



A survey of machine learning applications in HIV clinical research and care



Kuteesa R. Bisaso^{a,b,c,*}, Godwin T. Anguzu^a, Susan A. Karungi^b, Agnes Kiragga^a,
Barbara Castelnovo^a

^a Infectious Diseases Institute, College of Health Sciences, Makerere University, Kampala, Uganda

^b Department of Pharmacology and Therapeutics, College of Health Sciences, Makerere University, Kampala, Uganda

^c Breakthrough Analytics Limited, Kampala, Uganda

ARTICLE INFO

Keywords:

HIV
Machine learning
Clinical research
Application paradigms

ABSTRACT

A wealth of genetic, demographic, clinical and biomarker data is collected from routine clinical care of HIV patients and exists in the form of medical records available among the medical care and research communities. Machine learning (ML) methods have the ability to identify and discover patterns in complex datasets and predict future outcomes of HIV treatment. We survey published studies that make use of ML techniques in HIV clinical research and care. An advanced search relevant to the use of ML in HIV research was conducted in the PubMed biomedical database. The survey outcomes of interest include data sources, ML techniques, ML tasks and ML application paradigms.

A growing trend in application of ML in HIV research was observed. The application paradigm has diversified to include practical clinical application, but statistical analysis remains the most dominant application. There is an increase in the use of genomic sources of data and high performance non-parametric ML methods with a focus on combating resistance to antiretroviral therapy (ART). There is need for improvement in collection of health records data and increased training in ML so as to translate ML research into clinical application in HIV management.

1. Introduction

Machine learning (ML) is generating enormous buzz and gaining importance in many domains. It has given birth to applications such as self-driving cars, speech and language recognition, optical character recognition, e-commerce online recommender systems, fraud detection, email filtering and most recently precision medicine. This trend in popularity of ML is driven by improvements in data collection and storage, and advancement in computing power (processing, memory and storage) over the past decade. Together, these two factors have spurred the use of computers to tackle increasingly complex tasks [1,2].

Machine learning provides computers with the ability to learn without being explicitly programmed. It uses complex algorithms and techniques to recognize patterns in data in order to make predictions. Over the past two decades, there has been a steady increase in the medical research involving ML, progressing dramatically from laboratory curiosity to a practical clinical application. This is largely attributed to growing volumes of clinical, social, epidemiological, genetic and other types of medical data that are overwhelming for humans to infer from

and make decisions. Consequently, ML has been envisaged to improve medical practice through “better decision-making, optimized innovation, improved research/clinical trial efficiency, and new tool creation for physicians, consumers, insurers, and regulators” [3–6].

HIV/AIDS remains one of the world's most significant public health and developmental challenges. Despite tens of millions of AIDS related deaths since the beginning of the epidemic in 1981, approximately 36 million people current live with HIV. Approximately 19 million people with HIV are enrolled in routine care programs and receiving treatment [7]. The management of HIV is complicated by the wide variability in both host and viral genetic makeup, dozens of options of ART to chose from, multiple opportunistic infections, demographic differences in disease progression and response to ART as well as socio-cultural differences in acceptability and adherence to treatment, all of which necessitate personalization of care and treatment [8].

A wealth of demographic, clinical and biomarker data is collected from routine clinical care of HIV patients. This data exists in form of medical records available among the medical care and research communities. These large amounts of data make HIV research and treatment

* Corresponding author. Infectious Diseases Institute, College of Health Sciences, Makerere University, Kampala, Uganda.

E-mail address: kuteesar@gmail.com (K.R. Bisaso).

a potential beneficiary of ML. The extent of ML application in HIV research and treatment remains unclear. The purpose of this survey is to explore the nature and trends in ML application in HIV/AIDS clinical research and management.

2. Machine learning methods

The key goal of ML is to use an example dataset to map out the characteristics that are most helpful in predicting an outcome of interest, and apply those characteristics to accurately predict outcomes in a new situations not previously encountered [9]. This is referred to as prediction, a process made possible via Bayesian statistics which allows learning a probability distribution from data and utilization of inverse probability to infer the unknowns in future data [10].

A diverse array of machine-learning methods (models and algorithms) has been developed to tackle the wide nature of tasks. These methods are broadly classified into supervised and unsupervised learning [11].

Supervised learning methods search for a function $f(x)$ that predicts a target/output variable (y) given a set of predictor/input variables (x). The training data is called labeled data because it consists of (x,y) pairs of variables. The inputs x may be simple vectors or more complex objects such as texts, DNA sequences, molecular structures, images, graphs or videos. Outputs (or labels) may include continuous outcomes or the more common binary yes-or-no outcomes. Regression learning methods predict outcomes in a continuous spectrum while classification learning methods predict outcomes of a categorical (binary) nature [11]. Abundant research has been done on problems such as multiclass classification (where y takes on one of *more than 2* labels), multi-label classification (where y is labeled simultaneously by several of the K labels), ranking problems (where y provides a partial order on some set), and general structured prediction problems (where y is a combinatorial object such as a graph, whose components may be required to satisfy some set of constraints) [2].

A number of supervised learning methods have been developed. These include multiple linear regression (M-LR), decision trees and forests (DT, DF) [12], logistic regression (LR), support vector machines (SVM) [13–16], artificial neural networks (ANN) [15–17], bayesian classifiers (BC), classification and regression trees (CART) [18,19], K-nearest neighbors (KNN) among others [20]. Ensemble methods combine outputs of multiple independently trained weaker models to make an overall prediction. The selection of the combination of weaker learning methods is made in such a way as to maximize the prediction power of the ensemble algorithm. Ensemble methods include boosting, bootstrap aggregation (bagging), stacking/blending, random forests (RF) [16,21] and their modifications [22].

On the other hand, unsupervised learning involves the analysis of unlabeled (no distinction between input and output) data under assumptions about the structural properties of the data (e.g., algebraic, combinatorial, or probabilistic). Since there are no training examples used in this process, the learning algorithm aims to identify patterns and correlations in the given data. The main applications of these algorithms include clustering and dimensionality reduction. Dimensionality reduction algorithms, including principal components analysis, manifold learning, factor analysis, random projections, and autoencoders, identify and eliminate redundancies in the data so as to remain with only the variables that account for the most variability in the data. Clustering algorithms partition data into coherent clusters and determine the partitioning rule for predicting clusters in future data. The K-means clustering algorithm is the most commonly used method [20]. Computational complexity is a major concern in both clustering and dimension reduction since the datasets to exploit are large and unlabeled [2].

A third major ML paradigm is semi-supervised learning. Here, the data is a mixture of small amounts of labeled and large amounts of unlabeled training data. The algorithm learns the structures of the data from the labeled examples and makes assumptions about the unlabeled data in

order to make predictions. Semi-supervised learning is useful when the cost associated with labeling is too high to allow for a fully labeled training process [8]. Semi-supervised learning is subclassified into inductive learning whereby the goal is to learn from both the labeled and unlabeled dataset to predict labels for future datasets and transductive learning whereby the goal is to predict labels for the unlabeled portion of the data [23].

A more recent category of ML called reinforcement learning involves the algorithm discovering actions that yield the greatest rewards through trial and error. The algorithm is trained to choose actions that maximize reward. It is said to learn from past experiences and capture knowledge to make accurate decisions [24]. The most common example is the markov decision process [25].

In practice, current research and application blends unsupervised and supervised categories of ML. The choice of ML method to use is guided by the objectives of the analysis and the data available. Important data considerations include the number of predictor variables/features available in the data and quality of data. In general, a small but informative feature space results in higher generalizability of the model and avoids overfitting [26] while improving data quality and greatly improves the analysis. Therefore data preprocessing and feature selection, in often using the unsupervised learning methods, are often the initial steps in an ML analysis. The Feature selection process optimizes information gain and minimizes overfitting. Whereas some supervised learning algorithms (SVM) offer the advantage of sparsity and inherently select the most predictive features, explicit feature selection methods exist. These are broadly classified into filter methods (e.g. correlation coefficient scores, chi square tests, information gain), wrapper methods (e.g. step-wise covariate modeling, recursive feature elimination) and embedded methods (regularization algorithms, LASSO, elastic net and ridge regression) [27,28]. The extremely high dimensionality of biological data such as protein and peptide sequences exposes the inadequacy of the above feature selection methods during implementation of ML based approaches. Hence the need for methods that reduce the dimensionality of features so as to decrease computational running times while increasing classification accuracy. For this purpose, feature encoding techniques, which map original representations into new spaces, have been developed. This mapping makes separation of classes easier by condensing complex data patterns into fewer, easily manageable and statistically significant forms which makes the subsequent classification step easier and more accurate. A number of feature extraction methods have been developed and evaluated. These include the orthonormal encoding (OE), frequency based encoding (FE), Taylor's venn-diagram (TVD), residual couple encoding (RC) and a combination of OE and TVD [29,30].

Model training and validation may be done concurrently. A model is trained on one dataset (training set) and its prediction performance tested on another (test set). Prediction performance is a measure of how well a method is able to give correct prediction on unseen data. Model validation techniques include the holdout method, N-fold-cross-validation and bootstrap. With the holdout method, the data is conventionally split into a training (for model building) and test set (for performance evaluation) in a ratio of 2:1 respectively. In comparison, the N-fold-cross-validation method randomly splits the data into k subsets where the $k-1$ sets of the data are used to train the model while the k th set is used to assess the model's accuracy of prediction. The bootstrap method samples with replacement from the dataset to create n new datasets (bootstrap replicates). The replicates are used to test the model's predictive ability. In both N-fold-cross-validation and bootstrap, the accuracy measures are calculated as the average of all different validation cycles or bootstrap replicates [11,31].

Model performance is gauged using measures of accuracy e.g. root-mean-squared-error (RMSE), mean absolute error (MAR) and percentage prediction error (PPE) for regression with continuous numerical outcomes. With the classification of categorical outcomes, the percentage of correctly predicted observations, sensitivity and specificity, false

positive rate (FPR) and false negative rate (FNR), receiver operator characteristics (ROC) along with its area under the curve (AUROC) are the commonly utilized measures of performance [11,32,33].

A number of software have been developed and utilized for ML analyses. These include the statistical software R [34], Python, MATLAB and other dedicated ML software (WEKA, OpenAI, among others).

3. Survey methodology

The scope of this survey was limited to HIV/AIDS clinical research and medical care studies that utilized machine learning methodology. Clinical research was defined as any web-based, hospital based, social, clinical, imaging, genetic or molecular laboratory research involving the diagnosis and treatment of HIV/AIDS involving human participants or data from human participants. This survey aimed to establish the trend in application of ML techniques, the specific ML and clinical tasks involved, the most commonly used ML methods, their application paradigms and the appropriate data types and sources.

A focused search including all articles published online up to 11:59 p.m., 31st September 2017 and indexed in the PubMed biomedical database was conducted. PubMed database was chosen to limit the search to literature that is most likely to be accessed and applied by medical and health workers. The specific query search term used was '(((HIV patients) OR HIV-infected patients) OR HIV clinic) OR HIV medical records)) AND (((((((((((random forests) OR (classification and regression trees)) OR artificial neural networks) OR neural networks) OR support vector machines) OR SVM) OR bayes classifier) OR semi-supervised learning) OR supervised learning) OR unsupervised learning) OR machine learning)'. Studies in which HIV was neither the only nor the main disease of concern were excluded. Likewise, studies that used only the conventional statistical methods (e.g. *t*-test, chi-square, Cox regression) instead of ML methods in the analysis were excluded.

The search retrieved 115 publications out of which 39 were outside the scope of this survey (did not directly address HIV disease, or did not use ML methods, or did not directly involve human participants or data from human participants) and were thus eliminated from the survey. The remaining 76 publications were further categorized according to paradigm of ML application using a criterion adapted from Ref. [35] as follows:

1. Studies that review and summarize literature on the application and challenges of ML in HIV clinical research – (review).

2. Studies that utilized ML in the analysis of data to identify factors associated with outcomes of interest- (analysis).
3. Studies that demonstrate usable ML based tools and applications for clinical use to improve clinical care and research – (demonstration).
4. Studies that evaluate and use ML in the clinical diagnosis and management of HIV – (application)
5. Studies that present development of new ML methodologies, algorithms as well as new techniques improving on or enhancing already existing one – (development).
6. Studies that compare predictive performance of more than one ML techniques in HIV research – (comparison).

Additionally, the specific ML methods used the source and nature of data, the specific supervised learning task (Classification versus regression) and clinical objectives of each publication were extracted and summarized.

4. Key survey findings

An increase in the application of ML in published HIV clinical research from 1 publication in 1995 to 10 publications in 2016 was observed as shown in Fig. 1. While patient medical data was the primary source in earlier publications, a gradual diversification in data sources is observed, with genetic and imaging data becoming more common (Fig. 1). Overall, genetic data was the most common source (50%) followed by EMR (32.05%), imaging (5.13%), laboratory (3.85%) and other sources (16.67%). However, in most of the studies host or viral genetic profiles, clinical and social-economic variables and sometimes epidemiological variables were utilized in a complementary manner as model inputs.

The most commonly applied method was SVM (27.85%), followed by ANN (25.35%), LR (12.66%), CART (11.39%), RF (10.39%), LDA (2.53%) and other methods together (22.78%). A number of studies employed more than one method. Artificial neural networks and LR were the primary methods used in earlier publications but were gradually overtaken by SVM and RF together with custom made methods (Fig. 2). The most frequent ML task was classification (91.8%) followed by regression (6.3%) and unsupervised learning methods for dimension reduction (1.9%). The publication that utilized unsupervised methods for dimension reduction did not subsequently implement any supervised learning task [36].

Five of the publications were review articles discussing the

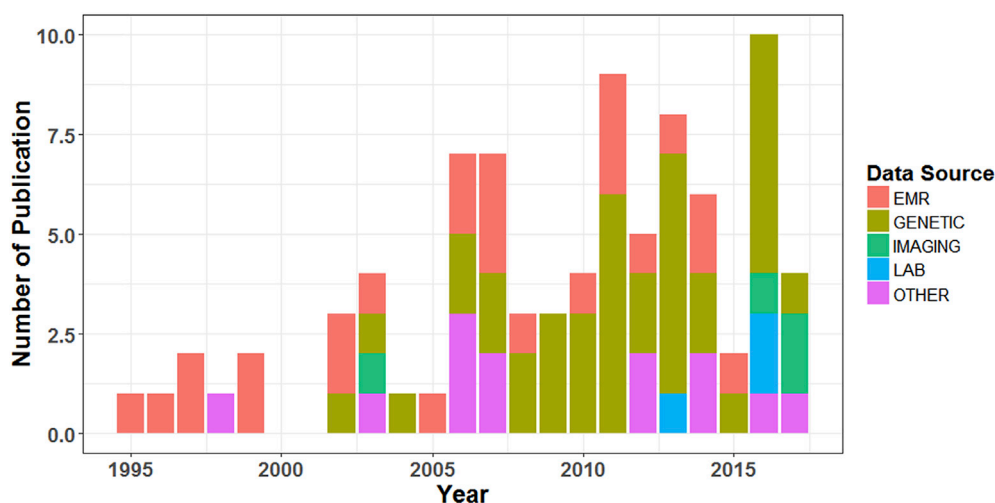


Fig. 1. Trend in PubMed indexed studies that employ ML techniques in HIV clinical research and care (n = 79). The different colors depict the different data sources used in the publications. EMR—electronic medical records (all clinical data about signs and symptoms of patients). IMAGING - radiological and other imaging data from patients. LAB - clinical laboratory data. GENETIC—human or pathogen genes, gene products, variants, phenotypes. OTHER—any other source of data including epidemiologic surveys, internet data, social data. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

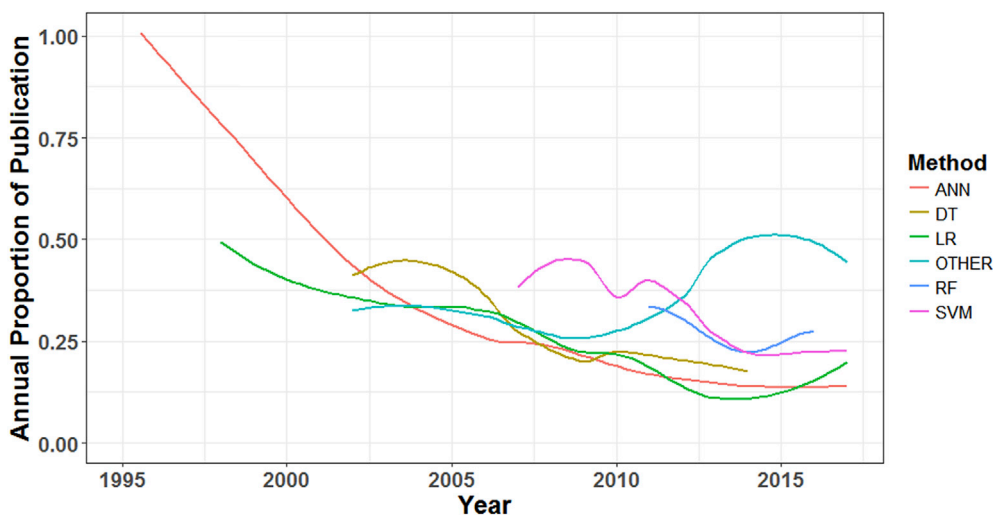


Fig. 2. Trend in ML methods employed in published HIV clinical research and care.

application of ML in optimization of HIV clinical research and treatment. The largest proportion of studies (41.77%) utilized ML methods for data analysis to identify predictor variables. Reports of clinical application of ML tools in form of web-servers, websites or software to assist scientists and clinicians in research and management of HIV patients were the least (7.59%). However, most of the studies fell in more than one category of ML application paradigm. More than 25% of the studies addressed resistance to ART. A gradual increase in development and clinical application of ML tools was observed. However, the volume of publications that utilized of ML for data analysis to identify predictive factors is still larger than that of publications using ML for clinical applications. (Fig. 3).

All studies employed the well-known evaluation techniques of data splitting during validation thus ensuring accuracy and reliability of the predictions made in these studies.

5. Discussion

Machine learning systems are expected to improve the speed and accuracy of diagnosis and clinical decision making among physicians thereby reducing costs, saving time and improving patients' health [5]. In this survey, we investigated the trend in application of ML to improve HIV clinical research and care.

Before the year 2000, published applications of ML focused on characterizing and predicting progression of HIV disease to AIDS (supplementary Table 1). However the studies utilized clinical data and application was limited to analysis and discovery of factors predicting disease progression. This choice of data and paradigm of ML mirrors the emphasis on prevention of opportunistic infections and AIDS progression in the pre- ART and pre-treatment guideline era of HIV management [37]. In addition, technology for genotyping was still very expensive and not wide spread.

The mass rollout of antiretroviral therapy (ART) in the early 2000s tremendously reduced mortality due to acquired immunity deficiency syndrome (AIDS) and improved quality of life. Unfortunately, the increased availability of ART coincided with emergency and spread of resistance to ART, currently between 5 and 10% and growing [38]. Resistance occurs when HIV virus undergoes mutations making it less susceptible to components of ART often resulting in virological failure. The slow pace of discovery of new effective anti-retroviral drugs has led to an increase in efforts to curb resistance such as resistance testing, optimization of choice of ART combination, improving adherence to ART and other interventions. Thus it is not surprising that an increase in application of ML to combat resistance to ART was observed in this survey.

As high-throughput sequencing technology became cheaper and more

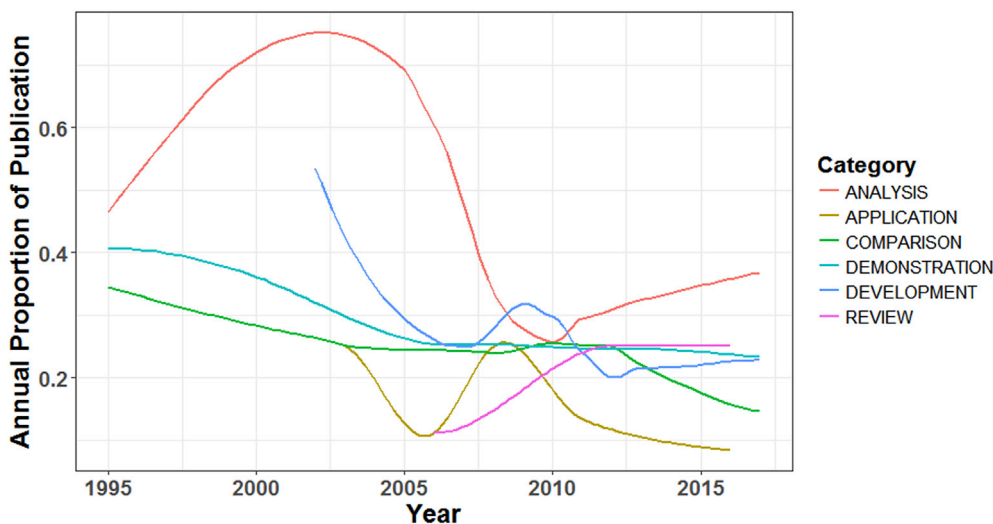


Fig. 3. Trend in patterns of application of ML in published HIV clinical research and care.

available in clinical settings, the use of genomic, proteomic, metabolomic and other such data with high information content has overtaken the use of clinical and other sources of data. Most data from EMRs are still extraordinarily messy, incomplete (having missing variables), incorrect (systematic random errors in the data), sparse (due to rare and unrepresentable patient records) and vague in terms of variable selection [4]. Indeed, some have described medical data as not being “big data” but “small data” arriving and changing quickly [39]. As observed in this survey, the solution to this problem has been to aggregate multiple types and sources of data during analysis so as to improve the predictive performance [40,41]. However, genetic data will likely remain the dominant source of ML data until such a time when aggregation of EMR data across health systems (or layers of health systems) to ensure completeness and sufficient variability is achieved. The recent advancement in imaging techniques such as high angular resolution diffusion imaging and resting state functional magnetic resonance imaging (MRI), that provide non-invasive yet fine-grained measures of biological structure and functionality will lead an increase in application of imaging data [46]. Coupled with genetic data this lead to an expansion of imaging genomics into HIV clinical research.

The availability of high information content data (genetic and imaging) could have necessitated the increased application of non-parametric methods (SVM, RF, DT) and ensemble method that require large amounts of data but result in higher predictive performance. The ever reducing cost of computing together with growing sizes of data will further buttress the use of such methods and warrant the development of more such methods. Just like in most medical problems, HIV clinical management mostly involves diagnosis or risk prediction. These are binary categorical tasks requiring classification machine learning techniques. Indeed the most commonly used techniques in HIV research were classification techniques.

The application of ML in the HIV related research is relatively young i.e. less than 30 years (Fig. 1) and medical practice is largely guided by evidence based practices whereby ideas and interventions are rigorously

tested to demonstrate benefit beyond mere chance, often through randomized controlled trials [42,43]. Therefore it is not surprising that most of the studies surveyed here focused on use of ML for discovery of predictive factors instead of direct practical clinical application. Machine learning is not about to replace human experts and as noted by Holzinger [44], interactive ML requiring a human in the loop is a necessity. It will however be important for clinicians to gain basic competence in the development and application of ML. Such competence should be built to combine both the quantitative modeling and the translational skills to bridge the discipline of ML with decision making [45]. An imbalance due to absence of either of the skills is likely to hinder the integration of ML into routine patient management. The human resource scarcity will have to be tackled in order to reap the benefits of ML in clinical management of HIV.

6. Conclusion and recommendations

The application of ML techniques in HIV clinical research and care was surveyed. A steady growth in the application of ML in HIV clinical research was observed. Despite the dominance of statistical analysis as an application paradigm, ML use is steadily diversifying to include practical clinical tools to assist in clinical decision making. The use of high dimension sources of data such as genomic data has overtaken other data sources and this has been attended by an increase in use of high performance non-parametric ML methods with a focus on combating resistance to ART. An improvement in clinical data collection and storage as well as ML training as a means to increase ML application in clinical care are recommended.

Acknowledgements

Part of this work is supported by Gilead through the Nelson Sewankambo scholarship program (O13F).

Appendix A. Supplementary data

Supplementary data related to this article can be found at <https://doi.org/10.1016/j.compbimed.2017.11.001>.

Author contributions

KRB: conceptualized the study, did the literature review, did the analysis, wrote the manuscript.

GTA - Read and edited the manuscript.

SAK - participated in the analysis, Read and edited the manuscript.

AK - conceptualized the study, Read and edited the manuscript.

BC - conceptualized the study, Read and edited the manuscript.

Conflict of interest

Non declared.

References

- [1] A. Agrawal, J. Gans, A. Goldfarb, The Simple Economics of Machine Intelligence, 2016 [cited 2017 18th/Jan/]; Available from: <https://hbr.org/2016/11/the-simple-economics-of-machine-intelligence>.
- [2] M. Jordan, T. Mitchell, Machine learning: trends, perspectives, and prospects, *Science* 349 (6245) (2015) 255–260.
- [3] S.L. Fultz, et al., Development and verification of a “virtual” cohort using the national VA health information system, *Med. Care* 44 (8 Suppl 2) (2006) S25–S30.
- [4] C. Paxton, A. Niculescu-Mizil, S. Saria, Developing predictive models using electronic medical records: challenges and pitfalls, *AMIA Annu. Symp. Proc.*, 2013, pp. 1109–1115, 2013
- [5] Z. Obermeyer, E.J. Emanuel, Predicting the future—big data, machine learning, and clinical medicine, *N. Engl. J. Med.* 375 (13) (2016) 1216.
- [6] R.C. Deo, Machine learning in medicine, *Circulation* 132 (20) (2015) 1920–1930.
- [7] WHO. HIV/AIDS fact sheet. July 2017 31-Sept-2017; Available from: <http://www.who.int/features/factfiles/hiv/en/>.
- [8] A.D. White, Complexity of human immunodeficiency virus management in developing countries, *Epidemiology* 9 (6) (1998) 593–595.
- [9] K. Kourou, et al., Machine learning applications in cancer prognosis and prediction, *Comput. Struct. Biotechnol. J.* 13 (2015) 8–17.
- [10] G.A. Barnard, T. B. Studies in the history of probability and statistics: IX. Thomas Bayes’s essay towards solving a problem in the doctrine of chances, *Biometrika* 45 (3/4) (1958) 293–315.
- [11] T. Hastie, R. Tibshirani, J. F. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, New York, 2009.
- [12] C. Shen, et al., Automated prediction of HIV drug resistance from genotype data, *BMC Bioinforma.* 17 (8) (2016) 278.
- [13] Y. Singh, M. Mars, Support vector machines to forecast changes in CD4 count of HIV-1 positive patients, *Sci. Res. Essays* 5 (17) (2010) 2384–2390.

- [14] M.H. Goldbaum, et al., Analysis with Support Vector Machine Shows HIV-positive Subjects without Infectious Retinitis Have MFERG Deficiencies Compared to Normal Eyes, vol 106, Transactions of the American Ophthalmological Society, 2008, p. 196.
- [15] Y. Singh, N. Narsai, M. Mars, Applying machine learning to predict patient-specific current CD 4 cell count in order to determine the progression of human immunodeficiency virus (HIV) infection, *Afr. J. Biotechnol.* 12 (23) (2013).
- [16] D. Wang, et al., A comparison of three computational modelling methods for the prediction of virological response to combination HIV therapy, *Artif. Intell. Med.* 47 (1) (2009) 63–74.
- [17] B. Larder, D. Wang, A. Revell, Application of artificial neural networks for decision support in medicine, *Artif. Neural Netw. Methods Appl.* (2009) 119–132.
- [18] Y. Li, B. Rapkin, Classification and regression tree uncovered hierarchy of psychosocial determinants underlying quality-of-life response shift in HIV/AIDS, *J. Clin. Epidemiol.* 62 (11) (2009) 1138–1147.
- [19] J.A. Muñoz-Moreno, et al., Classification models for neurocognitive impairment in HIV infection based on demographic and clinical variables, *PLoS One* 9 (9) (2014) e107625.
- [20] I. Choi, et al., Machine learning methods enable predictive modeling of antibody feature: function relationships in RV144 vaccinees, *PLoS Comput. Biol.* 11 (4) (2015) e1004185.
- [21] A.D. Revell, et al., The use of computational models to predict response to HIV therapy for clinical cases in Romania, *Germs* 2 (1) (2012) 6.
- [22] Dietterich, T.G. *Ensemble Methods in Machine Learning*. [cited 2016 30th/Aug/]; Available from: <http://www.cs.orst.edu/~tgd>.
- [23] X. Zhu, A.B. Goldberg, Introduction to semi-supervised learning, *Synthesis Lect. Artif. Intell. Mach. Learn.* 3 (1) (2009) 1–130.
- [24] V. Mnih, et al., Human-level control through deep reinforcement learning, *Nature* 518 (7540) (2015) 529–533.
- [25] C. Szepesvári, *Algorithms for Reinforcement Learning*, Morgan & Claypool Publishers, 2009.
- [26] Pang-ning Tan, Michael Steinbach, V. Kumar, *Introduction to Data Mining*, Pearson Education, Limited, 2014.
- [27] Y. Saeys, I. Inza, P. Larrañaga, A review of feature selection techniques in bioinformatics, *Bioinformatics* 23 (19) (2007) 2507–2517.
- [28] I. Guyon, A. Elisseeff, An introduction to variable and feature selection. *Journal of Machine Learning Research* 3 (Mar) (2003) 1157–1182.
- [29] M. Gök, A comparison of several feature encoding techniques for MHC class I binding prediction, *Int. J. Biosci. Biochem. Bioinforma.* 3 (2) (2013).
- [30] M. Gök, A.T. Özcerit, A new feature encoding scheme for HIV-1 protease cleavage site prediction, *Neural Comput. Appl.* 22 (7–8) (2013) 1757–1761.
- [31] B. Efron, *The Jackknife, the Bootstrap, and Other Resampling Plans*, Society for Industrial and Applied Mathematics, Philadelphia, Pa., 1982.
- [32] G. James, et al., *An Introduction to Statistical Learning*, vol 6, Springer, 2013.
- [33] N.R. Cook, Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve, *Clin. Chem.* 54 (1) (2008) 17–23.
- [34] R-Core-Team, *R: a Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2013.
- [35] E. Kolç, N. Frasheri, *A Literature Review of Data Mining Techniques Used in Healthcare Databases*, 2012.
- [36] J. Underwood, et al., Gray and white matter abnormalities in treated human immunodeficiency virus disease and their relationship to cognitive function, *Clin. Infect. Dis.* 65 (3) (2017) 422–432.
- [37] A.S. Fauci, The AIDS epidemic - considerations for the 21st century, *N. Engl. J. Med.* 341 (1999) 1045–1050.
- [38] K.M. Stadel, D. R. Rates of emergence of HIV drug resistance in resource -limited settings: a systematic review, *Antivir. Ther.* 18 (1) (2013) 115–123.
- [39] Sullivan, T. *Machine Learning: Changing Everything but Healthcare*. *Healthcare IT News* 2016 25th/October/2016, 12:16 [cited 2017 25th/January/]; Available from: <http://www.healthcareitnews.com/news/machine-learning-changing-everything-healthcare>.
- [40] M. Zazzi, et al., Predicting response to antiretroviral treatment by machine learning: the EuResist project, *Intervirology* 55 (2) (2012) 123–127.
- [41] M. Rosen-Zvi, et al., Selecting anti-HIV therapies based on a variety of genomic and clinical factors, *Bioinformatics* 24 (13) (2008) i399–406.
- [42] T. Greenhalgh, J. Howick, N. Maskrey, Evidence based medicine: a movement in crisis? *Bmj* 348 (2014) g3725.
- [43] E.-B.M.W. Group, Evidence-based medicine. A new approach to teaching the practice of medicine, *Jama* 268 (17) (1992) 2420.
- [44] A. Holzinger, Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Inf.* 3 (2) (2016) 119–131.
- [45] Dorian Pyle, C.S. Jose, *An Executive's Guide to Machine Learning* Mckinsey Quarterly, June 2015, 2015.
- [46] P.M. Thompson, et al., Imaging genomics, *Curr. Opin. Neurol.* 23 (4) (2010) 368–373.