Using machine learning models to plan HIV services: Emerging opportunities in design, implementation and evaluation

T Dzinamarira,[1] PhD; E Mbunge,[2] PhD; I Chingombe,[3] MSc; D F Cuadros,[4] PhD; E Moyo,[5] MB ChB; I Chitungo,[6] MSc;
G Murewanhema,[6] MMed(O&G); B Muchemwa,[2] MSc; G Rwibasira,[7] MD; O Mugurungi,[8] MD; H Herrera,[9] PhD; G Musuka,[10] PhD

[1] School of Health Systems and Public Health, University of Pretoria, Pretoria, South Africa
[2] Department of Computer Science, University of Eswatini, Manzini, Eswatini
[3] Chinhoyi University of Technology, Chinhoyi, Zimbabwe
[4] Department of Geography and Geographic Information Science, University of Cincinnati, Cincinnati, USA
[5] Department of Public Health, Oshakati Medical Center, Oshakati, Namibia
[6] College of Medicine and Health Sciences, University of Zimbabwe, Harare, Zimbabwe
[7] HIV, STIs, Viral Hepatitis and other Viral Diseases Control Division, Rwanda Biomedical Center, Kigali, Rwanda
[8] AIDS and TB Program, Ministry of Health and Child Care, Harare, Zimbabwe
[9] School of Pharmacy and Biomedical Sciences, University of Portsmouth, UK
[10] International Initiative for Impact Evaluation, Harare, Zimbabwe

Corresponding author: T Dzinamarira (u19395419@up.ac.za)

HIV/AIDS remains one of the world's most significant public health and economic challenges, with approximately 36 million people currently living with the disease. Considerable progress has been made to reduce the impact of HIV/AIDS in the past years through successful multiple HIV/AIDS prevention and treatment interventions. However, barriers such as lack of engagement, limited availability of early HIV-infection detection tools, high rates of HIV/sexually transmitted infections (STIs), barriers to access antiretroviral therapy, lack of innovative resource optimisation and distribution strategies, and poor prevention services for vulnerable populations still exist and substantially affect the attainment of the UNAIDS 95-95-95 targets. A rapid review was conducted from 24 October 2022 to 5 November 2022. Literature searches were conducted in different prominent and reputable electronic database repositories including PubMed, Google Scholar, Science Direct, Scopus, Web of Science, IEEE Xplore, and Springer. The study used various search keywords to search for relevant publications. From a list of collected publications, researchers used inclusion and exclusion criteria to screen and select relevant papers for inclusion in this review. This study unpacks emerging opportunities that can be explored by applying machine learning techniques to further knowledge and understanding about HIV service design, prediction, implementation, and evaluation. Therefore, there is a need to explore innovative and more effective analytic strategies including machine learning approaches to understand and improve HIV service design, planning, implementation, and evaluation to strengthen HIV/AIDS prevention, treatment, and awareness strategies.

Despite the significant progress made in previous years to reduce the catastrophic impact of HIV/AIDS, they are among the most threatening infectious diseases and continue to overburden public health systems worldwide. Key public health interventions including HIV screening, increasing universal and scaling-up of antiretroviral therapy (ART), and improving pre-exposure prophylaxis (PrEP) delivery[1] have been utilised to improve health outcomes in people living with HIV (PLHIV) and reduce new infections. However, these key HIV interventions and prevention measures continue to fall short of attaining the UNAIDS 95-95-95 targets. Emerging challenges, including a lack of effective innovative HIV awareness services,[2] lack of engagement to care, few early HIV-infection detection tools, high prevalence of HIV infection and sexually transmitted infections (STIs), lack of innovative resource optimisation tools and distribution strategies, poor linkage and retention,[3] lack of health workers to facilitate HIV testing and counselling, and poor prevention services for vulnerable populations, still exist, which consequently affect the attainment of UNAIDS 95-95-95 targets. For these reasons, there is a need

to integrate emerging digital health technologies such as machine learning (ML) to alleviate challenges while working towards UNAIDS 95-95-95 targets.

The increased availability of HIV/AIDS-related health data presents unprecedented opportunities to develop intelligent models that can substantially transform HIV service design and enhance targeted HIV prevention and detection. Data repositories, such as electronic health records (EHR), health information systems, surveys (bio-behavioural data, population-based HIV impact assessment (PHIA)), and clinical information, generate a tremendous amount of data that increase chances of applying ML techniques to identify and establish patterns of HIV-related insights, improve models' prediction capacity, and enhance HIV clinical research and care.[4]

Despite the increased availability of HIV-related data, the development of ML-based HIV data-driven applications is still nascent in many developing countries,[5] especially in sub-Saharan Africa. Therefore, there is a need to maximise presenting opportunities to apply ML techniques to develop data-driven HIV-intelligent tools to improve HIV response, particularly to emerging HIV-related

challenges in different high-risk communities,[6] and to provide targeted interventions to hard-to-reach vulnerable communities, such as men who have sex with men (MSM).

In this study, we conducted a comprehensive review to find ways that ML can be implemented to understand HIV service design, prediction, implementation, and assessment. The success of the HIV programmes can be enhanced because of ML's contribution to better programme design and execution. This would assist countries in achieving the UNAIDS goal of ending the HIV/AIDS pandemic as a global public health problem by 2030.

Methods

A rapid review was conducted from 24 October 2022 to 5 November 2022. Literature searches were conducted in different prominent and reputable electronic database repositories including PubMed, Google Scholar, Science Direct, Scopus, Web of Science, IEEE Xplore, and Springer. Two reviewers screened articles for eligibility. No quality appraisal was conducted.

Search strategy

The study used various keywords to search for relevant publications. From a list of collected publications, researchers used inclusion and exclusion criteria to screen and select relevant papers as shown in Fig. 1. The search keywords used were "HIV/AIDS" OR "HIV" OR "AIDS", AND "predicting" OR "machine learning" OR "forecasting" OR "service design implementation".

Inclusion and exclusion criteria

Using the PRISMA model (Fig. 1), the study included papers that were written in English or with an English translation and further excluded papers that did not specify ML models/algorithms used to predict HIV/AIDS. Incomplete articles, letters to the editor, and preprints were excluded from this study. The study included papers that applied ML techniques to improve HIV/AIDS service design, implementation, and prediction of HIV status.
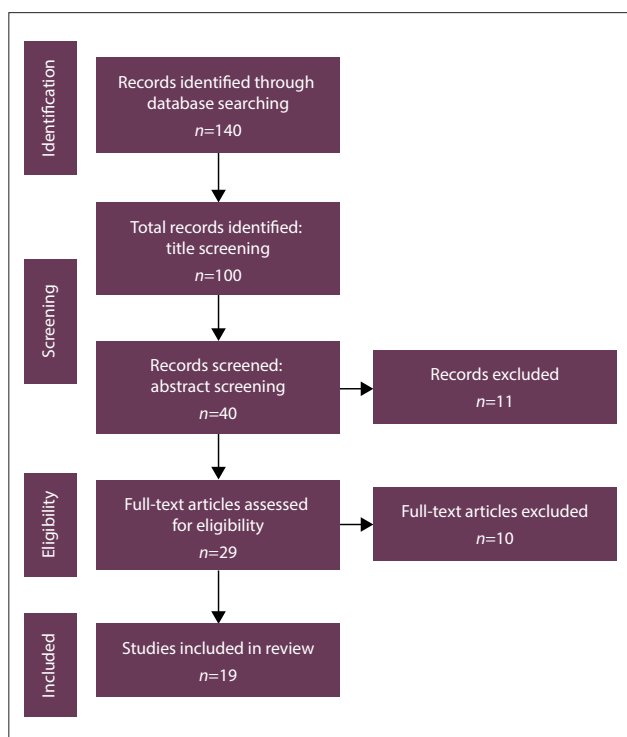

Fig. 1. PRISMA model.

Screening process

The titles, abstracts, and content of articles were screened, strictly based on the inclusion and exclusion criteria. Any discrepancies that arose during screening were resolved through a consensus by at least two leading authors.

Data synthesis

After the full-text screening, the authors extracted the reference, ML model(s) used, the purpose of the study, population groups/participants, and the limitations of the study. The selected papers are shown in Table 1.

Results

A total of 19 studies met the criteria to be included in this review. More details are presented in Table 1.

HIV/AIDS prediction models

Several techniques can be used to predict various elements of HIV care, including diagnosis, choice of medication, the success of treatment, risk of getting HIV, and risk of developing resistance to ART. The Gaussian Naïve Bayes, support vector machines (SVMs), bagging classifiers, gradient boosting classifiers, recurrent neural networks (RNNs), logistic regression, and random forests are some of the models in this category.[7]

Gaussian Naïve Bayes

A strong independence assumption and the application of the Bayes theorem form the basis of this probabilistic classification technique. This naïve Bayes classifier has a high level of expressiveness, scalability, and accuracy. Several characteristics contribute to this model's effectiveness. Some of the features include that the model scales well with the size of the training dataset, it can handle continuous features, and it does not need the classification model's parameters to be tuned.[8]

Support vector machines (SVMs)

SVMs are a form of supervised learning method used for classification, regression, and outliers' detection. SVMs are effective in high-dimensional spaces, memory efficient, and versatile. SVMs in Scikit-learn support both dense and sparse sample vectors as input.[9]

Bagging classifier

An ensemble meta-estimator called a bagging classifier model fits base classifiers one at a time to random subsets of the original dataset, and it then averages or votes on each classifier's predictions to produce a final prediction. Such a meta-estimator can typically be used as a way to reduce the variance of a black-box estimator like a decision tree, by introducing randomisation into its construction procedure and then making the ensemble out of it.[10]

Gradient boosting classifier

These are a collection of ML algorithms that pool several weak learning models to produce a strong predictive model. Gradient boosting frequently makes use of decision trees. These models are gaining popularity because they are good at classifying large, complicated datasets. This algorithm constructs an additive model in a stage-by-stage manner. The optimisation of any differentiable loss function is possible.[11] One kind of gradient boosting classifier is Extreme Gradient Boosting (XGBoost). It is a distributed gradient-boosting library that has been developed to be very effective, adaptable, and portable. It offers parallel tree boosting to address a variety of data science tasks quickly and accurately.[12] A study

Table 1. Findings from included studies

References	ML models	Purpose	Population	Limitations
Marcus et al., 2020[1]	• ML techniques	HIV prevention	General population	• The model applied systematic analysis of the ML application to enhance HIV prevention
Mutai et al., 2021[2]	• XGBoost algorithm • Elastic-net regression • K-nearest neighbours • Random forest • Support vector machine • Light gradient boosting	Identifying predictors for HIV status	PHIA data, in sub-Saharan countries	• Hypermeter turning and optimisation of the models is needed and deploy them in the real healthcare setting. • High degree of missingness and inconclusiveness from self-reported data
Olatosi et al., 2021[3]	• Bayesian network • Random forest • Decision tree • Neural network	HIV medical care status classification	PLWH data, South Carolina	• Limited dataset from PLWH
Balzer et al., 2020[6]	• Logistic regression	Construct HIV risk scores	General population data, Uganda and Kenya	• Data used were collected from individuals who at least took two HIV tests, other potential participants were omitted • HIV risk scores generalisability
Chingombe et al., 2022[7]	• Recurrent neural networks • Gradient boosting classifier • Support vector machines • Gaussian naïve Bayes	Predicting HIV status	MSM in Zimbabwe	• Survey data are subject to social desirability bias • The study does not show a temporal relationship between exposures and outcomes.
Marcus et al., 2019[19]	• ML and LASSO	Identify candidates for HIV PrEP	EHR data, USA	• The EHR-based HIV risk prediction model needs to be optimised
Young et al., 2017[22]	• Random forest • Logistic regression • Support vector machine • Ridge regression classifier	HIV identification	Tweets data	• Social media data tend to be biased • The study did not use clinical health data to identify HIV
Domínguez-Rodríguez et al., 2022[26]	• Logistic regression • Support vector machine • Naïve Bayes • K-nearest neighbour • Elastic-net regression	Predict paediatric HIV mortality	Early ART in HIV-positive children, South Africa and Mozambique	• Limited sample size and limited number of observations, which likely diminishes the prediction capacity of ML models
Orel et al., 2022[27]	• XGBoost algorithm • Support vector machine • Elastic-net regression	HIV status prediction	General population, Demographic and Health Survey, Southern and Eastern Africa	• Generalisability of the predictive models for countries that were not used to train the algorithms
Xu et al., 2022[30]	• Random forest • Gradient boosting machine • Elastic-net regression	HIV risk prediction	General population, Australia	• Data used for prediction depend on self-reported information. • HIV risk factors can change over time due to factors including PrEP
Chingombe et al., 2022[23]	• Logistic regression • Random forest classifier • Support vector machine	Predicting HIV status	Zimbabwe Population-Based HIV Impact Assessment	• The study did not include other HIV clinical predictors to optimise HIV status prediction models
Wray et al., 2019[29]	• Random forest	Identify HIV risk factors	Data were collected from gay-oriented smartphone dating applications	• The study used self-reported data which might be biased.
Ahlström et al., 2019[30]	• Random forest • Logistic regression	Predicting HIV status	Nation-wide electronic registry data, Denmark	• The study assumed that the undiagnosed individuals with HIV resemble the PLWH
Krakower et al., 2019[31]	• LASSO	HIV prediction and identify candidates for PrEP	EHR data, USA	• The models cannot be universally used because electronic registry data vary inevitably by country

...continued

Table 1. (continued) Findings from included studies

References	ML models	Purpose	Population	Limitations
Bao et al., 2021[24]	• Gradient boosting machine • Random forest • Multivariable logistic regression	Predicting HIV/STI diagnosis	Australian MSM	• Validity of predictors from self-reported data which is subject to recall and non-response bias • Data imputation for missing data may introduce uncertainty in model calibration and prediction inaccuracy
Feller, et al., 2018[25]	• Random forest	HIV risk estimation	General population, EHR, New York	• The dataset was small owing to the non-sharing of clinical data
Xu et al., 2022[24]	• Elastic-net regression • Support vector machine • Boosted GLM • Gradient boosting machine • XGBoost • Naïve Bayes • K-nearest neighbour	HIV risk prediction	EHR data, Melbourne	• Social-desirability bias • Highly imbalanced data
Paul et al., 2020[32]	• Linear gradient-boosted multivariate regression • Logistic regression	Frailty in HIV	Older PLWH data, St Louis	• The dataset used was small
Kozak et al., 2007[33]	• Support vector machine	Detecting in eyes of HIV patients	PLWH, San Diego, California	

GLM = general linear model.

by Mutai et al.[2] applied several ML algorithms to identify HIV predictors for screening and the XGBoost algorithm outperformed other models by achieving F1-scores of 90% and 92% for males and females, respectively.

Recurrent neural networks (RNNs)
RNNs, which are descended from feedforward networks, behave in a manner resembling behaviour of the human brain. RNNs contain internal memory that enables them to retain key details from the input they receive, enabling them to make extremely accurate predictions about what will happen next.[13] Sequential data are used by RNNs to provide prediction outcomes. RNNs can be categorised as one-to-one, one-to-many, many-to-one, and many-to-many.[14] For instance, a study conducted by Chingombe et al.[7] applied recurrent neural networks to predict HIV status among MSM using bio-behavioural data and achieved a high accuracy of 98%.

Logistic regression
A common use of supervised learning is the prediction of the likelihood that a binary event will occur. A good illustration of this is logistic regression. Several independent variables that can be categorical or numerical can be used to determine the likelihood of falling into one of the two groups.[15] Multinomial logistic regression is used when there are several outcomes, and ordinal logistic regression is used when the outcome is ordered. The independent variables in logistic regression must not be dependent on one another; that is, they must not be highly multicollinear.[16]

Random forest
Popular learning algorithms that fall under the category of supervised learning include the random forest. It can be applied to classification and regression issues in ML. It is a classifier that uses multiple decision trees on different subsets of the provided dataset and averages the results to increase the dataset's predicted accuracy. [17] Instead of relying solely on one decision tree, the random forest gathers forecasts from all the trees and predicts the result depending on which predictions received the most votes overall. Higher accuracy and prevention of overfitting result from the use of a large number of trees in the forest.[17]

Application of ML techniques in HIV/AIDS programmes
The capacity of ML approaches to accelerate the scaling-up of ART and PrEP in communities that are most likely to benefit will determine the extent to which they are useful for HIV prevention.[1] ML can find possible PrEP candidates by combining multiple HIV risk scores. The extensive data found in EHRs can be utilised to estimate the risk of contracting HIV and improve HIV/AIDS diagnoses.[18] EHRs contain rich information that can be used to predict HIV risk, including demographic characteristics, social history, diagnoses, laboratory tests and results, and prescriptions.[19] To identify individuals who are at a higher risk of contracting HIV and would benefit from PrEP there are validated HIV-risk-prediction algorithms that employ EHR data and ML.[1] When ML is used, it is crucial to consider the trade-offs between algorithms that can be applied to all healthcare settings but have lower predictive performance and those that are specific to each setting but require a higher investment of resources, since they are likely to be highly predictive for that setting's particular patient population.[1]

Sociodemographic data can be used by ML models to forecast a person's HIV status. Directing testing toward those who have been determined to be most likely to be HIV-positive can help the public become more aware of their own HIV status. These tailored screening techniques may lower the cost of HIV testing in settings with constrained resources.[2] ML can also use data from smartphones and social media to predict condomless sex and the real-time data can potentially be used to guide just-in-time interventions like text messages. ML can potentially be used to facilitate HIV serodisclosure.[10] Numerous computational approaches have been employed to identify mutations linked to drug resistance and to forecast HIV resistance in individuals on ART.[20] ML can also be used to identify content in social media that is associated with HIV risk behaviours, and this may help in targeting interventions for such individuals.[21]

Limitations and identified challenges

The study revealed that there are insufficient clinical HIV/AIDS datasets and that these are difficult to access; therefore, most studies used survey data (demographic and health survey data and PHIA data) and nationwide electronic registry data. For instance, a study conducted by Marcus et al.[19] applied ML to identify candidates for HIV PrEP using data from the EHR. Also, Mutai et al.[2] applied ML models to identify HIV predictors using PHIA data and stated that there was a high degree of missingness and inconclusiveness from self-reported data that potentially impacted the training data. A study by Chingombe et al.[7] also used survey data to predict HIV status among MSM using deep learning and ML models and achieved good prediction accuracy. Furthermore, datasets are often released late; hence, while on paper results may reveal some excellent model predictive capacity, when applied in real life they may perform poorly as a result of changes in context over time.

Survey datasets have several limitations, including social desirability bias (self-reported information subject to recall and non-response bias), and often do not show a temporal relationship between exposures and outcomes, which consequently affects the generalisability of the predictive models. In addition, owing to several key interventions to reduce new HIV infections, identifying HIV risk factors using survey data is proving not to be enough because HIV risk factors can change over time because of various factors, e.g. PrEP reduced HIV risk substantially but condom use declined in the PrEP era,[22] drug adherence, and drastic change of other social and individual behavioural patterns. Therefore, there is a need to integrate HIV clinical data as inputs to improve the performance of ML predictive models.

Our study also revealed that in most cases, especially in hard-to-reach communities such as MSM, the datasets are usually small, which consequently affects the performance of ML models and potentially causes the model to overfit. A study by Chingombe et al.[23] applied ML techniques to predict HIV status among MSM using a small dataset. With a limited dataset, policymakers solemnly rely on linear combinations; this then leads to the missed opportunity of applying advanced predictive models to improve the HIV care continuum.[3] However, it is difficult to gather data for communities like MSM because of multiple factors, including stigmatisation, discrimination,[23] and isolation in the community,[24] especially in countries where same-sex sexual behaviours are illegal.

Another limitation identified in this study is that several studies including Xu et al.[24] applied class imbalance techniques and data imputation methods to handle missing values. However, predicting HIV status, and identifying HIV predictors and candidates from PrEP using datasets with missing values or imputed data may introduce uncertainty in model calibration and prediction inaccuracy.[24]

A study by Olatosi et al.[3] highlighted that there is a lack of interoperability among datasets, especially for PLWH to improve the HIV care continuum and ultimately achieve complete viral suppression among PLWH. However, the lack of integrated evidence from different HIV data sources such as clinical, PHIA projects and community-based datasets has led to missed opportunities to extract HIV patterns to capture a complete portrait of PLWH's health status, improve HIV testing and treatments and proactively identify HIV drug defaulters. There is a need to integrate health data sources and improve the interoperability of health information systems and electronic health records to maximise the HIV care continuum and intensity of administration of HIV key interventions.

Overall, the study revealed that there is a lack of real-time data-driven applications integrated with national health information systems and EHR data to improve HIV service design, map vulnerable communities, automate risk assessment,[25] identify HIV risk factors in real-time, predict retention-in-care,[3] and intensify HIV awareness, testing and ART adherence. There is a need to integrate various health data sources and develop data-driven clinical ML-based applications that could be further used to improve HIV diagnosis tools, HIV early detection, and diagnosis.

Recommendations to improve HIV service design for key populations

Firstly, there is a need to improve data availability through the promotion of research, particularly among key populations which are criminalised, to generate data for the training of ML models to improve their accuracy and in turn the design of interventions. Secondly, we recommend that HIV intervention designs need to be informed by current or most-near-current data. There is a need for frequent updating of data sources, especially among the key populations that are criminalised and challenging to study. For instance, before and after the behavioural survey among MSM in Zimbabwe, which took place in 2018, there was no other study performed to inform HIV response targeting MSM.

The fact however remains that the criminalised key populations continue to be central for HIV transmission, within their circles, as well as in the general population, thereby eroding the gains that might have been achieved from all past efforts.

Thirdly, it is critical to understand and factor in the diversity that exists within populations and sub-populations when designing and implementing HIV interventions. With most countries nearing epidemic control, there is a need for additional interventions that target the needs of specific individuals to reach out to the 'few' remaining high-risk cases within subgroups, rather than spreading resources equally among those where the risks are low. Targeted interventions are relevant, particularly at this juncture where resources to support HIV responses are not only dwindling but also strained by the need to address other public health disasters like COVID-19 and others like cyclones and droughts.

Conclusion

Evidence from the current comprehensive review has shown the potential of applying ML models to understand HIV service design, planning, prediction, implementation, and evaluation, including among the key populations. The main challenges noted are around data management and the resources and skills necessary for this. There is limited availability of the training of ML models to improve the accuracy and generalisability of the results. To realise the full potential of the role of ML techniques in supporting the HIV service design, planning, implementation, and evaluation, it is imperative that data for both the general and all other hard-to-reach HIV high-risk subpopulations are systematically generated on a consistent regular basis. Failure to generate such data will continue to impact the effectiveness of the HIV interventions and consequently erode gains from past efforts.

1. Marcus JL, Sewell WC, Balzer LB, Krakower DS. Artificial intelligence and machine learning for HIV prevention: Emerging approaches to ending the epidemic. Curr HIV/AIDS Rep 2020;17(3):171. https://doi.org/10.1007/S11904-020-00490-6

2. Mutai CK, McSharry PE, Ngaruye I, Musabanganji E. Use of machine learning techniques to identify HIV predictors for screening in sub-Saharan Africa. BMC Med Res Methodol 2021;21(1):1-11. https://doi.org/10.1186/S12874-021-01346-2

3. Olatosi B, Sun X, Chen S, et al. Application of machine-learning techniques in classification of HIV medical care status for people living with HIV in South Carolina. AIDS 2021;35:S19-S28. https://doi.org/10.1097/QAD.0000000000002814

4. Bisaso KR, Anguzu GT, Karungi SA, Kiragga A, Castelnuovo B. A survey of machine learning applications in HIV clinical research and care. Comput Biol Med 2017;91:366-371. https://doi.org/10.1016/J.COMPBIOMED.2017.11.001

5. Mbunge E, Batani J. Application of deep learning and machine learning models to improve healthcare in sub-Saharan Africa: Emerging opportunities, trends and implications. Telemat Informatics Reports 2023;11:100097. https://doi.org/10.1016/J.TELER.2023.100097

6. Balzer LB, Havlir DV, Kamya MR, et al. Machine learning to identify persons at high-risk of human immunodeficiency virus acquisition in rural Kenya and Uganda. Clin Infect Dis 2020;71(9):2326-2333. https://doi.org/10.1093/CID/CIZ1096

7. Chingombe I, Dzinamarira T, Cuadros D, et al. Predicting HIV status among men who have sex with men in Bulawayo & Harare, Zimbabwe: Using bio-behavioural data, recurrent neural networks, and machine learning techniques. Trop Med Infect Dis 2022;7(9):231. https://doi.org/10.3390/TROPICALMED7090231

8. Anand MV, Kiranbala B, Srividhya SR, Kavitha C, Younus M, Rahman MH. Gaussian Naïve Bayes algorithm: A reliable technique involved in the assortment of the segregation in cancer. Mob Inf Syst 2022;2022. https://doi.org/10.1155/2022/2436946

9. Guenther N, Schonlau M. Support vector machines. 2016;16(4):917-937. https://doi.org/10.1177/1536867X1601600407

10. Breiman L. Bagging predictors. Mach Learn 1996;24(2):123-140. https://doi.org/10.1007/BF00058655

11. Biau G, Cadre B, Rouvière L. Accelerated gradient boosting. Mach Learn 2019;108(6):971-992. https://doi.org/10.1007/S10994-019-05787-1/TABLES/5

12. Zhong L, Hu L, Zhou H. Deep learning based multi-temporal crop classification. Remote Sens Environ 2019;221:430-443. https://doi.org/10.1016/J.RSE.2018.11.032

13. Mbunge E, Batani J, Mafumbate R, et al. Predicting student dropout in massive open online courses using deep learning model: A systematic review. Lect Notes Networks Syst 2022;503 LNNS:212-231. https://doi.org/10.1007/978-3-031-09073-8_20/COVER

14. Rezk NM, Purnaprajna M, Nordström T, Ul-Abdin Z. Recurrent neural networks: An embedded computing perspective. IEEE Access 2019;8:57967-57996. https://doi.org/10.1109/access.2020.2982416

15. Mbunge E, Chemhaka G, Batani J, et al. Predicting diarrhoea among children under five years using machine learning techniques. Lect Notes Networks Syst 2022;502 LNNS:94-109. https://doi.org/10.1007/978-3-031-09076-9_9/COVER

16. Maalouf M. Logistic regression in data analysis: An overview. Int J Data Anal Tech Strateg 2011;3(3):281-299. https://doi.org/10.1504/IJDATS.2011.041335

17. Breiman L. Random forests. Mach Learn 2001;45(1):5-32. https://doi.org/10.1023/A:1010933404324

18. Haas O, Maier A, Rothgang E. Machine learning-based HIV risk estimation using incidence rate ratios. Front Reprod Health 2021;3:96. https://doi.org/10.3389/FRPH.2021.756405

19. Marcus JL, Hurley LB, Krakower DS, Alexeeff S, Silverberg MJ, Volk JE. Use of electronic health record data and machine learning to identify candidates for HIV pre-exposure prophylaxis: A modelling study. Lancet HIV 2019;6(10):e688-e695. https://doi.org/10.1016/S2352-3018(19)30137-7

20. Blassel L, Tostevin A, Villabona-Arenas CJ, Peeters M, Hué S, Gascuel O. Using machine learning and big data to explore the drug resistance landscape in HIV. PLOS Comput Biol 2021;17(8):e1008873. https://doi.org/10.1371/JOURNAL.PCBI.1008873

21. Young SD, Yu W, Wang W. Toward automating HIV identification: machine learning for rapid identification of HIV-related social media data. J Acquir Immune Defic Syndr 2017;74(Suppl 2):S128. https://doi.org/10.1097/QAI.0000000000001240

22. Xu X, Ge Z, Chow EPF, et al. A machine-learning-based risk-prediction tool for HIV and sexually transmitted infections acquisition over the next 12 months. J Clin Med 2022;11(7):1818. https://doi.org/10.3390/JCM11071818/S1

23. Chingombe I, Musuka G, Mbunge E, et al. Predicting HIV status using machine learning techniques and bio-behavioural data from the Zimbabwe Population-Based HIV Impact Assessment (ZIMPHIA15-16). Lect Notes Networks Syst 2022;502 LNNS:247-258. https://doi.org/10.1007/978-3-031-09076-9_24/COVER

24. Bao Y, Medland NA, Fairley CK, et al. Predicting the diagnosis of HIV and sexually transmitted infections among men who have sex with men using machine learning approaches. J Infect 2021;82(1):48–59. https://doi.org/10.1016/J.JINF.2020.11.007

25. Feller DJ, Zucker J, Yin MT, Gordon P, Elhadad N. Using clinical notes and natural language processing for automated HIV risk assessment. J Acquir Immune Defic Syndr 2018;77(2):160. https://doi.org/10.1097/QAI.0000000000001580

26. Domínguez-Rodríguez S, Serna-Pascual M, Oletto A, et al. Machine learning outperformed logistic regression classification even with limit sample size: A model to predict pediatric HIV mortality and clinical progression to AIDS. PLoS ONE 2022;17(10):e0276116. https://doi.org/10.1371/JOURNAL.PONE.0276116

27. Orel E, Esra R, Estill J, et al. Prediction of HIV status based on socio-behavioural characteristics in East and Southern Africa. PLoS ONE 2022;17(3):e0264429. https://doi.org/10.1371/JOURNAL.PONE.0264429

28. Xu X, Yu Z, Ge Z, et al. Web-based risk prediction tool for an individual's risk of HIV and sexually transmitted infections using machine learning algorithms: Development and external validation study. J Med Internet Res 2022;24(8). https://doi.org/10.2196/37850

29. Wray TB, Luo X, Ke J, Pérez AE, Carr DJ, Monti PM. Using smartphone survey data and machine learning to identify situational and contextual risk factors for HIV risk behavior among men who have sex with men who are not on PrEP. Prev Sci 2019;20(6):904-913. https://doi.org/10.1007/S11121-019-01019-z

30. Ahlström MG, Ronit A, Omland LH, Vedel S, Obel N. Algorithmic prediction of HIV status using nation-wide electronic registry data. eClinicalMedicine 2019;17. https://doi.org/10.1016/J.ECLINM.2019.10.016

31. Krakower DS, Gruber S, Hsu K, et al. Development and validation of an automated HIV prediction algorithm to identify candidates for pre-exposure prophylaxis: A modelling study. Lancet HIV 2019;6(10):e696-–e704. https://doi.org/10.1016/S2352-3018(19)30139-0

32. Paul RH, Cho KS, Luckett P, et al. Machine learning analysis reveals novel neuroimaging and clinical signatures of frailty in HIV. J Acquir Immune Defic Syndr 2020;84(4):414. https://doi.org/10.1097/QAI.0000000000002360

33. Kozak I, Sample PA, Hao J, et al. Machine learning classifiers detect subtle field defects in eyes of HIV individuals. Trans Am Ophthalmol Soc 2007;105:111.