ORIGINAL PAPER

Application of machine learning and deep learning for the prediction of HIV/AIDS

Minyechil Alehegn

Mizan Tepi University, Tepi, Ethiopia

Abstract

Introduction: Nowadays human immunodeficiency virus (HIV)/acquired immunodeficiency syndrome (AIDS) is very dangerous. HIV targets the resistant system and weakens people's denial against many contaminations and some kinds of cancer. As the virus breaks up and impairs the function of immunity, infected people gradually become immunodeficient. Both deep learning and machine learning models play a great role in the prediction of diseases. The function of immunity is CD4 cell count. In this study both the machine learning and deep learning algorithms were applied.

Material and methods: In this paper the data were collected from data world. Support vector machine, random forest, naïve Bayes, gated recurrent unit, and long short-term memory were used to fit the incidence data. The performance of the model was evaluated by accuracy, precision, sensitivity, and F-score with respective errors.

Results: Based on the evaluation, deep learning models achieve better results in the metrics of accuracy, precision, and F-score than machine learning models. But in sensitivity metrics machine learning models achieve better result than deep learning. Machine learning algorithms SVM, RF, and NB provide accuracy of 89.00%, 87.00%, and 86.94%; precision of 75.89%, 74.97%, and 75.78%; sensitivity of 87.96%, 84.00%, and 84.12%; F-score of 82.87%, 80.03%, and 79.05%, respectively. LSTM, GRU provides accuracy of 97.65%, 96.00%, precision of 77.35%, 84.00%, sensitivity of 87.93%, 82.98%, and F-score of 82.03%, 83.20%, respectively.

Conclusions: The possibility survival of the illness is less than no illness. The existence of TB negative is higher than TB positive. In the machine learning model SVM provides better sensitivity with 87.96%, long short-term memory provides accuracy of 97.65%, precision of 77.35%, sensitivity of 87.93%, and F-score of 82.03%.

HIV AIDS Rev 2022; 21, 1: 17-23
DOI: https://doi.org/10.5114/hivar.2022.112852

Key words: deep learning, HIV/AIDS, machine learning, prediction.

Introduction

Human immunodeficiency virus (HIV)/acquired immunodeficiency syndrome (AIDS) is a community healthiness delinquent in the world. The World Health Organization (WHO) reported that more than 33 million people live with HIV/AIDS. In the report the great majority (85%) of pregnant and breastfeeding women living with HIV also received antiretroviral therapy (ART), which not only protects their health, but also ensures prevention of HIV transmission to their newborns. In addition to that, the WHO said that the number of new people starting treatment is far below expectation due to the reduction in HIV-testing and treat-

Address for correspondence: Minyechil Alehegn, Mizan Tepi University, Tepi, Ethiopia, e-mail: minyechil21@gmail.com

Article history:
Received: 05.06.2021
Received in revised form: 22.06.2021
Accepted: 22.06.2021
Available online: 15.01.2022

ment initiation and ARV disruptions that occurred during the COVID-19 pandemic. By the end of 2020, testing and treatment rates showed steady but variable recovery. Today it is one of the largest public health crises endangering the human race. In almost 3 decades since its first cases were recognized, it has claimed the lives of millions of people, making it one of the most devastating epidemics. Deep learning and machine learning models play an inordinate role in the prediction of sicknesses in health industries to save human life early. Deep learning is a chunk of machine learning which is defined as the knowledge representation of data. If I human delivers the system tons of information, it starts to understand it and respond. AIDS is a medical condition caused by the HIV and is a major problem worldwide. The discovery of the HIV as the causative organism of AIDS and the inability of modern medicine to find a cure for it have placed HIV as one of the most dreaded pathogens of the 21st century. The expansion of the epidemic has now become a burning issue globally, and this is particularly so in developing countries; especially in sub-Saharan Africa.

Material and methods

Related work

Yang *et al*. [1]: In this paper a CPM model was built and successfully predicted individual differences in global cognitive performance. Singh *et al*. [2]: Two machine learning methods were applied: SVM and neural network. Yin *et al*. [3]: DT, RF, and AdaBoost with DT (AdaBoost) were applied. AdaBoost showed the highest accuracy = 92.8%, precision = 91.5%, recall = 94.4%, $F - 1 = 93.0\%$, and AUC = 96%. The detection value of each variable was calculated using the optimal machine learning algorithm. Bisaso *et al*. [4]: Three models were used to multitask temporal logistic regression, patient-specific survival prediction, and simple logistic regression. Multitask logistic regression provided better results than other models. Bonet *et al*. [5]: Recurrent neural network models provided better accuracy of 94%. Singh [6]: Developing countries are characterized by poor infrastructure and limited resources. Dubey [7]: svm, DT, ensemble methods, and naive base classifier were applied; svm shows good result. Young *et al*. [8]: Machine learning can enable social big data to become a new and important tool in HIV research, helping to create a new field of "digital HIV epidemiology". Gelaw *et al*. [9]: The proportion of the population who were migrants or who had a low educational status was associated with a high risk of infection. Ramachandran *et al*. [10] concluded that retention in care is crucial for individual and public health. Two techniques were used: random forest and LR. Shen *et al*. [11]: Machine learning using a unified encoding of sequence and protein structure as a feature vector provides an accurate prediction of drug resistance. Marcus *et al*. [12]: Machine learning has strong ability to improve delivery of PrEP. Singh and Su [13]: Experiment results showed that combinations of sequence, structure, and physicochemical features performed better than single feature type for identification of HIV-1 protease cleavage sites. Tu *et al*. [14]: Application of machine learning methods to dissect the different variables. Singh *et al*. [15]: svm machine learning model was applied for the prediction of HIV/AIDS. Cole *et al*. [16]: Increased apparent brain aging, predicted using neuroimaging, was observed in HIV-positive adults, despite effective viral suppression. Ahlström *et al*. [17] identified that machine learning techniques can learn from nation-wide electronic registry data and help to identify undiagnosed PLWH with a fairly high level of accuracy. Zazzi *et al*. [18] concluded that high-quality training data play a role in predicting purpose. Nan and Gao [19]: Three criteria of forecasting performance, MAPE, RMSPE, and IA, all indicate that the MLP model of ANNs can result in accurate forecasting of concurrent AIDS incidences and deaths with Baidu search trend data. Lu *et al*. [20]: Support vector machine, random forest, and LSTM were applied. Moreover, Steiner *et al*. [21] used machine learning in studying HIV drug resistance, and developed a framework that has many important applications in viral genomics more broadly. Wang *et al*. [22]: Four models were applied: LSTM, NN, ARIMA, and GRNN – LSTM was the best. Rajpurkar *et al*. [23]: tuberculosis (TB) is the leading cause of preventable death in HIV/AIDS-positive patients. Mayr *et al*. [24]: Deep learning excelled in toxicity prediction. Li *et al*. [25]: BPNN, RNN, LSTM were applied. The performance of the MHP-SO-GRU network model was the best. Madigan *et al*. [26]: The main factors in understanding HIV/AIDS prevalence rates are physician density followed by female literacy rates and nursing density. Oliveira *et al*. [27]: Multilayer artificial neural networks, k-nearest neighbour algorithm, support vector machines, and naive Bayesian classifiers. MLP showed better results. Hajipour *et al*. [28]: Infant mortality is the consequence of a variety of factors, including factors related to infants themselves and their mothers and events during pregnancy.

I use both traditional and modern approaches, i.e. machine learning and deep learning models for scalable and accurate prediction of HIV/AIDS, which is effective in both time and cost.

I use a flexible universal, learnable framework for representing world, visual, and linguistic information.

I use the model that can learn both unsupervised and supervised.

In this proposed work there is the concept of flexibility to use a pivot date for training/testing; you can start training/prediction from any date of choice.

Data collection

A description of the attributes is shown in Table 1. The dataset used in this paper was obtained from public available world data [29].

Data preprocessing

In this work, Principal Component analysis was applied. In the dataset preprocessing data conversion (CSV to Arff, Excel to CSV, etc.), data cleaning, and elimination of missing data are included in data preprocessing.

Feature selection

In the investigative world, identifying the position of features is very important, to get an efficient result in every aspect. In this work 9 attributes are included. For feature selection to identify the importance of a feature in this PCA was applied.

Predictive models for proposed work

Deep learning models

LSTM

It is a superior gentle of RNN is accomplished in learning LT dependencies from the situation. This model is designed to avoid long-term dependency problems.

The 3 gates are computed as follows:

$$f_{(t)} = \sigma\,(W_{fx}X_{(t)} + Wf_h h_{(t-1)} + b_f) \quad (1)$$

$$i_{(t)} = \sigma\,(W_{ix}X_{(t)} + Wi_h h_{(t-1)} + b_i) \quad (2)$$

$$o_{(t)} = \sigma\,(W_{ox}X_{(t)} + Wo_h h_{(t-1)} + b_o) \quad (3)$$

where σ is a nonlinear activation function. Most of the time the sigmoid function can be used as an activation function for gates. A sigmoid layer decides what parts of the cell state will be output. Inside LSTM intermediate state $C_{(t)}$ can be generated as follow the input gates decide the updated value.

$$C_{(t)} = tanh\,(W_{cx}X_{(t)} + W_{ch}h_{(t-1)} + b_c) \quad (4)$$

A tanh layer creates a vector of new candidates. Then the memory cell and hidden state of LSTM are updated as follows:

$$C_{(t)} = f_{(t)}\,\Theta\,C_{(t)} + i_{(t)}\,\Theta\,C_{(t)} \quad (5)$$

$$h_{(t)} = o_{(t)}\,\Theta\,tanh\,C_{(t)} \quad (6)$$

where *tanh* = the non-linear tanh activation function, Θ is the pointwise that was used to denote multiplication operation for 2 vectors.

GRU

GRU is the novice generation of RNN. GRU is like LSTM except it is simple to compute and implement. The information to pitch away and the new information to add decided by this gate. The expanse of past information to Foregate is decided by reset gate. GRU as usual has the input and output layers.

$$z_{(t)} = \sigma\,(W_z\,[h_{(t-1)}, x_t]) \quad (7)$$

$$r_{(t)} = \sigma\,(W_r\,[h_{(t-1)}, x_t]) \quad (8)$$

$$h'_{(t)} = tanh\,(W\,[r_{(t)} \times h_{t-1}, x_t]) \quad (9)$$

$$h_{(t)} = (1 - z_{(t)}) \times [h_{(t-1)} + z_{(t)} \times h'_{(t)}] \quad (10)$$

where $z_{(t)}$ is the updated gate, $r_{(t)}$ represents the reset gate, and $h'_{(t)}$ represents new memory.

Machine learning models

Algorithm for SVM

Input:

Process:

Identification of right hyperplane

Exploiting the spaces between neighbor data point

Adding a feature $Z = X^2 + Y^2$.

Output:

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 produce D_1

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 possible values (F_1, ..., F_f) **for** i = 1 to f **do**

Set the contents of N_i to D_i, where D_i is all instances in N that match F_i

Call BuildTree (N_i)

end for

end if

Table 1. Attribute description

Attribute name	Description
Sex	Sex of the person
Marital status	Marital status of the HIV+
Place of residence	The place of the patient
Religion	Religion of the patient
Age	Age of the patient
TB coinfected	Check whether the patient is coinfected or not
Medication	Medication to give the patient
Illness	Illness/No illness
CD4	CD4 in count/mm^3

Discussion

The attributes of the data are described in Table 1. In this paper the obtained results from the experiment are shown in Figure 1 and Table 2.

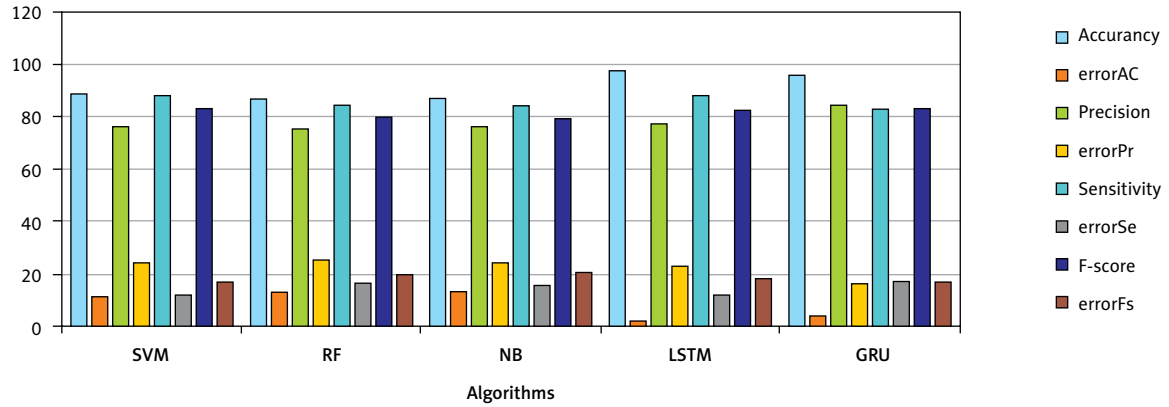

Figure 1. Performance of machine learning and deep learning model

Table 2. Performance of both machine learning and deep learning model

Model	Accuracy	Error	Precision	Error	Sensitivity	Error	F-score	Error
SVM	89.00	11.00	75.89	24.11	87.96	12.04	82.87	17.13
RF	87.00	13.00	74.97	25.03	84.00	16.00	80.03	19.97
NB	86.94	13.06	75.78	24.22	84.12	15.88	79.05	20.95
LSTM	97.65	2.35	77.35	22.65	87.93	12.07	82.03	17.97
GRU	96.00	4.00	84.00	16.00	82.98	17.02	83.20	16.80

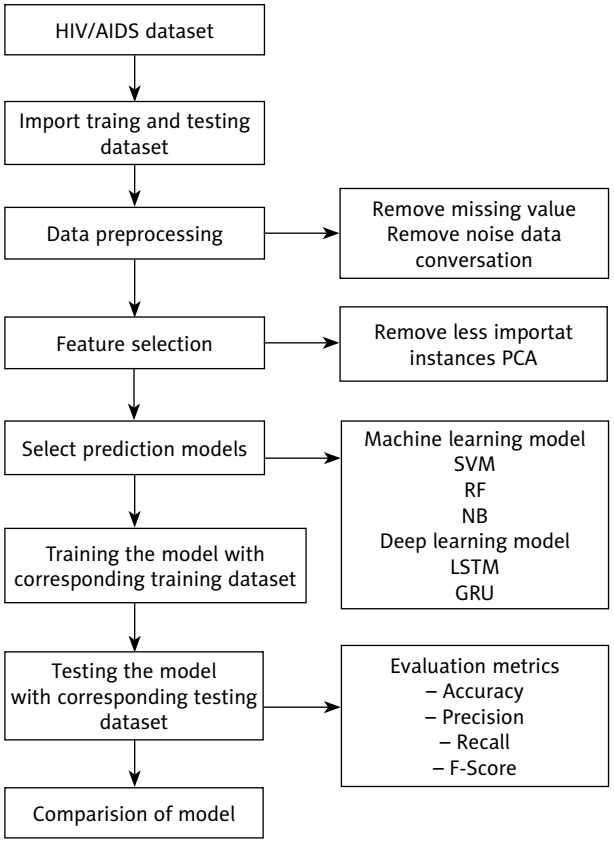

Figure 2. Architecture of proposed workflow

Figure 2 shows the proposed work flow.

Figure 3 shows the structure of LSTM, and Figure 4 shows the structure of GRU.

Figure 5 shows the data dependency of the machine learning and deep learning algorithms studied. I compared the results of the machine learning model and deep learning model.

Figure 6 shows that the loss over epoch machine learning algorithms (SVM, RF, NB) provides accuracy of 89.00%, 87.00%, and 86.94%, precision of 75.89%, 74.97%, and 75.78%, sensitivity of 87.96%, 84.00%, and 84.12%, and F-score of 82.87%, 80.03%, and 79.05%, respectively. In machine learning models SVM achieves better accuracy. In deep learning models the model achieves better results than machine models. LSTM and GRU provide accuracy of 97.65% and 96.00%, precision of 77.35% and 84.00%, sensitivity of 87.93% and 82.98%, and F-score of 82.03% and 83.20%, respectively. Based on the experiment results, the deep learning model shows better results than old models.

As Figure 7 shows, possible survival of the illness is less than no illness.

The existence of TB negative is higher than TB positive, as shown in Figure 8.

Conclusions

Finding the forecasting of something based on the current thing is satisfactory but difficult duty. An algorithm

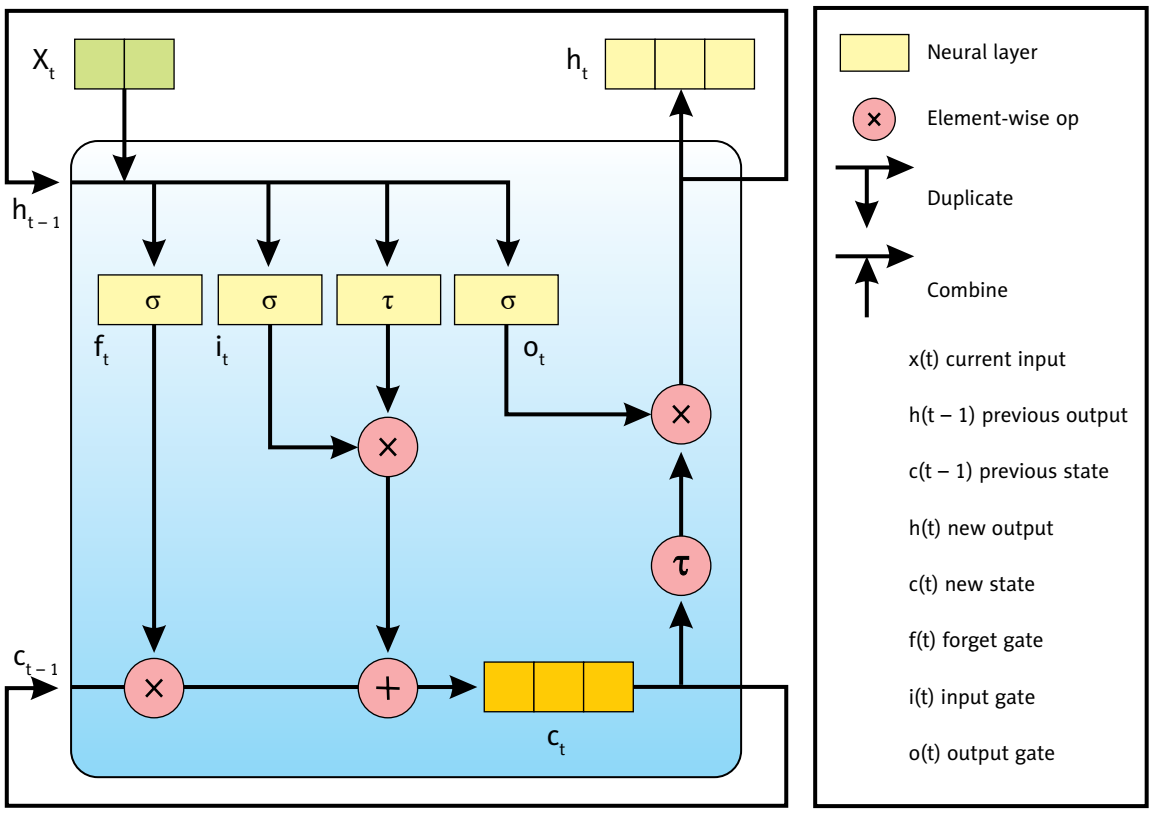

Figure 3. Structure of LSTM

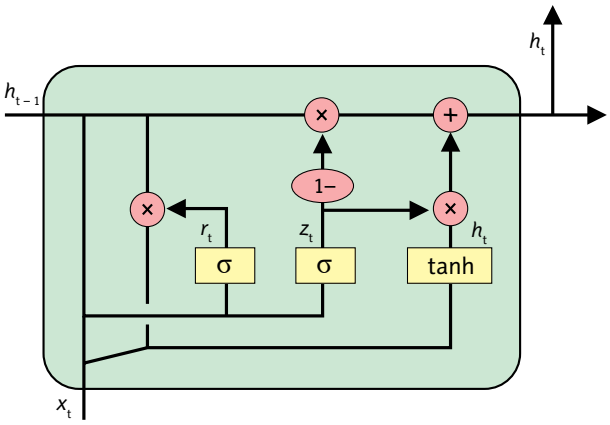

Figure 4. Structure of GRU

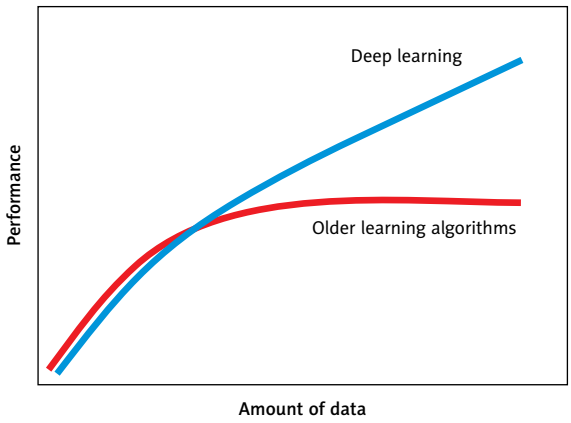

Figure 5. Data dependency of machine learning and deep learning algorithms

has the poIr to the prediction of the future in the health industry. Deep learning models are very influential for forecasting the future based on existing given data in the health industry. In machine learning models SVM achieves better accuracy of 89.00%. Based on the experiment results, the deep learning model showed better results than machine learning models by providing accuracy of 97.65% using LSTM. The possibility survival of the illness is less than no illness. The existence of TB negative is higher than TB positive. The proposed deep learning model LSTM provides the highest accuracy of 97.65% and low error of 2.35%.

Ethical approval

Permission to undertake this research was obtained from the university teaching hospital as primary data and use secondary data which is available online for researchers. Informed verbal consent was also obtained from study

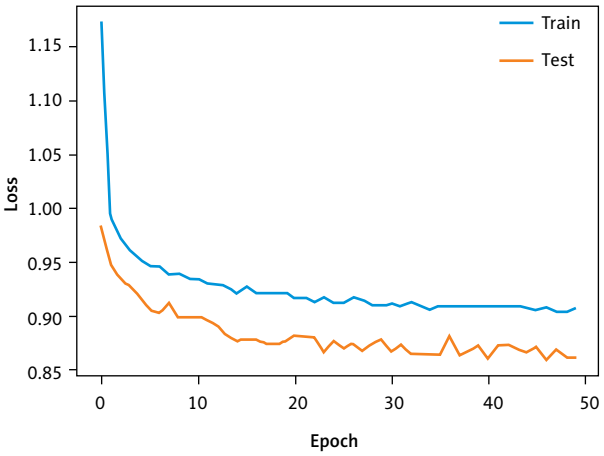

Figure 6. Loss over epoch

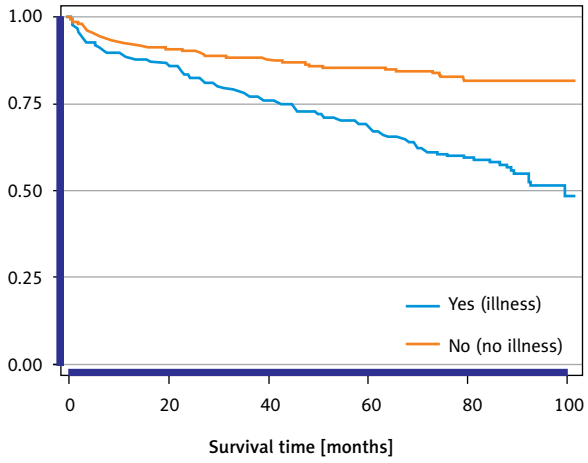

Figure 7. The possibility of subsist estimates by illness

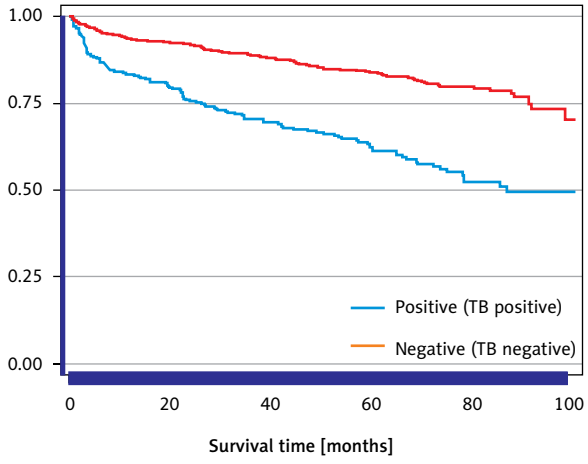

Figure 8. The possibility of subsist estimated by tuberculosis (TB)

participants in their local language after explaining the purpose of the study, the benefits of participating in the study, and the right to withdraw from the study at any time during data collection. The participants were also assured the confidentiality of their responses because their names were not included in the questionnaire.

Acknowledgements

The author would like to give heartfelt gratitude and appreciation to Mizan-Tepi University Research Directorate Office and Tepi Campus Research Coordinator Office and Department of Information Technology for granting logistical support to this study.

Conflict of interest

The author declares no conflict of interest.

References

1. Yang FN, Hassanzadeh-Behbahani S, Bronshteyn M, et al. Connectome-based prediction of global cognitive performance in people with HIV. Neuroimage Clin 2021; 30: 102677.
2. Singh Y, Narsai N, Mars M. 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 2013; 12: 3724-3730.
3. Yin Y, Xue M, Shi L, et al. A noninvasive prediction model for hepatitis B virus disease in patients with HIV: based on the population of Jiangsu, China. Biomed Res Int 2021; 2021:6696041.
4. Bisaso KR, Karungi SA, Kiragga A, Mukonzo JK, Castelnuovo B. A comparative study of logistic regression based machine learning techniques for prediction of early virological suppression in antiretroviral initiating HIV patients. BMC Med Inform Decis Mak 2018; 18: 77.
5. Bonet I, García MM, Saeys Y, Van de Peer Y, Grau R. Predicting human immunodeficiency virus (HIV) drug resistance using recurrent neural networks. In International Work-Conference on the Interplay Between Natural and Artificial Computation. Berlin, Heidelberg: Springer; 2007, pp. 234-243.
6. Singh Y. Machine learning to improve the effectiveness of ANRS in predicting HIV drug resistance. Healthc Inform Res 2017; 23: 271-276.
7. Dubey A. Machine learning approaches in drug development of HIV/AIDS. Int J Mol Biol Open Access 2018; 3: 23-25.
8. 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-S131.
9. Gelaw YA, Magalhães RJS, Assefa Y, Williams G. Spatial clustering and socio-demographic determinants of HIV infection in Ethiopia, 2015-2017. Int J Infect Dis 2019; 82: 33-39.
10. Ramachandran A, Kumar A, Koenig H, et al. Predictive analytics for retention in care in an urban HIV clinic. Sci Rep 2020; 10: 6421.
11. Shen C, Yu X, Harrison RW, Weber IT. Automated prediction of HIV drug resistance from genotype data. BMC Bioinformatics 2016; 17 Suppl 8: 278.
12. 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: 171-179.
13. Singh O, Su ECY. Prediction of HIV-1 protease cleavage site using a combination of sequence, structural, and physicochemical features. BMC Bioinformatics 2016; 17: 279-289.

14. Tu W, Chen PA, Koenig N, et al. Machine learning models reveal neurocognitive impairment type and prevalence are associated with distinct variables in HIV/AIDS. J Neurovirol 2020; 26: 41-51.
15. Singh Y, Mars M. Support vector machines to forecast changes in CD4 count of HIV-1 positive patients. Sci Res Essays 2010; 5: 2384-2390.
16. Cole JH, Underwood J, Caan MW, et al. Increased brain-predicted aging in treated HIV disease. Neurology 2017; 88: 1349-1357.
17. Ahlström MG, Ronit A, Omland LH, Vedel S, Obel N. Algorithmic prediction of HIV status using nation-wide electronic registry data. EClinicalMedicine 2019; 17: 100203.
18. Zazzi M, Incardona F, Rosen-Zvi M, et al. Predicting response to antiretroviral treatment by machine learning: the EuResist project. Intervirology 2012; 55: 123-127.
19. Nan Y, Gao Y. A machine learning method to monitor China's AIDS epidemics with data from Baidu trends. PLoS One 2018; 13: e0199697.
20. 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.
21. Steiner MC, Gibson KM, Crandall KA. Drug resistance prediction using deep learning techniques on HIV-1 sequence data. Viruses 2020; 12: 560.
22. Wang G, Wei W, Jiang 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.
23. Rajpurkar P, O'Connell C, Schechter A, et al. CheXaid: deep learning assistance for physician diagnosis of tuberculosis using chest x-rays in patients with HIV. NPJ Digit Med 2020; 3: 115.
24. Mayr A, Klambauer G, Unterthiner T, Hochreiter S. DeepTox: toxicity prediction using deep learning. Front Environ Sci 2016; 3: 80.
25. Li X, Xu X, Wang J, Li J, Qin S, Yuan J. Study on prediction model of HIV incidence based on GRU neural network optimized by MHPSO. Ieee Access 2020; 8: 49574-49583.
26. Madigan EA, Curet OL, Zrinyi M. Workforce analysis using data mining and linear regression to understand HIV/AIDS prevalence patterns. Hum Resour Health 2008; 6: 2.
27. Oliveira A, Faria BM, Gaio AR, Reis LP. Data mining in HIV-AIDS surveillance system. J Med Syst 2017; 41: 51.
28. Hajipour M, Taherpour N, Fateh H, et al. Predictive factors of infant mortality using data mining in Iran. Journal of Comprehensive Pediatrics 2021; 12: e108575.
29. https://data.world/datasets/hiv