ORIGINAL PAPER

Analysis and prediction of HIV and AIDS using machine learning: ensemble method

Minyechil Alehegn Tefera

Department of Electro-Optical Engineering, National Taipei University of Technology, Taipei 10608, Taiwan

Abstract

Introduction: At present, especially in developing countries, human immunodeficiency virus/acquired immune deficiency syndrome become very extreme. As we know, the virus break up and damages the function of immune system, and infected people gradually become immunodeficient. Artificial intelligence play a great role in the prediction of diseases. The function of immune system is the level of CD4+ cell count. In this study, ensemble method was applied.

Material and methods: Records used in the study were collected from world data available online. Random forest, naïve Bayes, J48, and KNN algorithms as well as the proposed ensemble method using bagging technique were employed to fit the rate data, with 10k cross-validation. The performance of the model was evaluated using accuracy, precision, sensitivity, and F-score.

Results: Based on the experiment, the proposed ensemble method achieved the best results in metrics of accuracy, precision, and F-score compared with other models by providing accuracy of 95.36%, precision of 90.23%, recall of 91.5%, and F-score of 92.35%.

Conclusions: Currently, machine learning plays a great role in the prediction of future, especially in health industry. The power of models is depending on the data. Based on the results of the study, the proposed ensemble method provided the best accuracy compared with other machine learning methods, with accuracy of 95.36% shown as very effective result.

HIV AIDS Rev 2025; 24, 4: 262-266
DOI: https://doi.org/10.5114/hivar/152465

Key words: ensemble method, prediction, HIV/AIDS, machine learning, bagging.

Introduction

Acquired immune deficiency syndrome (AIDS) is a severe medical condition caused by human immunodeficiency virus (HIV) infection that severely hampers functioning of the human immune system, making the human body more susceptible to minor infections [1]. In early stage of HIV disease (one week), the virus infects macrophages and enters the nerve center through the blood-brain barrier, causing chronic and long-term damage to nerve cells. Even with antiretroviral therapy used by HIV-infected patients, more than half of the patients still suffer from sensory, motoric, and neuro-cognitive impairments, namely HIV-related neuro-cognitive dysfunction syndrome. Currently, clinical diagnosis and treatment of HIV-related neurocognitive disorders (HAND) focus on the third stage. However, due to the irreversibility of neuronal damage, clinical treatment

Address for correspondence: Minyechil Alehegn, Department of Electro-Optical Engineering, National Taipei University of Technology, Taipei 10608, Taiwan, e-mail: minyechil21@gmail.com

Article history:
Received: 14.06.2022
Received in revised form: 25.07.2022
Accepted: 28.07.2022
Available online: 15.07.2025

at this stage lacks effective measures, leading to patients' disability or death. Global research and development on HIV and related infections has been carried out on a massive scale in the past few decades. Although HIV cannot be completely cured, researchers have developed certain drugs (antiretroviral drugs) over the past decade, which can control the multiplication of HIV in the human body, and thus prolong an infected individual's life. If a human delivers to the system tons of information, it understands it, and starts to respond [2]. AIDS is a medical condition caused by HIV, and is the major problem worldwide. The discovery of HIV as the causative organism of AIDS as well as the inability of modern medicine to find a cure for the disease, have placed HIV as one of the most dreaded pathogens of the 21st century. The expansion of the HIV/AIDS epidemic has now become a burning issue globally, especially in developing countries, such as sub-Saharan Africa. Today, HIV/AIDS is one of the largest public health crises endangering the human race. By the help of technology, data mining offers valued advantage in all healthcare research, e.g., diabetes, which leads to improvement and expanding of healthcare distribution, improvement of diseases supervision, and decision-making.

Related work

In Jiang *et al*. [1], the authors concluded that the nomogram is effective and accurate in predicting the survival of people living with HIV (PLHIV), and beneficial for medical workers in health administration. Jones *et al*. [3] reported that for PLHIV, an unique set of risk factors are relevant for prediction of preterm birth, including ART exposure, cervical length, cervico-vaginal biomarkers, vaginal microbiota, and imaging modalities. In Li *et al*. study [2], the researchers observed that the prediction model might be practical and easily applied to recognize HIV/AIDS individuals who are most likely to benefit from modern antiretroviral therapy. McKittrick *et al*. [4] explored the role of monocyte/macrophages in cardiovascular disease (CVD) pathogenesis, with some studies examining functional assays as better predictors of CVD risk. In their study, Tunc *et al*. [5] applied artificial neural network, while in Dong *et al*. [6], a total of 547 respondents were used. The impact of psycho-social factors on increasing the risk of HIV infection among men who have sex with men (MSM) has attracted researchers' attention. Belete *et al*. [7] used grid search for hyper-parameter optimization (GSHPO) on the considered models to strengthen the prediction power, while Zheng *et al*.'s [8] model achieved the highest accuracy/recall, with an average improvement of 38.5% over the other baseline models. Turbé *et al*. [9] demonstrated that deep learning algorithms showed an increasing promise for disease diagnosis; however, their use with rapid diagnostic tests performed was not extensively tested.

In the current study, deep learning was employed to classify images of rapid HIV tests acquired in rural South Africa.

Steiner *et al*. [10] identified convolutional neural networks as the best performing architecture, and demonstrated a connection between the importance of biologically relevant features in the classifier and the overall performance. Hu *et al*. [11] developed an attention-based deep learning framework, named DeepHINT. In Olatosi *et al*. [12] study, convolutional neutral network (CNN) was used, while in Lu *et al*. [13], CNN, RNN, and LSTM were applied. In Luckett *et al*. [14], deep neural network (DNN) was employed, and provided 86% of accuracy, whereas Kaku *et al*. [15] demonstrated that deep learning method is very important in the prediction of HIV/AIDS. In Xiang *et al*. [16] study, modern methods of computer-aided drug design significantly expanded the capabilities of pharmaceutical industry. Sun *et al*. study [17] showed that rough set-based algorithms are capable of dealing with the gaps and imperfections present in real-time data.

In recent years, a large number of studies have appeared on the use of machine learning methods to predict potential HIV-1 inhibitors and virus resistance to anti-HIV drugs. In Oliveira *et al*. study [18], based on the researchers experiment using multilayer artificial neural networks, *k*-nearest neighbor algorithm, support vector machines, and naïve Bayesian classifiers, with MLP demonstrating superior results. Lu *et al*. [13] applied support vector machine, random forest, and LSTM. Moreover, Steiner *et al*. [10] used machine learning in investigating HIV drug resistance, and more broadly developed a framework with many important applications in viral genomics. In Wang *et al*. study [19], four models were applied, such as LSTM, NN, ARIMA, and GRNN, proving superiority of LSTM model. Whereas Shi *et al*. [20] applied XGBoost and machine learning-based tool to predict the risk of death.

Material and methods

Data collection

The description of the attribute is shown in Table 1. Dataset used in this study were obtained from publicly available world data [21].

Data pre-processing

In this study, PCA was applied, while in data pre-processing, data conversion (data cleaning, eliminating of missing data) was used.

Feature selection

In the scientific world, identifying the position of features is very important to obtain an efficient result in every aspect. In the current study, nine attributes were included. For feature selection to identify the importance of a feature, PCA was applied.

Table 1. Attribute description

Attribute name	Description
Sex	Sex of the person
Marital status	Marital status of HIV-positive patient
Place of residence	Place of the patient's residence
Religion	Religion of the patient
Age	Age of the patient
TB co-infection	Yes/No
Medication	Medication given to the patient
Illness	Illness/no illness
CD4+ count level	CD4+ count/ mm^3

Models for proposed work

In this paper, random forest (RF), *k*-nearest neighbor (KNN), naïve Bayes (NB), and J48 algorithms as well as the proposed ensemble method were employed.

KNN (*k*-nearest neighbor)

It is grouped under the category of lazy prediction technique, which is easy and simple. This technique classifies new work based on similarity measure. The training data are sorted in this algorithm, and based on the nearest prediction of the test data, it is completed.

$$\text{Euclidean} = \sqrt{\left(\sum_{i=1}^{k}(X - y)^2\right)} \quad (1)$$

The value of variable *k* is always +ve integer (positive integer). When the *k*-value is large, it is more accurate. In most of the cases, the value of variable *k* ranges between 3 and 10.

Results and discussion

Performance metrics

$$\text{Acuracy} = 100 \times \left(\frac{TP + TN}{TP + FP + FN + TN}\right) \quad (2)$$

$$\text{Precision} = 100 \times \left(\frac{TP}{TP + FP}\right) \quad (3)$$

$$\text{Sensitivity} = 100 \times \left(\frac{TP}{TP + FP}\right) \quad (4)$$

$$\text{F-score} = 100 \times \left(\frac{2 \times (\text{recall} \times \text{precision}}{\text{recall} + \text{precision}}\right) \quad (5)$$

Discussion

In Table 1, the attributes of the data are described. Long-lasting experimental and computational efforts on HIV integration produced results shown in Table 2 and Figure 2. While Figure 1 indicates the proposed work flow, and Figure 2 display the graph of the results of models' performance. Based on the experiment, the comparison of the results is shown in Table 2 and Figure 2. J48 algorithm pro-

Algorithm for NB

Input:

Process:

Learning Algorithm
Train

Test $p(C|x) = \frac{P(X|C)P(C)}{p(x)}$

Output:

Algorithm For Random Forest

for i = 1 to c **do**
  Randomly sample the training data D with replacement to produ
  Create a root node, N_i containing D_i
  Call BuildTree(N_i)
end for

BuildTree(N):
if N contains instances of only one class **then**
  return
else
  Randomly select x% of the possible splitting features in N
  Select the feature F with the highest information gain to split on
  Create f child nodes of N, N_1,..., N_f, where F has f possibl
  for i = 1 to f **do**
    Set the contents of N_i to D_i, where D_i is all instances in N tha
    F_i
    Call BuildTree(N_i)
  end for
end if

Algorithm for Proposed Ensemble method

Bagging (forecast algorithm A, datasets Da, repetitions r)

a. model generation

For i= 1 to r:

Generate a bootstrap sample Da (i) from Da

Let M (i) be result of training A on Da (i)

b. prediction for a given test instance x

For i = 1 to r:

Let C (i) = output of M (i) on x

Return class that appears most often among C (1)...C (r)

Table 2. Performance of models

Model	Accuracy	Precision	Sensitivity	*F*-score
J48	88.98	83.01	85.76	83.23
Random forest	90.03	77.97	83.45	81.05
Nave Bayes	85.44	76.68	86.12	79.98
k-nearest neighbor	82.02	75.02	77.04	77.52
Proposed ensemble using bagging (PEUB)	95.36	90.23	91.50	92.35

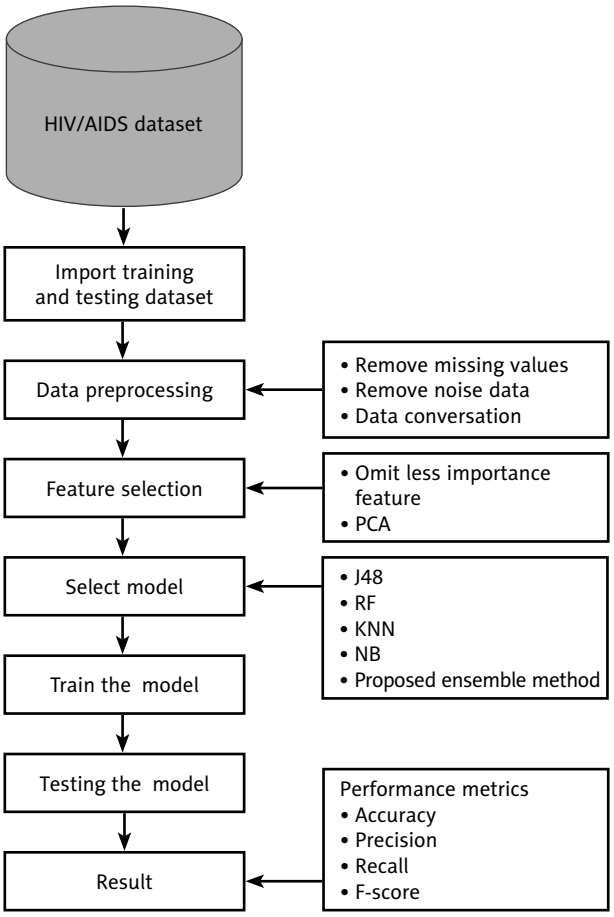

Figure 1. Architecture of the proposed work flow

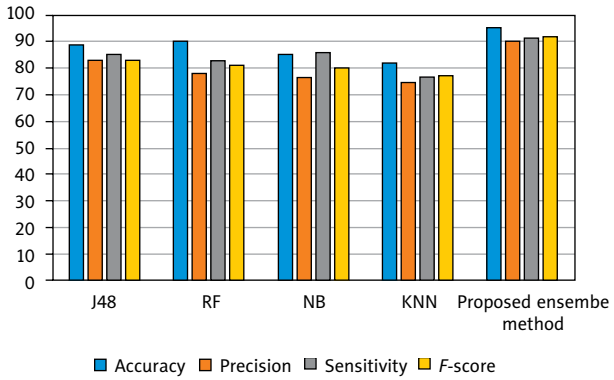

Figure 2. Performance models

vided accuracy of 88.98%, precision of 83.01%, sensitivity of 85.76%, and F-score of 83.23%. RF demonstrated 90.03% of accuracy, 77.97% of precision, 83.45% of sensitivity, and 81.05% of F-score. Naïve Bayes delivered accuracy of 85.44%, precision of 76.68%, sensitivity of 86.12%, and F-score of 79.98%. KNN provided accuracy of 82.02%, precision of 75.02%, sensitivity of 77.04%, and F-score of 77.52%. Finally, the proposed ensemble using bagging (PEUB) showed the best results compared with other models, providing 95.36% of accuracy, 90.23% of precision, 91.5% of sensitivity, and 92.35% of F-score.

Conclusions

Finding the prediction of something based on the current entity, is satisfying but very difficult work. A model has the power to forecast the future in health sector. Machine learning methods are very effective to predict the future based on a given data. Here, in machine learning models, the proposed ensemble method achieved the best results. Based on the study, the proposed ensemble using bagging method provided the highest accuracy of 95.23%, with lower error of 4.77%.

Disclosures

1. Institutional review board statement: Not applicable.
2. Assistance with the article: None.
3. Financial support and sponsorship: None.
4. Conflicts of interest: None.

References

1. Jiang F, Xu Y, Liu L, Wang K, Wang L, et al. Construction and validation of a prognostic nomogram for predicting the survival of HIV/AIDS adults who received antiretroviral therapy: a cohort between 2003 and 2019 in Nanjing. BMC Public Health 2022; 22: 30. DOI: 10.1186/s12889-021-12249-8.
2. Li B, Zhang L, Liu Y, Xiao J, Li C, Fan L, et al. A novel prediction model to evaluate the probability of CD4/CD8 ratio restoration in HIV-infected individuals. AIDS 2022; 36: 795-804.
3. Jones AJ, Eke UA, Eke AC. Prediction and prevention of preterm birth in pregnant women living with HIV on antiretroviral therapy. Expert Rev Anti Infect Ther 2022; 20: 837-848.

4. McKettrick P, Mallon PWG. Biomarkers to predict cardiovascular disease in people living with HIV. Curr Opin Infect Dis 2022; 35: 15-20.
5. Tunc H, Durdagi S, Sari M, Kotil S. ANN-based Drug-isolate-fold-change model predicting the resistance profiles of HIV-1 protease inhibitors. ChemRxiv 2022. DOI: 10.26434/chemrxiv-2022-h5krl.
6. Dong Y, Liu S, Xia D, Xu C, Yu X, Chen H, Wang R, et al. Prediction model for the risk of HIV Infection among MSM in China: validation and stability. Int J Environ Res Public Health 2022; 19: 1010. DOI: 10.3390/ijerph19021010.
7. Belete DM, Huchaiah MD. Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results. Int J Comput Applications 2021; 44: 875-886.
8. Zheng C, Wang W, Young SD. Identifying HIV-related digital social influencers using an iterative deep learning approach. AIDS 2021; 35 (Suppl 1): S85-S89. DOI: 10.1097/QAD.0000000000002841.
9. Turbé V, Herbst C, Mngomezulu T, Meshkinfamfard S, Dlamini N, Mhlongo T, et al. Deep learning of HIV field-based rapid tests. Nat Med 2021; 27: 1165-1170.
10. Steiner MC, Gibson KM, Crandall KA. Drug resistance prediction using deep learning techniques on HIV-1 sequence data. Viruses 2020; 12: 560. DOI: 10.3390/v12050560.
11. Hu H, Xiao A, Zhang S, Li Y, Shi X, Jiang T, et al. DeepHINT: understanding HIV-1 integration via deep learning with attention. Bioinformatics 2019; 35: 1660-1667.
12. Olatosi B, Vermund SH, Li X. Power of Big Data in ending HIV. AIDS 2021; 35 (Suppl 1): S1-S5. DOI: 10.1097/QAD.0000000000002888.
13. Lu X, Wang L, Jiang Z. The Application of Deep Learning in the Prediction of HIV-1 Protease Cleavage Site. In: 2018 5th International Conference on Systems and Informatics (ICSAI). IEEE, 2018; pp. 1299-1304.
14. Luckett P, Paul RH, Navid J, Cooley SA, Wisch JK, Boerwinkle A, et al. Deep learning analysis of cerebral blood flow to identify cognitive impairment and frailty in persons living with HIV. J Acquir Immune Defic Syndr (1999) 2019; 82: 496-502.
15. Kaku Y, Kuwata T, Gorny MK, Matsushita S. Prediction of contact residues in anti-HIV neutralizing antibody by deep learning. Japanese J Infect Dis 2020; 73: 235-241.
16. Xiang Y, Du J, Fujimoto K, Li F, Schneider J, Tao C. Application of artificial intelligence and machine learning for HIV prevention interventions. Lancet HIV 2022; 9: e54-e62. DOI: 10.1016/S2352-3018(21)00247-2.
17. Sun D, Peng Y, Li H. Construction of Knowledge Graph of HIV-associated Neurocognitive Disorders Syndrome based on Deep Learning. In: 2020 International Conference on Artificial Intelligence and Computer Engineering (ICAICE). IEEE, 2020; pp. 134-141.
18. Oliveira A, Faria BM, Gaio AR, Reis LP. Data mining in HIV-AIDS surveillance system. J Med Systems 2017; 41: 51. DOI: 10.1007/s10916-017-0697-4.
19. Wang G, Wei W, Jiang J, Ning C, Chen H, Huang J, et al. Application of a long short-term memory neural network: a burgeoning method of deep learning in forecasting HIV incidence in Guangxi, China. Epidemiol Infect 2019; 147: e194. DOI: 10.1017/S095026881900075X.
20. Shi M, Lin J, Wei W, Qin Y, Meng S, Chen X, et al. Machine learningbased in-hospital mortality prediction of HIV/AIDS patients with Talaromyces marneffei infection in Guangxi, China. PLoS Negl Trop Dis 2022; 16: e0010388. DOI: 10.1371/journal.pntd.0010388.
21. https://data.world/.