 10.15789/2220-7619-ITC-17896

IDENTIFYING THE CHARACTERISTICS OF LATE HIV DIAGNOSIS USING OPTIMIZED MACHINE LEARNING ALGORITHM

Farhadian M. [a],

Moslehi S. [a],

Mirzaei M. [b]

[a] Hamadan University of Medical Sciences, Hamadan, Iran.

[b] Center for Disease Control & Prevention, Hamadan, Iran.

 ISSN 2220-7619 (Print)
ISSN 2313-7398 (Online)

 10.15789/2220-7619-ITC-17896

Abstract

Background: Early detection of HIV infection is essential for clinical diagnosis, preventing transmission, and ensuring the safety of blood products. Individuals diagnosed late with HIV may unknowingly transmit the virus, and once diagnosed, they may experience worse health outcomes. Therefore, this study aims to identify the characteristics associated with late diagnosis of HIV patients.

Methods: In this retrospective cohort study, the information of 236 patients with HIV infection in Hamadan, the West of Iran, was collected by recording the CD4 count during 2011 to 2022 years. Late HIV diagnosis was considered with a CD4≤350/mm3. Initially, Extreme Gradient Boosting (XGBoost) and Random Forest (RF) algorithms identified important variables. Subsequently, models such as Logistic Model Tree (LMT), Classification and Regression Tree (CART), Deep Neural Network (DNN), and Support Vector Machine (SVM) were developed using a 70/30 training/test dataset split for clinical and demographic variables. Finally, the optimal model was selected based on accuracy and F1-score using Python software version 3.10.

Results: The age, logarithm of Viral Load (LVL), Wight Blood Cell (WBC), Red Blood Cell (RBC), Lymphocyte (Lym), Hematocrit (Hct), Platelet (PLT), Hemoglobin (Hb), and clinical stage variables had relative importance above 6%. Among the developed models for the importance variables, the CART with F1-score and Accuracy values of 0.887 and 0.801 and 0.897 and 0.822 for training data, respectively. The AUC value obtained for the CART was equal to 0.918.

Conclusions: Late diagnosis of HIV infection is a substantial problem, particularly in developing an algorithm that can accurately and interpretably detect disease characteristics, such as the CART, which could be essential for identifying characteristics that influence late HIV diagnosis and clinical and therapeutic decisions.

Keywords: Machine Learning, Deep Learning, Decision Tree, HIV/AIDS, Classification

 ISSN 2220-7619 (Print)
ISSN 2313-7398 (Online)

 10.15789/2220-7619-ITC-17896

ОПРЕДЕЛЕНИЕ ХАРАКТЕРИСТИК ПОЗДНЕЙ ДИАГНОСТИКИ ВИЧ С ИСПОЛЬЗОВАНИЕМ ОПТИМИЗИРОВАННОГО АЛГОРИТМА МАШИННОГО ОБУЧЕНИЯ

Фархадян М. [1],

Мослехи С. [1],

Мирзаи М. [2]

[1] Университет медицинских наук Хамадана, Хамадан, Иран.

[2] Центр по контролю и профилактике заболеваний, Хамадан, Иран.

 ISSN 2220-7619 (Print)
ISSN 2313-7398 (Online)

 10.15789/2220-7619-ITC-17896

Резюме

Введение: Раннее выявление ВИЧ-инфекции имеет важное значение для клинической диагностики, предотвращения трансмиссии и обеспечения безопасности продуктов крови. Лица с поздним диагностированием ВИЧ могут неосознанно передавать вирус, и после постановки диагноза у них могут возникнуть более неблагоприятные последствия для здоровья. Поэтому настоящее исследование направлено на выявление характеристик, связанных с поздней диагностикой ВИЧ-пациентов.

Методы: В настоящем ретроспективном когортном исследовании была собрана информация о 236 пациентах с ВИЧ-инфекцией в Хамадане (запад Ирана) путем оценки количества CD4 Т клеток периферической крови в период с 2011 по 2022 годы. Поздняя диагностика ВИЧ считалась при уровне CD4 Т клеток≤350/мм3. Первоначально алгоритмы Extreme Gradient Boosting (XGBoost) и Random Forest (RF) выявили основные переменные. Впоследствии были разработаны такие модели, как Logistic Model Tree (LMT), Classification and Regression Tree (CART), Deep Neural Network (DNN) и Support Vector Machine (SVM) с использованием 70/30 разделения набора данных для обучения/тестирования для клинических и демографических переменных. Наконец, оптимальная модель была выбрана на основе точности и F1-оценки с использованием программного обеспечения Python (версия 3.10).

Результаты: Показано, что возраст, логарифм вирусной нагрузки (LVL), содержание лейкоцитов (WBC), эритроцитов (RBC), лимфоцитов (Lym), гематокрит (Hct), уровень тромбоцитов (PLT), гемоглобина (Hb) и параметры клинической стадии имели относительную важность выше уровня в 6%. Среди разработанных моделей для переменных важности CART со значениями F1-оценки и точности 0,887 и 0,801 и 0,897 и 0,822 для

 ISSN 2220-7619 (Print)
ISSN 2313-7398 (Online)

 10.15789/2220-7619-ITC-17896

обучающих данных соответственно. Значение AUC, полученное для CART, было равно 0,918.

Выводы: Поздняя диагностика ВИЧ-инфекции является существенной проблемой, особенно при разработке алгоритма, который может точно и интерпретируемо определять характеристики заболевания, такие как CART, что может быть важно для выявления характеристик, влияющих на позднюю диагностику ВИЧ и клинические и терапевтические решения.

Ключевые слова: Машинное обучение, Глубокое обучение, Дерево решений, ВИЧ/СПИД, Классификация.

.

 ISSN 2220-7619 (Print)
ISSN 2313-7398 (Online)

 10.15789/2220-7619-ITC-17896

1 Introduction

The human immunodeficiency virus (HIV) is a type of slowly spreading virus (retroviral virus with a low replication rate) and the cause of AIDS [31]. People infected with HIV experience psychological consequences. The virus leaves negative physical, psychological, and social impacts and threatens the personal and social lives of those affected [8]. Late referral and diagnosis of HIV remains a major unresolved problem with serious consequences at individual, social, and economic levels. Early detection of HIV enables people to protect themselves from opportunistic infections in time and receive prompt treatment to prevent secondary complications [26].

The late diagnosis of HIV in humans is an important factor in the occurrence of new infections and the resulting mortality. Factors such as sexual contact, contact with blood or infected tissue, transmission from mother to child during pregnancy, childbirth or breastfeeding, and co-injection of drugs are the main causes of virus transmission among people. Despite the efforts made worldwide to control and prevent HIV, 48% of newly infected HIV patients are still diagnosed too late, and as many as 27% are at an advanced stage of the disease [6, 14]. Late diagnosis requires rapid clinical assessments, increases mortality, leads to poor therapeutic response and even higher toxicity of antiretroviral therapies [5].

A CD4 cell count below 200 cells per microliter significantly increases susceptibility to opportunistic infections, which raises the risk of contracting additional viruses and can ultimately lead to death if the disease remains untreated and its stage is not properly identified [9]. In recent years, HIV-infected individuals have had many tests and their CD4 count has been relatively low. As for the problem of late HIV diagnosis (CD4 count less than 350 cells per microliter), these people are mislabeled as HIV-infected, leading to late diagnosis [7]. Based on public health significance, late HIV diagnosis remains a significant barrier to effective HIV management and control. Individuals who are diagnosed at a late stage often face poorer clinical outcomes, experience delays in starting antiretroviral therapy (ART),

 ISSN 2220-7619 (Print)
ISSN 2313-7398 (Online)

 10.15789/2220-7619-ITC-17896

and are at a higher risk of HIV-related health issues and mortality. Additionally, late diagnosis contributes to ongoing transmission within communities, as those who are undiagnosed may unknowingly spread the virus. Treating advanced HIV also demands more resources and is more costly due to increased hospitalizations and complex care needs. The high rates of late diagnosis indicate weaknesses in screening, education, and access to healthcare. This situation highlights populations that may be underserved or face stigma and discrimination [33, 3].

Several studies on late diagnosis of HIV infection have cited demographic, behavioral and clinical factors. In these studies, using statistical tests or GLM and GAM models, factors such as being female, older, living in rural areas, alcohol dependent, smoker, understanding the stigma associated with HIV, contact with commercial sex workers, and risky sexual behavior were introduced as factors for late diagnosis [10, 22]. Classical statistical models, like the generalized linear model (GLM), particularly linear regression, assume a linear relationship between the predictors and the outcome, handling multicollinearity, and outlier sensitivity. However, this assumption may not be valid in real datasets, where relationships are often nonlinear or more complex [27].

More recently, machine learning (ML) algorithms have been used in the diagnosis of disease, helping to improve advanced treatments for healthcare professionals [20, 21]. Key applications of these algorithms include predicting HIV status, identifying risk factors or characteristics affecting the disease, improving HIV testing algorithms and personalized treatment recommendations [29, 32]. In general, machine learning has the potential to improve the accuracy and efficiency of diagnosis and treatment, which can lead to better outcomes for people living with HIV [4]. Several studies have used ML algorithms in the diagnosis and identification of risk factors affecting outcome with high accuracy, e.g. decision trees, DNN, SVM and boosting [16, 19]. The ML models offer significant advantages over traditional statistical models, especially when working with nonlinear, complex, and high-dimensional data. Then, the advantages of ML models, include Handling Nonlinear

 ISSN 2220-7619 (Print)
ISSN 2313-7398 (Online)

 10.15789/2220-7619-ITC-17896

Relationships, Automatically Capture Interactions, High Predictive Accuracy, Robustness to Outliers and Noise, Feature Importance and Selection, Model Tuning, and Regularization Options [15].

Therefore, application of ML in late HIV diagnosis employs advanced algorithms to analyze large patient datasets, identifying patterns, characteristics and risk factors. Given the variability in characteristics influencing late HIV diagnosis such as clinical significance, public health implications of early diagnosis, timely treatment initiation, and CD4 cell counts across different regions, further investigation is needed. This study aims to identify the indicators or characteristics associated with late HIV diagnosis in Hamadan.

2. Materials and methods

2.1. Data collection

In this retrospective study, demographic, clinical and laboratory biomarkers of 236 patients with HIV infection were collected as risk factors or characteristics for late diagnosis in patients referred to medical centers and counseling centers in Hamadan province, West of Iran, between 2011 and 2022 at the third visit of the patients. All patient data were completely anonymized before analysis to ensure participants' confidentiality and privacy. No identifiable personal information was employed at any stage of the study. Also, informed written consent was obtained from all participants involved in the study. The retrospective nature of the study, combined with the use of de-identified data, led the ethics committee to waive the requirement for informed consent. Demographic characteristics such as age, gender, marital status, education, history of drug addiction (H_addiction), history of injecting drug use (H_injection), prison history, employment status and condom use, as well as clinical and laboratory biomarkers such as disease stage, Hepatitis, hemoglobin (Hb), hematocrit (Hct), lymphocytes (Lym), platelets (PLT), red blood cells (RBC), white blood cells (WBC), receipt of ARV treatment, logarithm of viral load (LVL), and CD4 cell count were collected from patient records in collaboration with the treatment deputy of Hamadan College of Medical Sciences. The outcome

 ISSN 2220-7619 (Print)
ISSN 2313-7398 (Online)

variable in this study was the number of CD4 cells below 350 and above 350 as a dichotomous condition.

2.2. Data preparation

The RF and XGboost models were used to identify important variables. First, a RF algorithm was used that generated 1000 decision trees and considered entropy and Gini criteria in determining important variables. Also, to ensure the selection of important variables, the XGBoost model was used with a learning rate of 0.001 and a maximum depth of 40. Significant variables affecting the number of CD4 cells below 350 and above 350 were selected with a relative importance of more than 6%. The dataset considered was divided into two subsets of training/test data in a 70:30 ratio. Machine learning models such as XGBoosting, CART, LMT, DNN and SVM were developed for the training dataset. Finally, the classification accuracy of these models was evaluated based on the test dataset. A 5-fold cross-validation was used for the efficiency of the models.

2.3. Decision Trees (DT)

The DT structure consists of a root node, internal nodes, and output nodes (leaves). Each internal node tests a variable, branches indicate test results, and leaf nodes represent characteristics. Internal nodes divide the number of samples into sub-samples using probability metrics like entropy and Gini. The paths from root to leaf define classification rules. Various algorithms, such as CART, LMT, and RF, can be used to build decision trees for classification and regression tasks [19].

2.4. Classification and Regression tree (CART)

CART trees, developed in 1984 by Breiman et al., stand for classification and regression trees. They generate binary trees, with two output edges for each internal node. Partitions are chosen based on the splitting criterion, and the tree is pruned using cost-complexity methods (weakest link pruning or error complexity pruning). CART accounts for misclassification costs during tree construction and visualizes prior probability distributions. A key feature of CART is its ability to create regression trees, where the leaves predict continuous numerical values rather than

 10.15789/2220-7619-ITC-17896

classes. In regression, CART seeks subdivisions that minimize the squared error of predictions, with each leaf's prediction derived from the weighted average of the nodes [11]. In CART, entropy and Gini index are used to divide the nodes into several other nodes. Then These indices for a given feature, such as X for class C, $i = 1, \ldots, C$, are as follows:

$$\text{Entropy}(X) = -\sum_{i=1}^{C} \hat{p}_i log_2(\hat{p}_i),$$

$$\text{Gini}(X) = 1 - \sum_{i=1}^{C} \hat{p}_i^2,$$

Where, $\hat{p}_i$ is the probability of a data set of class i.

2.5. Logistic Model Tree (LMT)

LMT is a classification algorithm that integrates a decision tree with logistic regression. It segments the data into subsets using a decision tree and applies a logistic regression model to each subset.In LMT, a logistic regression model is fitted for each tree node using the LogitBoost algorithm. In other words, the posterior class probabilities for several classes $C_i; i = 1, \ldots, I$ are modelled and the maximum value of this probability is estimated. For any variable such as Y of class I, the LMT model with parameter β_i^T for class i is as follows:

$$P(C = I|Y = y) = \frac{\exp(\beta_I^T y)}{\sum_{i=1}^{I-1} \exp(\beta_i^T y)}.$$

The LMT tree is pruned using the CART algorithm and uses cross-validation to find the number of LogitBoost iterations to avoid overfitting the tree [12].

2.6. Deep Neural Networks (DNN)

The DNN is an artificial neural network consisting of an input layer, several hidden layers, and an output layer, designed to classify and predict complex data patterns. Its neurons employ activation functions like ReLU, sigmoid, or tanh to introduce nonlinearity, enabling the model to learn complex relationships. DNNs are trained on large datasets using optimization algorithms like stochastic gradient

 ISSN 2220-7619 (Print)
ISSN 2313-7398 (Online)

descent (SGD) to minimize a loss function that measures the disparity between predicted and actual outputs. The Backpropagation algorithm adjusts the network's weights and biases by calculating gradients from the output error and propagating them back through the network [24].

2.7. Support Vector Machine (SVM)

SVMs are supervised learning models for classification and regression that identify the hyperplane best separating classes in feature space. In n-dimensional space, this hyperplane is an (n-1)-dimensional subspace that maximize the margin, the distance to the nearest data points of each class, enhancing the model's generalization to unseen data. Support vectors, the closest data points to the hyperplane, are crucial for determining its position and orientation. SVMs can apply kernel functions to map input space into higher dimensions for non-linear separations, with common kernels being linear, polynomial, and radial basis functions (RBF). A tuning parameter(C) manages the trade-off between maximizing the margin and minimizing classification error, with variations in C affecting the model's fit to the training data [28].

2.8. Extreme Gradient Boosting (XGBoost)

XGBoost is a powerful and efficient algorithm from the gradient boosting framework that is widely used in machine learning for classification and regression tasks. This algorithm builds models sequentially, with each new model correcting the previous errors. It combines predictions from multiple weak learners (usually DT) to create a robust predictive model. XGBoost incorporates L1 (lasso) and L2 (ridge) regularization techniques to avoid overfitting, making it more robust compared to traditional gradient boosting methods. To simplify the pre-processing of the data, this algorithm can automatically check missing values during training. XGBoost uses a deep-first approach for tree construction and performs backpruning of the trees to optimize the model and reduce complexity. These algorithms can identify and prioritize important variables [23].

2.9. Evaluation Metrics

 10.15789/2220-7619-ITC-17896

The confusion matrix reveals the characteristics of a classification rule by showing the counts of correctly and incorrectly classified instances for each class. The main diagonal represents correct classifications, while the sub-diagonals indicate incorrect ones. For a binary classification, the confusion matrix can be constructed by inputting the actual and predicted values from a chosen model, as shown in Table 1.

[Table 1]

Based on the values in Table 1, the evaluation metrics can calculate defined above [19]:

Accuracy: (a + d) / (a + b + c + d)

Precision: d / (b + d)

Recall: d / (c + d)

F1-score: 2 * (Precision * Recall) / (Precision + Recall)

We conducted analyses using Python software version 3.10 with scikit-learn library.

3. Result

In this study, out of 236 patients, late HIV diagnosis in 127 (53.8%) cases. It can be seen that 69.5% of patients with late diagnosis were men. Also, the majority of patients with 65.3% and 55.5% had a history of drug addiction and injection, respectively. Then, the other descriptive information of patients, including demographic variables and blood and laboratory biomarkers, is reported in Table 2.

[Table 2]

The results of the Table 2 indicated that only the variables of clinical stage, Lym, WBC, RBC, and LVL were statistically significant at different levels of CD4. Figure 1 shows the results of the relative importance of the variables for the entire dataset based on RF and XGboost methods in CD4 detection.

[Figure 1]

According to Figure 1, the variables LVL, WBC, RBC, Lym, Hct, PLT, Hb, Age, and clinical stage have more than 6% relative importance. The proposed LMT,

 ISSN 2220-7619 (Print)
ISSN 2313-7398 (Online)

 10.15789/2220-7619-ITC-17896

CART, DNN, and SVM models classify CD4. They use variables with importance over 6% from RF and XGBoost. The criteria of recall, precision, and F1-score evaluation for the training, test, and total dataset to evaluate the performance of the introduced models are shown in Table 3.

[Table 3]

The CART decision tree emerged as the top choice for late diagnosis, based on Table 3's evaluation criteria. A flowchart illustrated the CART model's use of key characteristics: LVL, WBC, RBC, Lym, Hct, PLT, Hb, age, and clinical stage. Figure 2 displays the final CART model's accuracy across the entire dataset, encompassing all samples and their diverse attributes. Also, the ROC curve visualized classification performance (Figure 3).

[Figure 2]

[Figure 3]

According to Figure 2, the interpretability of the CART can be expressed in the identification of variables affecting the late diagnosis of HIV infection based on the priority of their importance. The CART with a maximum depth of 8 consists of 50 nodes including a root node (node zero), 24 internal nodes, and 25 leaf nodes. In the root node, there are 236 subjects with HIV infection, of which 127 subjects have CD4 less or equal than 350 and 109 subjects have CD4 more than 350. For example, the root nodes, 2, 4, 36, and 48 from the first branch on the right side of the tree show that if an individual has LVL above 5.84, age above 25 years, and RBC above 1.58, CD4 detection less than 350 will be 43 person (18.2% of subjects). Meanwhile, for root nodes, 2, 4, 36, and 47, for an individual with LVL above 5.84, age above 25 years, and RBC less than 1.58, CD4 diagnosis is less than 350 for this individual, 8 person (3.4% of subjects).

4. Discussion

In this study, of the 18 demographic and biomarker variables associated with late HIV infection diagnosis, age stood out with a relative importance exceeding 6%. These findings align with those of Mohammadi et al., Gesesew et al., and Bath et al

 ISSN 2220-7619 (Print)
ISSN 2313-7398 (Online)

[10, 2, 17]. Additionally, biomarker variables such as LVL, WBC, RBC, PLT, Lym, Hb, and Hct exhibited significant influence in the late diagnosis of HIV, each with relative significance above 6%. This supports previous work by Weissman et al. and Lee et al., [13, 30] which indicated that deviations in these variables from their normal ranges can significantly affect in late HIV diagnosis.

This study utilized variables with relative importance above 6% to create an easily interpretable tree flowchart, enabling specialists and physicians to make quick treatment decisions. The LMT, CART, DNN, SVM, and XGBoost models were developed based on the total dataset using the variables age, LVL, WBC, RBC, PLT, Lym, Hb, and Hct. The accuracy of these models in classifying late HIV diagnosis was 78.9%, 82.2%, 79.2%, 78.2%, and 79.2%, respectively. Also, the evaluation criteria revealed F1-scores of 87.7%, 89.7%, 87.2%, 87.7%, and 86.5%, respectively. Among the developed trees, the CART exhibited the highest accuracy at 82.2% and was selected as the optimal decision tree for the late diagnosis of HIV infection, with an F1-score of 89.7%. A flowchart of the CART tree was created to guide treatment decisions for patients diagnosed late with HIV (AUC=91.8%). In the study of Morales-Sánchez et al., [18] and Romero-Rodríguez et al., [25] machine learning models such as LMT and LASSO regression in determining variables related to the recovery of CD4 cells in patients with HIV infection have higher predictive accuracy compared to classical models. In Mohammadi et al.'s study [33], variables such as age, transmission method, drug injection, gender, and marital status were effective in late HIV diagnosis. Compared to this studies that primarily employed traditional statistical models, our use of ML techniques such as Random Forest and XGBoost yielded significantly improved selective Features and sensitivity in identifying risk factors or characteristics associated with late HIV diagnosis. Also, Traditional models typically assume linear relationships, which can limit their ability to detect complex interactions. In contrast, the CART model automatically captures nonlinear relationships and interactions between variables.

 10.15789/2220-7619-ITC-17896

This capability allows us to identify subtle patterns and high-risk profiles that may have been missed in earlier analyses.

According to Figures 2, the CART tree's interpretability aids clinical specialists in prioritizing patient treatment and making quicker decisions. The tree consists of 50 nodes and 25 branches from the root to the leaves. In five branches leading to leaves with numbers 10, 22, 24, 48, and 49, the highest count of CD4 cells was below 350, indicating a greater probability of late HIV diagnosis. The branch of nodes zero, 2, 4, 36, and 48 indicates the highest risk of late HIV diagnosis, namely, if a person is LVL>5.84, age> 25, and RBC>1.58, then the risk of late HIV diagnosis is 18.2%. Also, the branch of nodes zero, 2, 3, 6, 9, 12, 13, and 24 indicates the highest risk of late HIV diagnosis, namely, if a person is LVL≤5.84, PLT> 98.5, Lym>25.1, and 3.05<Hb≤10.9, then the risk of late HIV diagnosis is 6.8%. In contrast, the risk of late HIV diagnosis in other branches is reduced. For instance, the branch of nodes zero, 2, 4, 36, 47, and 50 indicates the lowest risk of late HIV diagnosis, namely, if a person is LVL>5.84, age> 25 years old, and 1.52<RBC≤1.59, then his risk of late HIV diagnosis is 0 %. In study of Lee et al. identified blood biomarkers like Lym and PLT, along with age over 45, as key factors in the late diagnosis of HIV infection [13]. Similarly, Adler et al. found that LVL, RBC, WBC, age, and gender were significant variables contributing to late HIV diagnosis [17].

By identifying key risk factors and populations disproportionately affected by late diagnosis, this work contributes to the scientific understanding of HIV transmission dynamics and informs targeted public health interventions aimed at improving early detection, reducing transmission, and enhancing patient outcomes. Also, this methodological advancement enhances the field by providing a more scalable and precise framework for identifying late HIV diagnosis cases, which is essential for timely intervention, resource allocation, and effective public health surveillance. The main limitation of this study was the lack of sufficient information about the repetitions of subjects in later times such as two years after visiting the treatment centers and recording the information of these individuals.

 ISSN 2220-7619 (Print)
ISSN 2313-7398 (Online)

 10.15789/2220-7619-ITC-17896

Acknowledgements

This work is a part of a research work in Biostatistics at the Hamadan University of Medical Sciences in Iran. We would like to thank the patients and vice-chancellor of research and technology of Hamadan University of Medical Sciences in Iran (No.140105113518).

Ethics approval and consent to participate

The procedures used in this study strictly comply with the ethical standards formulated by the Ethics Committee of Hospitals in Iran. Ethical approval for the start of data collection and use of the dataset in this study was obtained from the Ethics Committee of Hamadan University of Medical Science with the approved ethical code: IR.UMSHA.REC.1401.258. All experimental protocols that included human data adhered to the guidelines of the University of Medical Science in Hamadan, Iran, as well as the Data Protection Committee, and the Internal Review Board.

Availability of data and materials

The dataset used for analysis during the current study are not publicly available due to restrictions related to our internal review board policy. However, the dataset is available from the corresponding author on reasonable request.

Competing interests

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

 ISSN 2220-7619 (Print)
ISSN 2313-7398 (Online)

 10.15789/2220-7619-ITC-17896

ТАБЛИЦЫ

Table 1. A confusion matrix for two classes.

		Predicted outcome	
		Late diagnosis	**Without Late diagnosis**
Real outcome	**Late diagnosis**	a	b
	Without Late diagnosis	c	d

Table 2. Descriptive statistics of investigated variables in cases of late diagnosis in HIV patients.

Variable		**CD4 <= 350** Frequency (%)	**CD4 > 350** Frequency (%)	P-value*
Gender	Male	89 (70.1)	75 (68.8)	0.833
	Female	38 (29.9)	34 (31.2)	
ART	No	27 (21.3)	20 (18.3)	
	Yes	100 (78.7)	89 (81.7)	
Education	Elementary-school	56 (44.1)	42 (38.5)	0.629
	Middle-school	42 (33.1)	42 (38.5)	
	High-school	29 (22.8)	25 (22.9)	
H_addiction	No	44 (34.6)	38 (34.9)	0.540
	Yes	83 (65.4)	71 (65.1)	
H_injection	No	54 (42.5)	51 (46.8)	0.515

 ISSN 2220-7619 (Print)
ISSN 2313-7398 (Online)

 10.15789/2220-7619-ITC-17896

	Yes	73 (57.5)	58 (53.2)	
H_prison	No	60 (47.2)	54 (49.5)	0.523
	Yes	67 (52.8)	55 (50.5)	
Transmission	Injection	76 (59.8)	63 (57.8)	0.131
	Sexual	33 (26.0)	38 (34.9)	
	Other	18 (14.2)	8 (7.3)	
Job	Season worker	20 (15.7)	17 (15.6)	0.811
	Employee	41 (32.3)	33 (30.3)	
	Other	66 (52.0)	59 (54.1)	
Marital status	Married	55 (43.3)	47 (43.1)	0.541
	Single	72 (56.7)	62 (56.9)	
Clinical stage	I	59 (46.5)	64 (58.7)	0.029
	II	38 (29.9)	30 (27.5)	
	III	30 (23.6)	15 (13.8)	
Condom use	No	105 (82.7)	84 (77.1)	0.328
	Yes	22 (17.3)	25 (22.9)	
Hepatitis	No	61 (48.0)	49 (45.0)	0.695
	Yes	66 (52.0)	60 (55.0)	
		Mean (SD)	Mean (SD)	P-value**
Age	-	36.61 (8.78)	35.43 (9.01)	0.309
Hb	-	8.28 (2.45)	8.72 (2.19)	0.153
Hct	-	34.15 (7.26)	35.23 (5.79)	0.215
Lym	-	30.2 (9.75)	32.77 (10.77)	0.045
PLT	-	214.3 (128.7)	190.5 (72.54)	0.088

 ISSN 2220-7619 (Print)
ISSN 2313-7398 (Online)

 10.15789/2220-7619-ITC-17896

WBC	-	70.16 (88.66)	40.33 (27.12)	0.001
RBC	-	6.56 (9.95)	4.24 (5.28)	0.030
LVL	-	5.78 (0.54)	4.96 (0.59)	< 0.001

ART: Antiretroviral Treatment; H_addiction: History of Addiction; H_injection: History of injection drug; Hb: Hemoglobin level; Hct: Hematocrit; Lym: Lymphocytes; PLT: Plaquette; WBC: Wight blood cell; RBC: Red blood cell; LVL: logarithm of viral load

*: Chi-square test; **: two sample T test

 ISSN 2220-7619 (Print)
ISSN 2313-7398 (Online)

 10.15789/2220-7619-ITC-17896

Table 3. Classifying accuracy of proposed models.

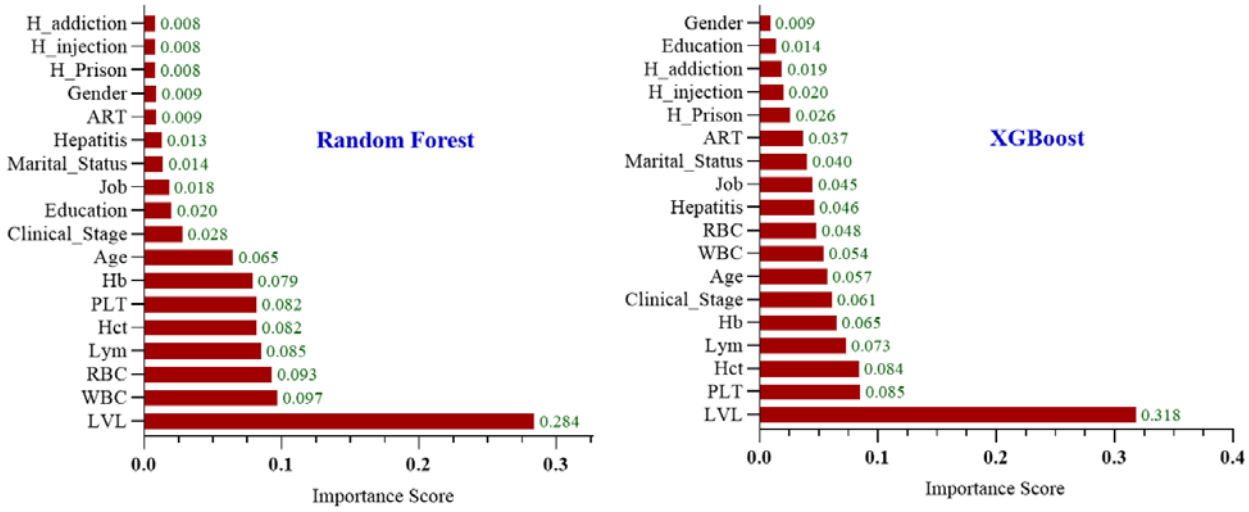
		Evaluation metric			
Model	**Subset**	Precision	Recall	F1-score	Accuracy
LMT	Train	0.854	0.913	0.883	0.793
	Test	0.836	0.896	0.865	0.775
	Total	0.848	0.907	0.877	0.789
DNN	Train	0.848	0.908	0.877	0.798
	Test	0.829	0.889	0.858	0.784
	Total	0.843	0.904	0.872	0.792
SVM	Train	0.828	0.888	0.857	0.787
	Test	0.836	0.896	0.865	0.779
	Total	0.848	0.908	0.877	0.789
XGBoost	Train	0.820	0.881	0.849	0.786
	Test	0.811	0.871	0.840	0.774
	Total	0.836	0.896	0.865	0.792
CART	Train	0.848	0.908	**0.877**	**0.801**
	Test	0.835	0.905	**0.869**	**0.798**
	Total	0.869	0.929	**0.897**	**0.822**

 ISSN 2220-7619 (Print)
ISSN 2313-7398 (Online)

 10.15789/2220-7619-ITC-17896

РИСУНКИ

Figure 1. Relative importance variables based on RF and XGBoost.



 ISSN 2220-7619 (Print)
ISSN 2313-7398 (Online)

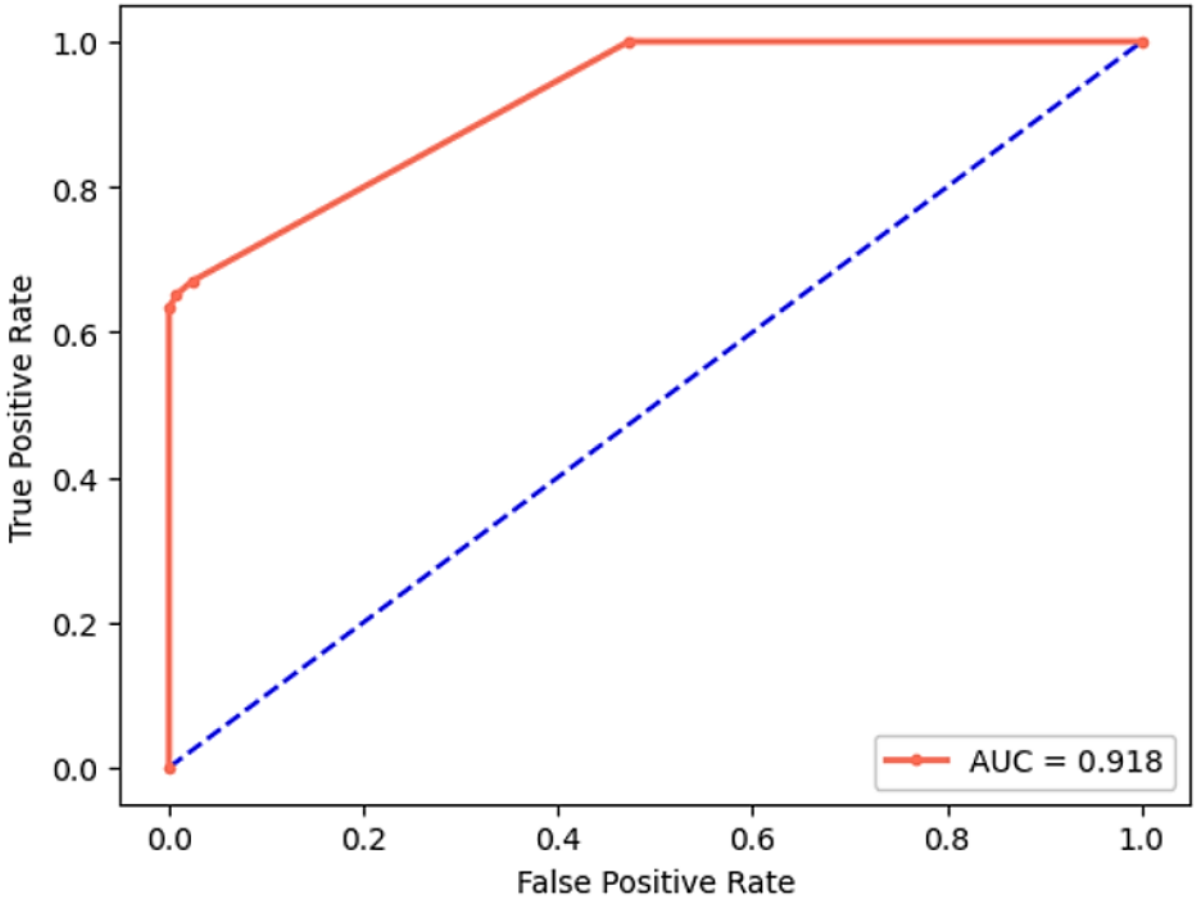
 10.15789/2220-7619-ITC-17896

Figure 2. CART flowchart in CD4-level classification.

 ISSN 2220-7619 (Print)
ISSN 2313-7398 (Online)

 10.15789/2220-7619-ITC-17896

Figure 3. ROC curve for the CART model.



 ISSN 2220-7619 (Print)
ISSN 2313-7398 (Online)

 10.15789/2220-7619-ITC-17896

ТИТУЛЬНЫЙ ЛИСТ_МЕТАДАННЫЕ

Блок 1. Информация об авторе ответственном за переписку

Samad Moslehi, Ph.D, Assistant Professor of Biostatistics, Department of Biostatistics, School of Public Health, Modeling of Noncommunicable Diseases Research Center, Hamadan University of Medical Sciences, Hamadan, Iran;

Tel: +98(81)38380090;

Fax: +98(81)38380509;

E-mail: samadmoslehi999@gmail.com.

Блок 1. Информация об авторе ответственном за переписку

Самад Мослехи, доктор философии, доцент кафедры биостатистики, кафедра биостатистики, Школа общественного здравоохранения, Исследовательский центр моделирования неинфекционных заболеваний, Медицинский университет Хамадана, Хамадан, Иран.

Тел.: +98(81)38380090;

Факс: +98(81)38380509;

Электронная почта: samadmoslehi999@gmail.com.

Блок 2. Информация об авторах

Maryam Farhadian, Ph.D, Associate Professor of Biostatistics, Department of Biostatistics, School of Public Health and Research Center for Health Sciences, Hamadan University of Medical Sciences, Hamadan, Iran. Email: maryam_farhadian80@yahoo.com. ORCID: 0000-0002-6054-9850

Mohammad Mirzaei, MS.c, Disease Control Expert, Center for Disease Control & Prevention, Deputy of Health Services, Hamadan University of Medical Sciences, Hamadan, Iran. Email: mirzaei3589@gmail.com. ORCID: 0000-0001-9428-059X

 ISSN 2220-7619 (Print)
ISSN 2313-7398 (Online)

 10.15789/2220-7619-ITC-17896

Марьям Фархадиан, доктор философии, доцент кафедры биостатистики, кафедра биостатистики, Школа общественного здравоохранения и исследовательский центр медицинских наук, Медицинский университет Хамадана, Хамадан, Иран. Электронная почта: maryam_farhadian80@yahoo.com. ORCID: 0000-0002-6054-9850

Мохаммад Мирзаи, магистр наук, эксперт по контролю заболеваний, Центр по контролю и профилактике заболеваний, заместитель службы здравоохранения, Медицинский университет Хамадана, Хамадан, Иран. Электронная почта: mirzaei3589@gmail.com. ORCID: 0000-0001-9428-059X

 ISSN 2220-7619 (Print)
ISSN 2313-7398 (Online)

 10.15789/2220-7619-ITC-17896

Блок 3. Метаданные статьи

IDENTIFYING RISK FACTORS OF LATE HIV DIAGNOSIS USING OPTIMIZED MACHINE LEARNING ALGORITHM

ОПРЕДЕЛЕНИЕ ХАРАКТЕРИСТИК ПОЗДНЕЙ ДИАГНОСТИКИ ВИЧ С ИСПОЛЬЗОВАНИЕМ ОПТИМИЗИРОВАННОГО АЛГОРИТМА МАШИННОГО ОБУЧЕНИЯ

Сокращенное название статьи для верхнего колонтитула:

MACHINE LEARNING IN LATE HIV DIAGNOSIS

МАШИННОЕ ОБУЧЕНИЕ ПРИ ПОЗДНЕЙ ДИАГНОСТИКЕ ВИЧ

Keywords: Machine Learning, Deep Learning, Decision Tree, HIV/AIDS, Classification.

Ключевые слова: Машинное обучение, Глубокое обучение, Дерево решений, ВИЧ/СПИД, Классификация.

Оригинальные статьи.

Количество страниц текста – 10,

количество таблиц – 3,

количество рисунков – 3.

23.03.2025

 ISSN 2220-7619 (Print)
ISSN 2313-7398 (Online)

 10.15789/2220-7619-ITC-17896

СПИСОК ЛИТЕРАТУРЫ

Number	Authors, title of a publication and source where it was published, publisher's imprint	Reference's URL
1	Adler A., Mounier-Jack S., Coker R. Late diagnosis of HIV in Europe: definitional and public health challenges. *AIDS care. 2009, vol. 21, no. 3, pp: 284-293.*	https://pubmed.ncbi.nlm.nih.gov/19031304/
2	Bath RE., Emmett L., Verlander NQ., Reacher M. Risk factors for late HIV diagnosis in the East of England: evidence from national surveillance data and policy implications. *International journal of STD & AIDS. 2019, vol. 30, no. 1, pp: 37-44.*	https://pubmed.ncbi.nlm.nih.gov/30170527/
3	Bendera, A., Baryomuntebe, D. M., Kevin, N. U., Nanyingi, M., Kinengyere, P. B., Mujeeb, S., & Sulle, E. J. (2024). Determinants of Late HIV Diagnosis and Advanced HIV Disease Among People Living with HIV in Tanzania. HIV/AIDS-Research and Palliative Care. *2024, vol. 26, no. 16, pp: 313–323.*	https://pubmed.ncbi.nlm.nih.gov/39220740/

ISSN 2220-7619 (Print)
ISSN 2313-7398 (Online)

 10.15789/2220-7619-ITC-17896

4	Bisaso KR., Anguzu GT., Karungi SA., Kiragga A., Castelnuovo B. A survey of machine learning applications in HIV clinical research and care. *Computers in biology and medicine. 2017, vol. 91, pp: 366-371.*	https://pubmed.ncbi.nlm.nih.gov/29127902/
5	Buetikofer S. Prevalence and risk factors of late presentation for HIV diagnosis and care in a tertiary referral center in Switzerland. *University of Zurich. 2014, pp: 1-8.*	https://pubmed.ncbi.nlm.nih.gov/24723302/
6	Camoni L., Raimondo M., Regine V., Salfa MC., Suligoi B. Late presenters among persons with a new HIV diagnosis in Italy, 2010–2011. *BMC Public Health. 2013, vol. 13, no. pp. 1-6.*	https://pubmed.ncbi.nlm.nih.gov/23537210/
7	Croxford S, Stengaard AR., Brännström J., Combs L., Dedes N., Girardi E., Grabar S., Kirk O., Kuchukhidze G., Lazarus JV., Noori T. Late diagnosis of HIV: an updated consensus definition. *HIV medicine. 2022, vol. 23, no. 11, pp:1202-1208.*	https://pubmed.ncbi.nlm.nih.gov/36347523/
8	Gallo RC. A reflection on HIV/AIDS research after 25 years. *Retrovirology. 2006, vol. 3, no. 1, pp.1-7.*	https://pubmed.ncbi.nlm.nih.gov/17054781/
9	Gelaw YA., Senbete GH., Adane AA., Alene KA. Determinants of late presentation to HIV/AIDS care in Southern Tigray Zone, Northern Ethiopia: an	https://pubmed.ncbi.nlm.nih.gov/26633988/

 ISSN 2220-7619 (Print)
ISSN 2313-7398 (Online)

	institution-based case–control study. *AIDS research and therapy. 2015, vol. 12, no. 1, pp: 1-8.*	
10	Gesesew HA., Ward P., Woldemichael K., Mwanri L. Late presentation for HIV care in Southwest Ethiopia in 2003–2015: prevalence, trend, outcomes and risk factors. *BMC infectious diseases. 2018, vol. 18, pp: 1-11.*	https://pubmed.ncbi.nlm.nih.gov/29378523/
11	Holzinger A. Data mining with decision trees: theory and applications. *Online Information Review. 2015, vol. 39, no. 3, pp: 437-448.*	https://www.worldscientific.com/worldscibooks/10.1142/6604#t=toc
12	Landwehr N., Hall M., Frank E. Logistic model trees. *2005, vol. 59, pp: 161-205.*	https://doi.org/10.1007/s10994-005-0466-3
13	Lee C-Y., Lin Y-P., Wang S-F., Lu P-L. Late CART initiation consistently driven by late HIV presentation: A multicenter retrospective cohort study in Taiwan from 2009 to 2019. *Infectious Diseases and Therapy. 2022, vol. 11, no. 3, pp: 1033-1056.*	https://pubmed.ncbi.nlm.nih.gov/35301666/
14	Likatavicius G., Van de Laar M. HIV and AIDS in the European Union, 2011. *Eurosurveillance. 2012, vol. 17, no. 48, pp. 1-17.*	https://pubmed.ncbi.nlm.nih.gov/23218388/

ISSN 2220-7619 (Print)
ISSN 2313-7398 (Online)

15	Madakkatel I, Zhou A, McDonnell MD, Hyppönen E. Combining machine learning and conventional statistical approaches for risk factor discovery in a large cohort study. Scientific reports. 2021, vol. 11, no. 1, pp :22997.	https://pubmed.ncbi.nlm.nih.gov/34837000/
16	Mi JX., Li AD., Zhou LF. Review study of interpretation methods for future interpretable machine learning. *IEEE Access. 2020, vol. 8, pp: 191969 -191985.*	https://ieeexplore.ieee.org/document/9234594
17	Mohammadi Y., Mirzaei M., Shirmohammadi-Khorram N., Farhadian M. Identifying risk factors for late HIV diagnosis and survival analysis of people living with HIV/AIDS in Iran (1987–2016). *BMC infectious diseases. 2021, vol. 21, no. 1, pp: 1-9.*	https://pubmed.ncbi.nlm.nih.gov/33906638/
18	Morales-Sánchez, R., Montalvo, S., Riaño, A., Martínez, R. and Velasco, M. Early diagnosis of HIV cases by means of text mining and machine learning models on clinical notes. *Computers in Biology and Medicine. 2024, vol. 179, pp: 108830.*	https://pubmed.ncbi.nlm.nih.gov/38991321/
19	Moslehi S., Rabiei N., Soltanian AR., Mamani M. Application of machine learning models based on decision trees in classifying the factors affecting mortality of COVID-19 patients in Hamadan, Iran. *BMC medical informatics and decision making. 2022, vol. 22, no. 1, pp: 192.*	https://pubmed.ncbi.nlm.nih.gov/35871639/

ISSN 2220-7619 (Print)
ISSN 2313-7398 (Online)

20	Najafi-Vosough R., Faradmal J., Hosseini SK., Moghimbeigi A., Mahjub H. Predicting Hospital Readmission in Heart Failure Patients in Iran: A Comparison of Various Machine Learning Methods. *Healthcare informatics research. 2021, vol. 27, no. 4, pp: 307-14.*	https://pubmed.ncbi.nlm.nih.gov/34788911/
21	Najafi-Vosough R., Faradmal J., Tapak L., Alafchi B., Najafi-Ghobadi K., Mohammadi T. Prediction the survival of patients with breast cancer using random survival forests for competing risks. *Journal of preventive medicine and hygiene. 2022, vol. 63, no. 2. pp: 298-303.*	https://pubmed.ncbi.nlm.nih.gov/35968067/
22	Nyika H., Mugurungi O., Shambira G., Gombe NT., Bangure D., Mungati M., Tshimanga M. Factors associated with late presentation for HIV/AIDS care in Harare City, Zimbabwe, 2015. *BMC Public Health. 2016, vol. 16, no. 369, pp: 1-7.*	https://pubmed.ncbi.nlm.nih.gov/27142869/
23	Osman, A.I.A., Ahmed, A.N., Chow, M.F., Huang, Y.F. and El-Shafie, A. Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia. *Ain Shams Engineering Journal. 2021, vol. 12, no. 2, pp: 1545-1556.*	https://doi.org/10.1016/j.asej.2020.11.011

ISSN 2220-7619 (Print)
ISSN 2313-7398 (Online)

24	Reyad M., Sarhan AM., Arafa M. A modified Adam algorithm for deep neural network optimization. *Neural Computing and Applications. 2023, vol. 35, no. 23, pp: 17095-17112.*	https://doi.org/10.1007/s00521-023-08568-z
25	Romero-Rodríguez, D.P., Ramírez, C., Imaz-Rosshandler, I., Ormsby, C.E., Peralta-Prado, A., Olvera-García, G., Cervantes, F., Würsch-Molina, D., Romero-Rodríguez, J., Jiang, W. and Reyes-Terán, G. Machine learning-selected variables associated with CD4 T cell recovery under antiretroviral therapy in very advanced HIV infection. *Translational Medicine Communications. 2020, vol. 5, pp: 1-10.*	https://doi.org/10.1186/s41231-020-00058-x
26	Rotily M., Bentz L., Pradier C., Obadia Y., Cavailler P. Factors related to delayed diagnosis of HIV infection in southeastern France. *International journal of STD & AIDS. 2000, vol. 11, no. 8, pp. 531-535.*	https://pubmed.ncbi.nlm.nih.gov/10990338/
27	Roustaei N. Application and interpretation of linear-regression analysis. Med Hypothesis Discov Innov Ophthalmol. *2024, vol. 13, no. 3, pp :151-159.*	https://pubmed.ncbi.nlm.nih.gov/39507810/
28	Valkenborg D., Rousseau AJ., Geubbelmans M., Burzykowski T. Support vector machines. *Official Journal of the American Association of Orthodontists. 2023, vol. 164, no. 5, pp: 754-757.*	https://pubmed.ncbi.nlm.nih.gov/37914440/

29	Wang D., Larder B., Revell A., Montaner J., Harrigan R., De Wolf F., Lange J., Wegner S., Ruiz L., Pérez-Elías MJ., Emery S. A comparison of three computational modelling methods for the prediction of virological response to combination HIV therapy. *Artificial intelligence in medicine. 2009, vol. 47, no. 1, pp: 63-74.*	https://pubmed.ncbi.nlm.nih.gov/19524413/
30	Weissman S., Yang X., Zhang J., Chen S., Olatosi B., Li X. Using a Machine Learning Approach to Explore Predictors of Health Care Visits as Missed Opportunities for HIV Diagnosis. *AIDS (London, England). 2021, vol. 35, no. 1, pp: S7-S18.*	https://pubmed.ncbi.nlm.nih.gov/33867485/
31	World Health Statistics 2023: monitoring health for the SDGs, sustainable development goals.	*https://www.who.int/publications/i/item/9789240074323*.
32	Xiang Y., Du J., Fujimoto K., Li F., Schneider J., Tao C. Application of artificial intelligence and machine learning for HIV prevention interventions. *The Lancet HIV. 2022, vol. 9, no. 1, pp: 54-62.*	https://pubmed.ncbi.nlm.nih.gov/34762838/
33	Zhao J, Gao M, Zhao D, Tian W. Prevalence of late HIV diagnosis and its impact on mortality: A comprehensive systematic review and meta-analysis. HIV Med. *2025, vol. 26, no. 4.*	https://pubmed.ncbi.nlm.nih.gov/40181601/

ISSN 2220-7619 (Print)
ISSN 2313-7398 (Online)