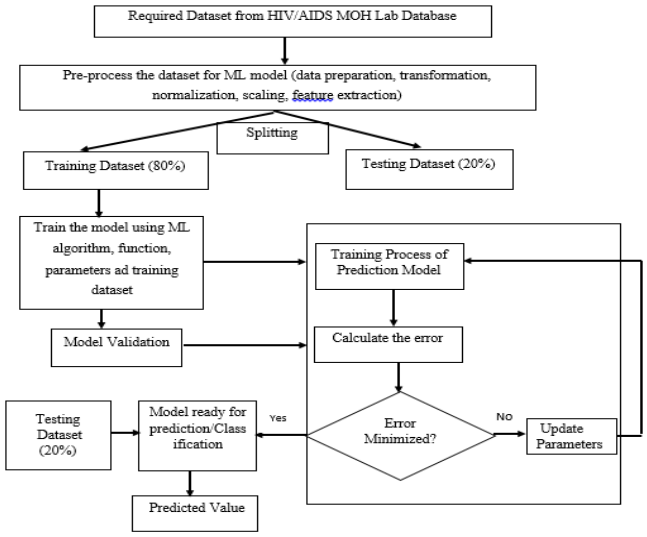

Figure 3: Flowchart for predictive Analytics model

10. CLASSIFICATION ACCURACY

The random forest classifier performed better than other binary classifier algorithms because it has bagging and ensemble approach consisting of a collection of different randomly composed decision trees whose results are aggregated for the cohort datasets. The ensemble methods also have the ability to produce both a prediction for an unseen sample, as well as a probability rating of its prediction. Their random nature often limits overfitting whilst controlling error making them attractive modelling tools for complex hyperspaces with non-linear separations in classes. The accuracy of the models for the study is given below.

Model	Accuracy Score (%)
Naïve Bayes (NB)	20%
Decision Trees (DT)	68%
Random Forest (RF)	81%
K Nearest Neighbor (KNN)	71%

Table 2: Classification accuracy score

F1-Score

It combines precision and recall relative to a specific positive class. The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1.0 and worst value at 0.0 as a weighted harmonic mean of precision and recall. So, f1-score is always lower than accuracy measures as they embed precision and recall into their computation. The weighted average of f1-score should be used to compare classifier models, not global accuracy. We used ensemble random forest and other binary classifier method in identification of the most predictive factors of retention on ART, ensemble model delivered better results than the other three algorithms with f1 score on the sample test of 79%.

Model	F1-Score (%)
Decision Trees (DT)	68%
Naïve Bayes (NB)	20%
Random Forest (RF)	79%
K Nearest Neighbor (KNN)	42%

Table 3: F1-Score for classification model performance

For one to be classified as retention on ART, the probability threshold set is 50 as the ratio of true positives (TP) to the sum of true positives and false negatives (TP + FN).

F1-Score combines both precision and recall metrics into one metric. If precision and recall are high f1-score will be high, and if they are low, the f1-score will be lower.

The classification algorithms was trained with an unbalanced 80:20 sample of the modelling set. This translated into 15085 patients' retention on ART. The model correctly classified 6466 of the test set yielding accuracy score of 55%.

Total cases of patients enrolled on ART were correctly identified, yielding a higher recall (or *sensitivity*) of 100% all positives. The model's precision of predicting a negative (or *specificity*) also remained high at 100%, further suggesting that total enrolled cases were higher.

Classification Report	Score (%)
Precision (Specificity)	100%
Recall (Sensitivity)	100%
True Positive Rate	95%
False Positive Rate	0%

Table 4: Classification report for precision and Recall

Cross Validation, Selection and Model Training

We tested the classification performance of machine learning models to cover a large spectrum of classifiers such as trees, ensemble trees, Euclidean distance, and Gaussian distribution. Hyperparameter combinations for each model were tested, then fit to each training set.

K-Fold cross-validation was performed to account for correlation and patterns in the data and correctly replicate the modeling workflow in deployment.

The data were divided into sets of model building cohorts and validation cohorts (training set and test set). This allowed models developed to account ART treatment outcomes occurring before the year of prediction and tested on treatment occurring during the year of prediction. Model performance was evaluated using accuracy and positive predictive value (PPV). The PPV is the percentage of patient cases correctly identified by the model starting and enrolled for ART.

Model	Training Accuracy (%)	Test Accuracy (%)	Cross-Validation (%)
K Nearest Neighbor	100%	43%	45%
Naïve Bayes	20%	18%	19%
Decision Tree	61%	56%	56%
Random Forest	77%	69%	69%

Table 5: Accuracy score of training/test set

The best performing model for ART treatment and access to care was a random forest with 400 estimators, maximum tree depth of 6, each leaf node having at least 2.5% of the samples, and each tree split requiring at least 10 samples. A simple decision tree had a lower performance with test and cross validation score of 56% with no specified maximum tree depth, each leaf node having at least 2.5% of the samples, and each tree split requiring at least 10 samples.

The ROC - AUC Curve Analysis for the Classifiers

The ROC Curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold levels. In the ROC Curve, we focused on the TPR (True Positive Rate) and FPR (False Positive Rate) of a single point. This gave us the general performance of the ROC curve consisting of the TPR and FPR. So, an ROC Curve plots TPR vs FPR at different classification threshold levels. If we lower the threshold levels, it may result in more items being classified as positive. It will increase both True Positives (TP) and False Positives (FP).

We utilized the area under the curve (AUC) of a receiver operating characteristic (ROC) to evaluate the broad predictive classification performance of the model. A range of 0.5 indicated no predictive power while 1.0 indicated perfect predictive power. The figure below shows AUC curves for the training and validation data of simple logistic regression, best performing random forest algorithm.

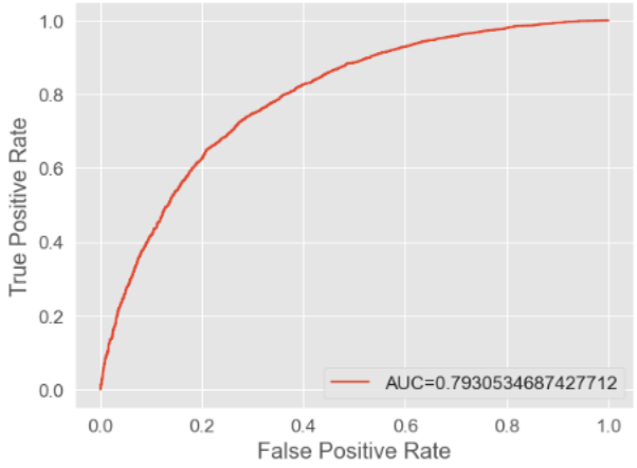

Figure 4: AUC graph for TPR vs FPR

AUC Curve represents a measure of separatability, the higher the AUC, the better the model is at predicting 0s and 1s. By analogy, higher the AUC, better the model is at distinguishing

between patients on ART adherence and those who are defaulting ART.
When AUC is approximately 0.5, the model has no discrimination capacity to distinguish between positive class and negative class.

The bigger the area covered, the better the machine learning models is at distinguishing the given classes. Ideal value for AUC is 1.

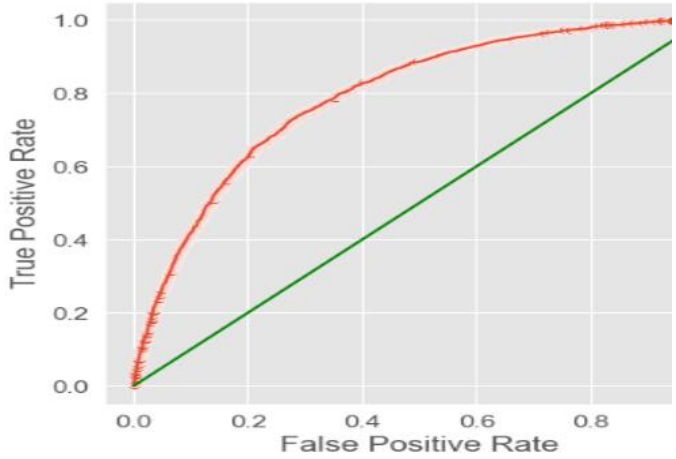

Figure 5: ROC curve to choose a threshold level.

The vertical axis (Y-axis) of ROC curve represents the true tested rate. The horizontal axis (X-axis) represents the false-tested rate. This indicates given that the attributes as input, the classifiers are better than the random model to predict individuals tested or not because all the three classifiers (five algorithms) have a ROC curve values above 50%. The ROC curve analyses for all experiments displayed showed that the curves moves sharply up from zero showing that there are more true tested than false tested rates. Then the curve starts to become more horizontal as it encounters less true tested and more false tested rates. The areas under the curve for the models are closer to 1.

Observations

To understand and quantify our machine-learning predictions, we compared our model to baseline class categories which corresponds to 47 Counties. Class0 represents patients on retention and ART care, class1 label represents persons virally suppressed. We ranked the observations by probability of whether a person is on ART care or virally suppressed.
We predicted the probabilities and chose the class with highest probability. The classification threshold level was 1.

```
Text(0, 0.5, 'Frequency')
```

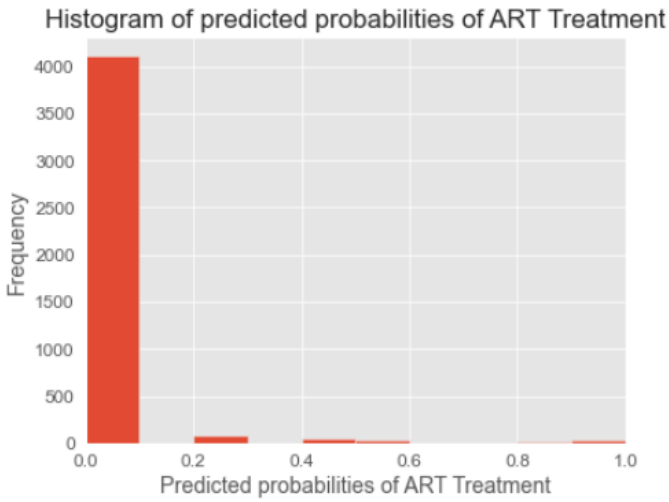

Figure 6: Histogram of Predicted probabilities

The above histogram is highly positive skewed. The first column tell us that there are approximately 4000 observations with probability between 0.0 and 0.1 where patients are likely to go for ART Treatment and switching first-line and second-line. There are relatively small number of observations with probability > 0.5. So, these small number of observations predicts persons virally suppressed. Majority of observations predict that patients are likely to start and enroll for ART Treatment.

11. Clinical Characteristics of Study Participants

During the three-year follow-up period, 3130 patients were hospitalized and had started ART, of which 2739 patients had taken first line antiretroviral therapy at 12 months on survival and retention. 329 switched to second-line antiretroviral therapy at 12 months due to failure of the first-line ART on survival and retention. Regarding medication adherence and total therapy at 12 months, patients had good adherence while on ART Net cohort at 12 months retention. After at least 12 months exposure of second-line therapy, 329 patients showed high survival and retention on ART with high viral load measurements and enrolled to enhance adherence support. Patients who switched timely to alternative first-line at 12 months were 313 and more likely to have viral re-suppression at any time compared to delayed patients on original first-line ART at 12 months. The median with inter-quartile ranges (IQR) for continuous variables were computed after considering distributional assumption tests between the alternative first-line therapy and start of second-line therapy where participants delayed to switch after first virological failure.

Clinical Study Variables	Total ART and Therapy Switch Outcome
Total Currently on ART	3243
Total Ever on ART	3207
Total Enrolled in Care	3179
Total Starting on ART	3130
Original 1st Line at 12 months Retention on ART	2739
ART Net Cohort at 12 months Survival and Retention on ART	2734
Total on therapy at 12 months	2677

2nd Line (or higher) at 12 months Survival and Retention on ART	329
alternative 1st Line at 12 months Survival and Retention on ART	313

Table 7: Clinical outcome and therapy switch

The table below demonstrates Orgunit Name and Health Facilities with highest number of patients retained for ARTs across 47 Counties.

Orgunit Name	Patients Retained for ART care
Sex Workers Outreach Program (Kibra)	25
Swop Korogocho	22
SWOP Clinic Donholm	22
Hoymas VCT (Nairobi)	21
Balambala Sub-District Hospital	17

Table 6: Health Facilities with highest number of patients retained for ARTs

12. DISCUSSION

This study demonstrates the potential of machine learning models to identify patients on ART within County health clinics, allowing busy HIV care clinics to direct limited resources towards patients who need them the most. Clinicians have difficulty predicting patients on ART treatment, risking missing appointments and may be subject to bias in determining which patients would benefit from resource intensive retention interventions. We developed ART predictive model to predict the ART treatment and adherence status from the MOH laboratory database sampling unit. This data comprised of PLWH who are on ART treatment, which constitutes total patients enrolled for ART, those currently on ART, patients starting ART, Ever on ART, original first line and second line We examined four data mining algorithms (Decision tree, Naïve Bayes, K Nearest Neighbor, Random Forest) to build a model that predicts retention on ART and viral suppression. The results of the experiment performance were evaluated based on their accuracy, sensitivity, specificity and area under the ROC curve.

We used the ensemble, trees, Gaussian distribution, Euclidean distance in identification of the most predictive factors of ART treatment outcome, ensemble method delivered better results than the other three algorithms with f1 score on the sample test of 79% and 77% when all variables are included.

For testing options, the experimentation results indicated that the ensemble method (random forest algorithm) performed the best with accuracy score of 81%, K Nearest neighbor came out to be the second best with a classification accuracy of 71%, the decision tree induction method followed (68%). Naïve Bayes classifier achieved the least classification accuracy of 20%. This implies that random forest, k nearest neighbor algorithm and decision tree predictive models were able to predict whether patients were retained in care for ART treatment. We created calibrated confusion matrices of the final model to mimic two clinical scenarios. The first scenario is when a high certainty of patients are retained in care (model calibrated to a specificity of 100%). The second scenario is when a large coverage and optimal trade-off between sensitivity and specificity of ART for viral suppression is required (model calibrated to a sensitivity of 100.

The calibration was done by calculating the average score that yields the desired performance characteristic (eg. sensitivity of 100%) in the training data set. Of note is that the calibration of the model was made on the training set and the confusion matrix calculated on the validation dataset. We found that an increasing number of variables included in the model resulted in increasing predictive performance. The confusion matrix shows that the Naïve Bayes classifier performed poorly on the cohort datasets because it introduces bias and independence assumptions on how weights are estimated given a data sample. The model was therefore not suitable for predictive modelling of retention and viral suppression. On the other hand, random forest classifier gave an exemplary performance on ART cohort dataset because of random subspace method and bagging preventing overfitting, outshining its classical counterpart on both datasets.

The model has a balanced representation of the classification metrics making it less biased and an optimal model for machine learning in predicting ART treatment outcomes.

13. CONCLUSION

In this study, we demonstrate how machine learning algorithm can derive an optimal model to identify patients on ART for continuous care and improve retention. We retrospectively follow 21551 data sample records within county health facilities located in 47 Counties for conducting the experiment.

The research investigated patient on ART treatment from huge data in order to identify patients enrolled in care, starting ART, those currently on ART, Original and alternative 1st Line at 12 months Survival and Retention on ART, which is a base to develop model and also to help medical staff by early identifying patient's ART status to get information about patient for better planning and formulating medical policies. ART coverage was high especially among participants who diagnosed after the treatment strategy. The research can be used for point -of-care interventions in a clinic as well as proactive outreach by a public health department [18]. Emphasis is placed on prediction and tracing patients on ART and those retained in care, especially if patients have recently started ART, including searching non/participating facility records for patients not on ART. Additionally, by anticipating future outcomes (e.g. suppressed VL) and specific targeted interventions can be designed on identified subsets of the treatment cohorts allowing for cost-effective differentiated models on care and treatment to be applied across the cascade.

Retention in care of patients and adherence to the prescribed antiretroviral therapy (ART) regimen is vital for the virological success of the treatment in which the PLHIV fall in viral suppression. This implies that the patient would require to be strictly following the regimen as indicated in order to have an optimal viral suppression to prevent transmission of the virus. This study has shown the use of supervised learning, classification algorithms in prediction of ART adherence and treatment status. The model developed was able to produce predictions that had high accuracy and low errors, illustrating the ability of supervised learning to be used in dispensing centers and comprehensive care centers to perform prediction of ART for patients who are retained in care.

14. RECOMMENDATIONS FOR FUTURE WORKS

Integrating ART adherence and treatment prediction model to end users to make nurses and general practitioners work easy. It is crucial to apply different data mining techniques for the same setting and compare the model performances based on ROC-AUC curve, accuracy, sensitivity and specificity. More research would be done on how to estimate PLHIV CD4 counts by using other treatment features like virological, clinical and

immunological treatment features. Other type of machine learning algorithms such as associative neural networks and recurrent neural network can be used to develop the model. This can ultimately lead to the determination of the optimal neural network to be used in training and developing the model of the system. Efficient distribution of healthcare facilities offering ART and HIV testing services among urban and rural settings are required. Attributes are selected based on business process analysis and facts from literatures. It can also be possible to utilize other feature selection algorithms for the purpose of comparison. Redefining or including predictors and exploring interactions/extensions for behavioral data or clinical measures in Nation-wide settings.

More research could be done on how to estimate the patient's ART adherence to the prescribed regimen by use of their lifestyle. This could provide a basis for computing the adherence percentage which could be used as a baseline data to the developed model.

The developed system should also be interlinked with the currently existing systems to provide seamless integration of the systems.

This will eliminate the need to have to manually transfer data, and ultimately eliminate the errors attributable to human intervention.

The data could be used to train the model periodically thus resulting to improved prediction accuracy.

15. ACKNOWLEDGMENT

The authors wish to thank CEMA-Africa who funded the project and to Prof Thumbi Mwangi, Dr Evans Mirithi and Mrs Selina Atwani who were involved in the validation for this research project.

16. REFERENCES

[1] World Granich RM, Gilks CF, Dye C,(2009); Williams BG. Universal voluntary HIV testing with immediate antiretroviral therapy as a strategy for elimination of HIV transmission: a mathematical model.

[2] Ministry of Health, MOH, (2020), *Kenya AIDS Strategic Framework II, Sustain Gains, Bridge Gaps and Accelerate Progress.*

[3] Slaymaker E, Hosegood V, (2014) Scale and distribution of excess deaths among HIV positive adults by diagnosis, care and treatment history in African population based cohorts 2007 – 2011.

[4] Kremer H, Ironson, (2006); Why people with HIV share or don't share with their physicians whether they are taking their medications as prescribed.

[5] Wachira J, Naanyu V, Koech B, Akinyi J, (2014) Health facility barriers to HIV linkage and retention in western Kenya.

[6] Rakhi Karwa, Mercy Maina, (2017); Leveraging peer-based support to facilitate HIV care in Kenya.

[7] Renju J, et al; (2017). Side effects are central effects; a multi-country qualitative study to understand the challenges of retention in HIV care and treatment programmes in sub–Saharan Africa.

[8] Miller WC, Powers Ka, Smith MK, (2013) Community viral load as a measure for assessment of HIV treatment as prevention.

[9] R.S Butt, & I Ahmad, (2019); Integral backstepping and synergetic control for tracking of infected cells during early antiretroviral therapy.

[10] T. Heestermans, J. L Browne, (2016); Determinants of adherence to antiretroviral therapy among HIV-positive adults in Sub-Saharan Africa; A systematic review.

[11] J. Estill, & N. Ford, L. Salazar- Vizcaya et al (2016); The need for second-line antiretroviral therapy in adults in sub-Saharan Africa up to 2030*; A mathematical modelling study*.

[12] J. Davey, & A. Nguimfack, S. Hares, W. Ponce, (2012); Evaluating SMS reminders in improving ART and PMTCT adherence in Mozambique: Challenges in achieving scale.

[13] R. Baraldi, K. Cross, C. McChesney et al, (2014); Uncertainty quantification for a model of HIV-1 patient response to antiretroviral therapy interruptions , *" in proceedings of the 2014 American Control Conference.*

[14] T. D Chala, (2019): Data mining technology enabled antiretroviral therapy (ART) for HIV positive patients in Gondar university hospital, Ethiopia, Bioinformation.

[15] Barradas DT, Gupta S, Moyo C, Sachathep K, Nkumbula T, et al (2017). *Findings from the 2016 Zambia Population-based HIV Impact Assessment (ZAMPHIA):* HIV prevalence, incidence and progress towards the 90-90-90 goals. Abstract TUAC0301.

[16] Geng EH, Bangsberg DR, Bwana MB, Yiannoutsos CT, et al (2010). Understanding reasons for and outcomes of patients lost to follow-up in antiretroviral therapy programs in Africa through a sampling-based approach. *https://doi.org/10.1097/QAI.0b013e3181b843f0*.

[17] Maskew M, & Sharpey-Schafer K, (2010). Machine learning to predict retention and viral suppression in South African HIV treatment cohorts.

[18] Arthi Ramachandran, Avishek Kumar (2020) Predictive Analytics for Retention in care in an Urban HiV clinic https://doi.org/10.1038/s41598-020-62729.