BMJ Health & Care Informatics

Evaluating machine learning algorithms for predicting HIV status among young Thai men who have sex with men

Krittaka Soha,[1] Sadiporn Phuthomdee,[2,3] Thanapat Srichai,[4] Lanchakorn Kittiratanawasin,[1,5] Win Min Han,[6,7] Sirinya Teeraananchai [1,2]

To cite: Soha K, Phuthomdee S, Srichai T, *et al* . Evaluating machine learning algorithms for predicting HIV status among young Thai men who have sex with men. *BMJ Health Care Inform* 2025;**32**:e101189. doi:10.1136/bmjhci-2024-101189

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/bmjhci-2024-101189).

Received 06 July 2024
Accepted 06 May 2025

[1]Master of Biomedical Data Science program, Kasetsart University, Bangkok, Thailand
[2]Department of Statistics, Kasetsart University, Bangkok, Thailand
[3]Panyananthaphikkhu Chonprathan Medical Center, Srinakharinwirot University, Nonthaburi, Thailand
[4]National Health Security Office, Bangkok, Thailand
[5]Department of Mathematics, Kasetsart University, Bangkok, Thailand
[6]HIV-NAT, Thai Red Cross AIDS Research Centre, Bangkok, Thailand
[7]Kirby Institute, University of New South Wales, Sydney, New South Wales, Australia

Correspondence to
Dr Sirinya Teeraananchai; sirinya.te@ku.th

ABSTRACT

Objective This study aimed to develop machine learning (ML) models to predict HIV status and assessed the factors associated with HIV infection among young men who have sex with men (MSM) under the Universal Health Coverage (UHC) programme in Thailand.

Methods Young MSM aged 15–24 years who underwent HIV testing through the UHC programme from 2015 to 2022 were included. Data were divided into training (70%) and testing (30%) sets, with the Synthetic Minority Oversampling Technique (SMOTE) applied to address data set imbalance. ML models, including logistic regression, k-nearest neighbour (KNN), random forest, extreme gradient boosting (XGB) and AdaBoost, were used to predict HIV infection.

Results Among 146 813 young MSM, 11% were diagnosed with HIV. While KNN initially outperformed other ML models, the sensitivity of all models using the original data set was low due to imbalanced data. After applying SMOTE, the XGB model showed the best performance with an accuracy of 0.72, sensitivity of 0.73, specificity of 0.72 and the area under the curve of 0.72. The top predictors of HIV infection were the year of HIV testing (68%), age (55%) and targeted HIV testing (54%).

Discussion This study demonstrates the potential of ML models, particularly XGB, in predicting HIV infection among young MSM in Thailand under the UHC programme. The application of SMOTE improved model sensitivity, addressing data imbalance and enhancing predictive accuracy.

Conclusions ML models have the potential to enhance HIV risk assessment and inform targeted prevention strategies for high-risk populations.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Machine learning algorithms are becoming more widely used in healthcare to assess disease risk and improve diagnoses.
⇒ This can process large data sets to uncover patterns and risk factors that traditional methods might miss for HIV infection prediction.

WHAT THIS STUDY ADDS

⇒ The XGBoost model, processed by applying Synthetic Minority Oversampling Technique to address the imbalance in the data set, showed the highest accuracy in predicting HIV infections among young men who have sex with men from real-world setting database.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ These findings suggest that applying the XGBoost model to real-world data can effectively estimate HIV prevalence, aiding in the planning of targeted interventions and prevention programmes.

INTRODUCTION

In 2022, there were an estimated 480 000 new HIV infections among individuals aged 10–24, with 140 000 of these cases occurring mainly in people aged 10–19, which increasingly represented a significant segment of the global population with HIV.[1 2] Despite declining HIV infections in Asia-Pacific, transmission among key populations (KPs) remains a concern, especially among men who have sex with men (MSM) with high and rising prevalence in some countries. A previous study reported a 2.5-fold increase in HIV prevalence among young MSM in Malaysia between 2014 and 2017.[3] A recent study from Indonesia conducted a respondent-driven sampling survey of 211 young MSM between 2018 and 2019 to estimate HIV prevalence and associated risk factors, finding that 30% of young MSM were HIV antibody positive.[4] HIV surveillance has shown an increasing prevalence among Bangkok MSM aged 22 years or younger, rising from 13% in 2003 to 22% in 2007, and 24% in 2014.[5] Briefly, the Universal Health Coverage (UHC) programme in Thailand offers free HIV testing two times a year as a target HIV test to early detect HIV infection.[6 7] Scaling up HIV testing among KPs in Thailand has resulted in an increase in the diagnosis of new HIV infections among young MSM.[8] In 2018, approximately 45.1% of MSM who were diagnosed with HIV infections were at age 15–24 years in Thailand, which had the highest proportion compared with other age

groups.[9] There have been limited studies on HIV prevalence in Thailand in recent years.

Previous studies employed logistic regression (LR) models to predict HIV infection and its associated factors among MSM in cross-sectional studies,[4 10–12] as well as Poisson regression and survival analysis to assess newly HIV infections in cohort studies.[13 14] Recently, machine learning (ML) algorithms have seen widespread use in the medical field, particularly for HIV infection risk research.[15 16] Several studies indicate that ML algorithms are advantageous over multivariable LR models in predicting HIV infection.[15–17] Previous study, based on 5 yearly surveys from Zimbabwe, applied supervised ML models to predict HIV status, reporting that extreme gradient boosting (XGB) was the best-performing algorithm for identifying HIV infection.[18] Similarly, a study from China indicates that ML algorithms have been used to predict HIV infection among MSM using cross-sectional data, demonstrating that ML models with random forest (RF) algorithms after adjusting the Synthetic Minority Oversampling Technique (SMOTE), which is applied to solve the problem of unbalanced data, perform significantly better than conventional LR models. [19]These algorithms could be applied to all populations including the general population and KPs, with variations depending on the characteristics and dynamics of the epidemic and the behaviour risks.[20] However, ML techniques have not yet been used to predict HIV infections among MSM in Thailand. This study aims to apply ML algorithms to predict HIV status among young MSM and to assess HIV infection and its associated factors through the UHC programme in Thailand.

METHODS

Data source

Briefly, the UHC programme serves as the primary health insurance initiative for HIV testing in Thailand, providing HIV care services to people living with HIV.[7] The database collects information as the basis for reimbursing hospitals for laboratory tests. HIV tests are free two times a year for all Thai people. The data is required for reimbursement under the UHC programme in Thailand. If the information is incomplete, data monitoring processes are implemented to recheck and validate it. The data set has already been cleaned, and missing data were addressed by removing rows with missing values when the proportion of missing data was small (eg, <5%), ensuring minimal impact on the analysis. For variables with a higher proportion of missing data (eg, >15%), the missing values were categorised as unknown. The management of the database is overseen by the National Health Security Office (NHSO).[7]

Study population

This cross-sectional study included young MSM aged 15–24 years who received HIV testing through the UHC programme in Thailand between 2015 and 2022. HIV status was defined as HIV infection based on a positive result recorded by hospitals using the testing algorithms implemented at each facility. Individuals with HIV were those diagnosed as HIV-positive for the first time, while individuals without HIV were those diagnosed as HIV-negative during their most recent HIV test when they returned for testing. HIV prevalence was defined as the proportion of young MSM diagnosed with HIV divided by the total number of young MSM who underwent HIV testing during the study period.

Study covariates

The main predictors of this data include age, year at HIV testing, targeted HIV testing, main insurance care system, President's Emergency Plan for AIDS Relief (PEPFAR) in Thailand, region of resident, non-occupational post-exposure prophylaxis (nPEP), number of HIV testing, partner of people with HIV and prisoners. Age was categorised into adolescents and young adults (15–19 years) and those aged 20–24 years, respectively. We classified individuals into two groups for targeted HIV tests, along with offering free testing through the UHC benefits as yes for clients who had HIV tests for 1–2 times in a year and defined as no for those who did not do a test based on programme offering. Year at HIV testing was categorised into 2015–2017, 2018–2020 and 2021–2022 and main insurance care system was categorised into four classes based on the predominant insurance care system: Social Security Scheme (SSS), Universal Coverage Scheme (UCS), Welfare Scheme (WEL) and others. We also classified the number of HIV testing into three groups (1, 2 and more than 3) during study periods. There was no missing data in this study.

STATISTICAL ANALYSIS

The characteristics of young MSM were described by HIV status (young MSM with HIV vs young MSM without HIV). Categorical characteristics were compared using Pearson's Chi-square (χ^2) test, and continuous characteristics were compared using the Wilcoxon rank-sum test, as appropriate.

Model establishing

First, the data set was randomly divided into two groups using stratified random sampling for training data (70%) and testing data (30%). We used the fivefold cross-validation method within the training set by dividing the data into five groups for ensuring that each group represented a proportional distribution of individual characteristics. Each group was used as the internal validation set for one of the five replications. This approach helps evaluate the ability of the model to generalise to unseen data and provides a robust estimate by reducing variability.[21 22] This study used SMOTE, which is a method for balancing the training data set to improve the ability of classification and reduce the risk of overfitting for the minority class sample.[23] Second, five ML algorithms were

compared for predicting HIV infection among young MSM including LR, k-nearest neighbour (KNN), RF, XGB and AdaBoost (ADB). These algorithms are widely used and effective for classification problems, each offering unique features and advantages. The hyperparameters of each model were optimised using Grid-search to identify the optimal configuration. The RF and XGB models were tuned with the hyperparameters 'n_estimators' and 'max_depth'; the ADB model was tuned with 'n_estimators' and 'learning_rate'; the LR model was tuned with 'C'; and the KNN model was tuned with 'n_neighbors'.

Model evaluation

The performance of the ML models was evaluated using various metrics including accuracy, precision, recall, F1-score, sensitivity, specificity and the receiver operating characteristic (ROC) curve by calculating the area under the curve (AUC). F1-score, precision and recall with weighted average were used as evaluation metrics to account for class imbalance, as the data set used in HIV prediction was imbalanced. The performance of these models was evaluated on the testing data (figure 1). SHapley Additive exPlanations (SHAP) were used to calculate the contribution of each feature to the model's prediction, providing SHAP values that indicate the importance of individual features. SHAP is a powerful tool for interpreting complex ML models, as it helps to understand how input features influence predictions. In this study, SHAP was applied to identify the key predictors of HIV infection by ranking features based on their contribution to the performance of the best-performing ML models.[24] LR analysis was used to calculate odd raito (OR) with a 95% CI for interpreting the association between risk factors and HIV infection among young MSM in Thailand.

ML analyses were performed using Python (V.3.9), and statistical analyses were conducted using Stata (V.18, StataCorp, MP). Statistical significance was defined as a two-sided p value of less than 0.05.

RESULTS

Population characteristics

A total of 146 813 individuals were included, 11% (16 268) were people with HIV. Characteristics of participants are shown in table 1, stratified by people with or without HIV. The median age was 20 years (IQR: 17–22), with 49% (71 850) aged between 15 and 19 years, and 51% (74 963) aged between 20 and 24 years. 11% (15 484) were tested between 2015 and 2017, 35% (51 454) between 2018 and 2020 and 54% (79 875) between 2021 and 2022. Most young MSM were diagnosed with HIV through targeted HIV testing (92%).

HIV prevalence outcome

The HIV prevalence among young MSM was 11%. We found that HIV prevalence was higher among young MSM aged 20–24 years compared with those aged 15–19 years (16% vs 6%). Online supplemental figure 1 shows the prevalence of HIV across provinces in Thailand by calendar year of HIV testing. There was a decreasing trend in HIV prevalence over time by year of HIV testing: 32% in 2015–2017, 14% in 2018–2020 and 4% in 2021–2022. The highest HIV prevalence among young MSM was found in Bangkok, followed by the Northeastern region.

Performance comparison of the models

The model construction approach involves three stages. We performed the original data set by stratified randomly sampling for model development by splitting into training set and testing set. The training set was then prepared in two forms: the original data set and the SMOTE-processed data set (online supplemental table 1 and 2) to address the issue of data set imbalance. The performance of five ML classification models in predicting HIV infection among young MSM is described in table 2. For models trained on the original imbalanced data set, KNN, RF, XGB, ADA and LR demonstrated AUC values ranging from 0.52 to 0.57. Precision values ranged from 0.35 to 0.55, indicating a moderate ability to correctly identify HIV-positive cases. Notably, recall values were low for all models (0.04–0.16) along with F1-score in a range of 0.07–0.22, suggesting poor detection of positive cases in the imbalanced data set. After addressing the imbalance using SMOTE-processed data, model performance improved significantly, particularly in sensitivity and recall. The RF, XGB and ADA models achieved the highest AUC values (0.72), while LR had an AUC of 0.70. Sensitivity increased notably across all models, with RF, XGB and ADA achieving 0.72–0.73. This improvement was reflected in a slight increase in the F1-score (0.23–0.37), while precision values decreased slightly (0.24–0.30) due to the increased identification of positive cases. Overall F1-scores average improved to 0.76–0.85, indicating a better balance between precision and recall with weighted average for imbalance data. Among the models evaluated, XGB and ADA achieved the best overall performance, with an AUC of 0.72 and sensitivity of 0.73, while the confusion matrix showed that the XGB model had higher true positive and true negative values than the ADA model (online supplemental figure 2 and 3). The ROC of the five ML algorithms using both the original and SMOTE-processed data is shown in figure 2.

The evaluation of feature importance

The evaluation of SHAP was conducted using the XGB model, which outperformed other ML models. Figure 3 displays the ranked features, ordered from highest to lowest importance, highlighting their significance in predicting HIV infection. Features importance included the year of HIV testing, age, targeted HIV testing, region of residence, primary insurance care system, PEPFAR support, number of HIV tests, prisoner status, nPEP usage and having a partner living with HIV. The top five important features in the model, including the year of HIV testing (68%),

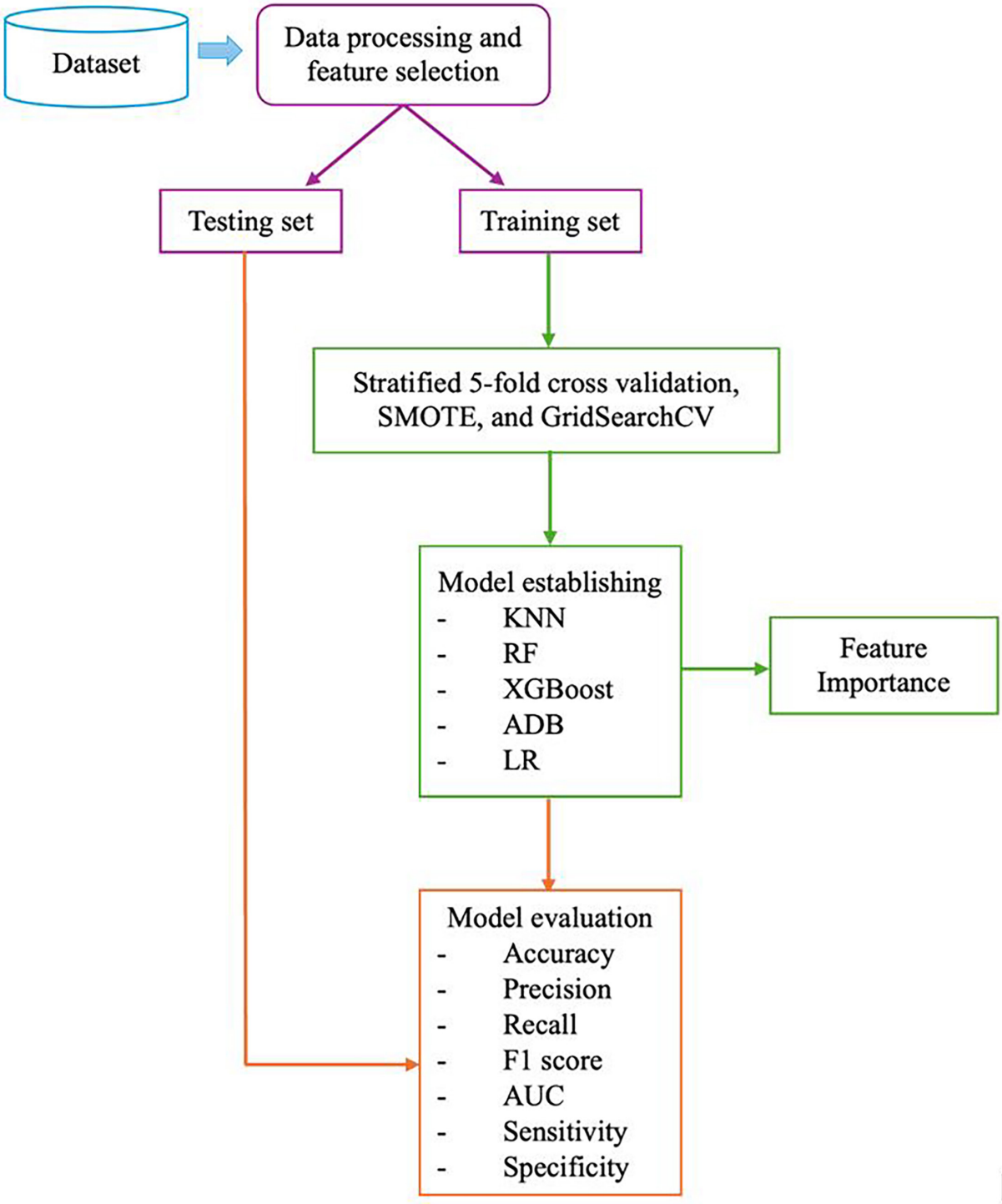

Figure 1 Study flow diagram of data preparation and model prediction. AUC, area under the curve; ADB, AdaBoost; KNN, k-nearest neighbor; LR, logistic regression; RF, random forest; SMOTE, Synthetic Minority Oversampling Technique; XGB, extreme gradient boosting.

age group (55%), targeted HIV testing (54%), region of residence (46%) and the primary insurance care system (35%), all of which demonstrate high importance but contribute negatively to the model's predictions. In contrast, the age feature (55%) also shows high importance but contributes positively to the model's predictions.

Interpretation of factors associated with HIV infection

A multivariate LR analysis was performed to study the predictors associated with HIV infection (online supplemental table 3). Young MSM aged 20–24 years (aOR 2.63, 95% CI 2.53 to 2.74) demonstrated higher odds of HIV infection compared with those aged 15–20 years. Young MSM who received HIV testing in the calendar year

Table 1 Characteristics of HIV infection in young men who have sex with men (MSM)

	Young MSM without HIV	Young MSM with HIV	Overall	P value
N (%)	130 545 (89)	16 268 (11)	146 813 (100)	
Median (IQR) age (years)	19 (17–21)	21 (19–23)	20 (17–22)	<0.001
Age group (years)				<0.001
15–19 years	67 385 (52)	4465 (27)	71 850 (49)	
20–24 years	63 160 (48)	11 803 (73)	74 963 (51)	
Year at HIV testing				<0.001
2015–2017	10 596 (8)	4888 (30)	15 484 (11)	
2018–2020	44 428 (34)	7026 (43)	51 454 (35)	
2021–2022	75 521 (58)	4354 (27)	79 875 (54)	
Targeted HIV testing				<0.001
No	36 119 (28)	1354 (8)	37 473 (26)	
Yes	94 426 (72)	14 914 (92)	109 340 (74)	
Main insurance care system				<0.001
SSS	13 382 (10)	3488 (22)	16 870 (12)	
UCS	88 594 (68)	10 806 (66)	99 400 (68)	
WEL	22 546 (17)	1596 (10)	24 142 (16)	
Others	6023 (5)	378 (2)	6401 (4)	
PEPFAR in Thailand				<0.001
No	59 430 (46)	6416 (39)	65 846 (45)	
Yes	71 115 (54)	9852 (61)	80 967 (55)	
Region of resident				<0.001
Bangkok	16 946 (13)	4396 (27)	21 342 (15)	
Central	15 785 (12)	1693 (10)	17 478 (12)	
Eastern	12 875 (10)	1312 (8)	14 187 (10)	
Northern	28 085 (22)	2610 (16)	30 695 (21)	
Northeastern	33 359 (26)	4596 (28)	37 955 (26)	
Southern	17 218 (13)	1230 (8)	18 448 (12)	
Western	6277 (4)	431 (3)	6708 (4)	
nPEP use				0.012
No	130 438 (99)	16 246 (99)	146 702 (99)	
Yes	107 (1)	4 (1)	111 (1)	
Number of HIV testing (times)				<0.001
1	82 125 (63)	13 180 (81)	95 305 (65)	
2	30 196 (23)	1913 (12)	32 109 (22)	
More than 3	18 224 (14)	1175 (7)	19 399 (13)	
Partner of people with HIV				<0.001
No	130 447 (99)	16 216 (99)	146 663 (99)	
Yes	98 (1)	52 (1)	150 (1)	
Prisoners status				<0.001
No	130 188 (99)	16 259 (99)	146 447 (99)	
Yes	357 (1)	9 (1)	366 (1)	

Categorical characteristics were compared using Pearson's χ^2 test, and continuous characteristics were compared using the Wilcoxon rank-sum test.
nPEP, non-occupational postexposure prophylaxis; PEPFAR, President's Emergency Plan for AIDS Relief; SSS, Social Security Scheme; UCS, Universal Coverage Scheme; WEL, Welfare Scheme.

Table 2 Comparison of different machine learning models for prediction of HIV status among young men who have sex with men

Data source	Models	AUC	Sensitivity	Specificity	Precision			Recall			F1-score		
					Class 0	Class 1	Weighted average	Class 0	Class 1	Weighted average	Class 0	Class 1	Weighted average
Original imbalance data	K-nearest neighbour	0.57	0.16	0.96	0.91	0.35	0.85	0.96	0.16	0.88	0.94	0.22	0.86
	Random forest	0.52	0.07	1.00	0.90	0.55	0.86	1.00	0.07	0.89	0.94	0.12	0.85
	Extreme gradient boosting	0.52	0.09	0.99	0.90	0.51	0.86	1.00	0.09	0.89	0.94	0.15	0.85
	AdaBoost	0.53	0.09	0.99	0.90	0.50	0.86	1.00	0.09	0.89	0.94	0.15	0.85
	Logistic regression	0.52	0.04	0.89	0.89	0.46	0.84	1.00	0.04	0.89	0.94	0.07	0.84
SMOTE processing data	K-nearest neighbour	0.56	0.18	0.95	0.90	0.30	0.83	0.95	0.18	0.86	0.93	0.23	0.85
	Random forest	0.72	0.72	0.72	0.96	0.25	0.88	0.72	0.72	0.72	0.82	0.37	0.77
	Extreme gradient boosting	0.72	0.73	0.72	0.96	0.24	0.88	0.72	0.73	0.72	0.82	0.36	0.77
	AdaBoost	0.72	0.73	0.72	0.96	0.24	0.88	0.72	0.73	0.72	0.82	0.36	0.77
	Logistic regression	0.70	0.69	0.71	0.95	0.24	0.87	0.71	0.69	0.71	0.81	0.36	0.76

Note: Class=1—young MSM with HIV; Class=0—young MSM without HIV.
F1-score, precision and recall with weighted average were used as evaluation metrics to account for class imbalance, as the data set used in HIV prediction was imbalanced.
AUC, area under the curve; MSM, men who have sex with men; SMOTE, Synthetic Minority Oversampling Technique.

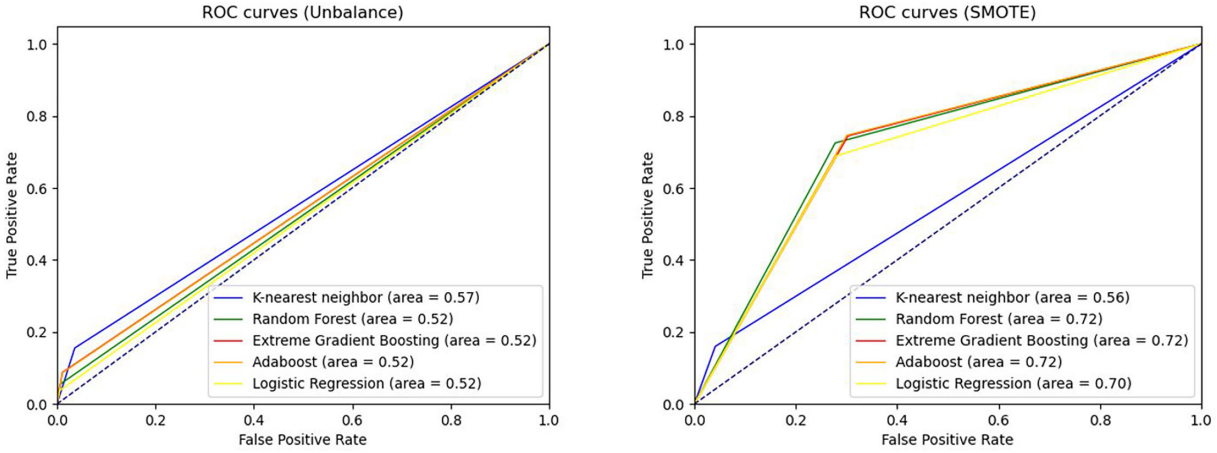

Figure 2 The receiver operating characteristic (ROC) curves of five models for predicting HIV infection for original unbalanced data (A) and SMOTE-processed data (B). Note: these models include k-nearest neighbour (KNN), random forest (RF), extreme gradient boosting (XGB), AdaBoost (ADA) and logistic regression (LR). SMOTE, Synthetic Minority Oversampling Technique.

periods of 2015–2017 (aOR 6.31, 95% CI 6.01 to 6.63) and 2018–2020 (aOR 2.51, 95% CI 2.41 to 2.62) had higher odds of being diagnosed with HIV infection compared with those in the period 2021–2022. Young MSM who had targeted HIV testing (aOR 3.23, 95% CI 2.99 to 3.49) were more likely to be diagnosed with HIV infection than those without routine HIV testing. The clients who had SSS as main insurance (aOR 2.26, 95% CI 2.01 to 2.53) and UCS (aOR 1.56, 95% CI 1.39 to 1.74) were more likely to be diagnosed with HIV infection compared with those in others. Moreover, residents of Bangkok had the highest OR of HIV infection compared with those residing in the Western region.

DISCUSSION

Our study shows that ML models can aid HIV epidemic control using real-world data. Among the evaluated models, XGB with SMOTE-processed data achieved the highest accuracy in predicting HIV infection among young MSM. SHAP analysis identified key risk factors: earlier calendar year, older age at diagnosis and targeted HIV testing. These findings support the use of ML for HIV prediction, targeted interventions and prevention planning in Thailand. Integrating ML with real-world data can enhance prediction accuracy, inform public health strategies and optimise prevention efforts for KPs.

In our study, we used electronic health record data from the UHC programme, which showed low precision after applying SMOTE, likely due to the low prevalence of HIV and high engagement in HIV testing. Similar findings from studies in the USA[25] and Denmark[15] highlight the challenge of class imbalance in predicting HIV infection. To address this, we applied a weighted average approach, adjusting class weights inversely to their frequencies. This improved sensitivity to rare HIV-positive cases while maintaining balance with precision. In health settings, minimising both false positives and false negatives is crucial. The F1-score with weighted average accounts for both errors, making it a better metric for assessing model performance in HIV prediction than accuracy alone.[26] Our model has the potential to support clinicians in identifying individuals at higher risk of acquiring

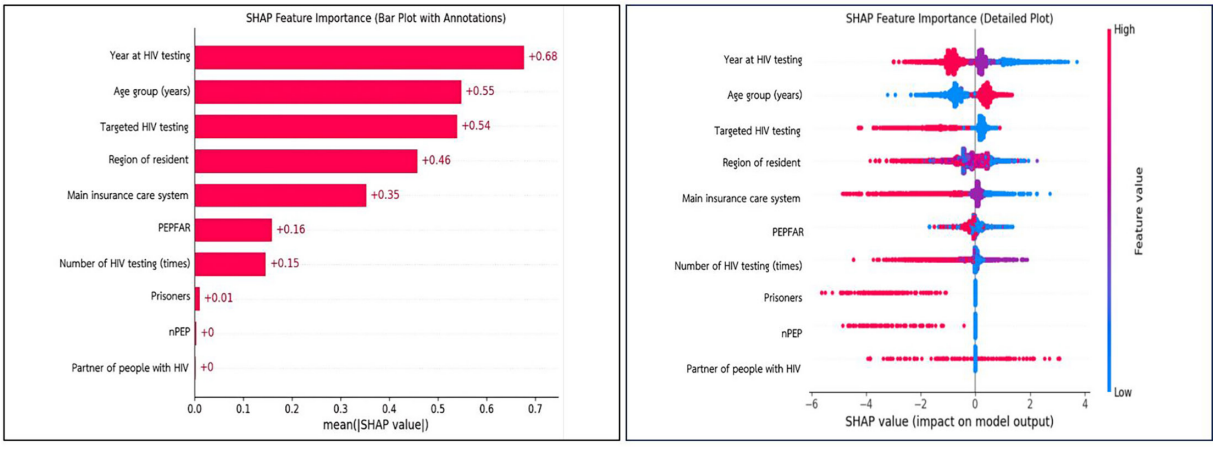

Figure 3 The importance of features in the extreme gradient boosting (XGB) model by using SHAP. nPEP, non-occupational postexposure prophylaxis; PEPFAR, President's Emergency Plan for AIDS Relief; SHAP, SHapley Additive exPlanations.

HIV and linking them to preventive services such as the pre-exposure prophylaxis registry programme.[27] Moving forward, real-world validation and optimisation of ML algorithms will be crucial to improving their practical application in public health settings. Our findings show that the XGB model achieved a weighted average F1-score of 0.77, demonstrating its ability to balance precision and recall while also achieving the highest AUC and sensitivity. These results are consistent with previous studies that highlight the effectiveness of ML algorithms in predicting HIV infection.[16] A study that applied ML approaches to predict HIV and sexually transmitted infections (STIs) among MSM in Australia reported that Gradient Boosting achieved the highest AUC for HIV prediction (76.3%), followed by XGB, RF, deep learning and LR. These results highlight the advantages of ML approaches over traditional LR models in predicting HIV among MSM.[17] More recently, a study from Zimbabwe also found that the XGB model demonstrated the highest performance in predicting HIV infection in the general population.[18] Similarly, our findings align with a study conducted among MSM in Zhejiang, China, from 2018 to 2020, which applied SMOTE to address data set imbalance. That study reported an HIV infection rate of 6% and identified the RF model as the best-performing algorithm (recall=0.775, and AUC=0.942) when compared with conventional LR models.[19] The usefulness of the SMOTE process for generating synthetic samples and addressing imbalanced biomedical data is further supported by findings from studies predicting HIV status in Danish registries.[15 28]

Additionally, our study observed an increasing trend in HIV testing among young MSM under the UHC programme in Thailand, accompanied by a decrease in the proportion of HIV infections during the study period. This decline is likely attributable to the effectiveness of the test-and-treat intervention and access to the PEPFAR programme in high-risk regions.[29 30] Our findings indicate that HIV prevalence among MSM during the study period was lower than the prevalence reported in previous studies from China.[19 31] Findings from conventional LR analysis further revealed that young MSM aged 20–24 years had higher odds of HIV infection, consistent with findings from studies conducted in China and Mozambique.[11 31] Moreover, the recent advancements in HIV testing and scaling up of treatment underscore the commitment to achieve better treatment coverage and higher long-term viral suppression rates among people with HIV in Thailand.[7 32] These significant factors were also reflected in the feature importance rankings identified in our study using the XGB model. Targeted efforts for MSM, the most affected group, are crucial to reducing HIV transmission and achieving global targets.

Our study supports the existing ML research that focused on predicting HIV infection. These studies collectively demonstrate the effectiveness of ML in detecting HIV infection among MSM through the real-world data sets. There were some limitations in our study. First, the inclusion of a limited number of variables, primarily demographic factors, restricted the utilisation of other important sexual behaviour factors such as condomless anal sex, substance abuse and history of STIs, which may have limited the depth of analysis in understanding the predictors for HIV infection among MSM. Second, ML models require large amounts of high-quality data and might not have exactly interpreted the association between outcome and predictors in detail as conventional regression analysis. Lastly, in low HIV positivity populations, ML models may have limited predictive power, especially with weak predictive features. Other sampling strategies, such as Adaptive Synthetic Sampling, were explored to oversample the minority class, but these resulted in lower accuracy. Additionally, adjusting decision thresholds (to better classify individuals at high risk for HIV), incorporating cost-sensitive learning (to minimise unnecessary testing and follow-up visits or prioritise detecting truer HIV-positive cases) and optimising the F1-score or balanced accuracy could further enhance the utility of the model in real-world HIV screening programmes. Incorporating ML in HIV prediction enhances disease understanding and supports public health goals with guiding targeted interventions.

CONCLUSION

The widespread application of ML in the medical domain, particularly in diagnosis and predictive classification, underscores its potential for enhancing healthcare outcomes. The XGB model and other ML techniques emerge as potentially effective tools for predicting HIV infection among young MSM in Thailand, enabling the implementation of timely interventions and tailored preventive measures.

Acknowledgements The data set was provided by the National Health Security Office and the Ministry of Health in Thailand. The content of this publication is solely the responsibility of the authors. This work (Grant No. RGNS 65 – 041) was partially financially supported by the Office of the Permanent Secretary, Ministry of Higher Education, Science, Research and Innovation, Thailand, from 2022 to 2024. The study was also supported by the Visiting Research Scholar Program, International SciKU Branding (ISB) grants, and the Department of Statistics, Faculty of Science, Kasetsart University. The content of this publication is solely the responsibility of the authors and does not necessarily represent the official views of the governments or institutions mentioned above.

Contributors KS, SP, TS, LK, WMH and ST created the study concept and study design. ST was responsible for data collection or oversaw programme implementation. KS, SP and ST prepared the data and conducted the analysis. ST and WMH advised on the analysis. ST and KS drafted the manuscript. ST, LK and WMH edited the manuscript. All authors critically reviewed the manuscript and approved the manuscript for submission. ST is guarantor, responsible for the overall content.

Funding ST was funded as a grantee (Grant No. RGNS 65 – 041) from the Office of the Permanent Secretary, Ministry of Higher Education, Science, Research and Innovation, Thailand, from 2022 to 2024.

Disclaimer Data were presented as oral presentation at the Asia-Pacific AIDS & Co-Infections Conference (APACC) 2024 in Hongkong on 27/06/2024-29/06/2024.

Map disclaimer The inclusion of any map (including the depiction of any boundaries therein), or of any geographical or locational reference, does not imply the expression of any opinion whatsoever on the part of BMJ concerning the legal status of any country, territory, jurisdiction or area or of its authorities. Any such

expression remains solely that of the relevant source and is not endorsed by BMJ. Maps are provided without any warranty of any kind, either expressed or implied.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval The study was approved by Kasetsart University Research Ethics Committee (KUREC-HSR65/023), Thailand. Consent was waived, and the NHSO de-identified the HIV database before analysis.

Provenance and peer review Part of a Topic Collection; Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request. No data are available. The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID iD

Sirinya Teeraananchai http://orcid.org/0000-0001-9100-2930

REFERENCES

1 UNICEF. Adolescent HIV prevention, 2023. Available: https://data.unicef.org/topic/hivaids/adolescents-young-people
2 UNAIDS. Fact sheet world aids day 2023: UNAIDS. 2023. Available: https://www.unaids.org/sites/default/files/media_asset/UNAIDS_FactSheet_en.pdf
3 UNAIDS. UNAIDS data 2019. 2019.
4 Johnston LG, Soe P, Widihastuti AS, *et al*. Alarmingly High HIV Prevalence Among Adolescent and Young Men Who have Sex with Men (MSM) in Urban Indonesia. *AIDS Behav* 2021;25:3687–94.
5 Division ATaSC. HIV/aids epidemic and status of prevention and alleviation efforts, bangkok, 2010–2014. Bangkok, Thailand, 2015.
6 Wongkanya R, Pankam T, Wolf S, *et al*. HIV rapid diagnostic testing by lay providers in a key population-led health service programme in Thailand. *J Virus Erad* 2018;4:12–5.
7 Chaivooth S, Bhakeecheep S, Ruxrungtham K, *et al*. The challenges of ending AIDS in Asia: outcomes of the Thai National AIDS Universal Coverage Programme, 2000-2014. *J Virus Erad* 2017;3:192–9.
8 Weir BW, Dun C, Wirtz AL, *et al*. Transactional sex, HIV and health among young cisgender men and transgender women who have sex with men in Thailand. *Ann Epidemiol* 2022;72:1–8.
9 HUB HI. Epidermic: HIV INFO HUB, Available: https://hivhub.ddc.moph.go.th/epidemic.php
10 Gangcuangco LMA, Tan ML, Berba RP. Prevalence and risk factors for HIV infection among men having sex with men in Metro Manila, Philippines. *Southeast Asian J Trop Med Public Health* 2013;44:810–7.
11 Ribeiro Banze Á, Muleia R, Nuvunga S, *et al*. Trends in HIV prevalence and risk factors among men who have sex with men in Mozambique: implications for targeted interventions and public health strategies. *BMC Public Health* 2024;24:1185.
12 Hessou PHS, Glele-Ahanhanzo Y, Adekpedjou R, *et al*. Comparison of the prevalence rates of HIV infection between men who have sex with men (MSM) and men in the general population in sub-Saharan Africa: a systematic review and meta-analysis. *BMC Public Health* 2019;19:1634.
13 Ditangco R, Mationg ML. HIV incidence among men who have sex with men (MSM) in Metro Manila, the Philippines: A prospective cohort study 2014-2018. *Medicine (Baltimore)* 2022;101:e30057.
14 Kimani M, van der Elst EM, Chiro O, *et al*. PrEP interest and HIV-1 incidence among MSM and transgender women in coastal Kenya. *J Int AIDS Soc* 2019;22:e25323.
15 Ahlström MG, Ronit A, Omland LH, *et al*. Algorithmic prediction of HIV status using nation-wide electronic registry data. *EClinicalMedicine* 2019;17:100203.
16 Fieggen J, Smith E, Arora L, *et al*. The role of machine learning in HIV risk prediction. *Front Reprod Health* 2022;4:1062387.
17 Bao Y, Medland NA, Fairley CK, *et al*. Predicting the diagnosis of HIV and sexually transmitted infections among men who have sex with men using machine learning approaches. *J Infect* 2021;82:48–59.
18 Birri Makota RB, Musenge E. Predicting HIV infection in the decade (2005-2015) pre-COVID-19 in Zimbabwe: A supervised classification-based machine learning approach. *PLOS Digit Health* 2023;2:e0000260.
19 He J, Li J, Jiang S, *et al*. Application of machine learning algorithms in predicting HIV infection among men who have sex with men: Model development and validation. *Front Public Health* 2022;10:967681.
20 UNAIDS. Gender disparities related to hiv emerge in adolescence 2023. 2023. Available: https://data.unicef.org/topic/hivaids/adolescents-young-people
21 Team GL. Hyperparameter tuning with gridsearchcv. 2024. Available: https://www.mygreatlearning.com/blog/gridsearchcv
22 Berrar D, Cross-Validation. ResearchGate. 2018;1:542–5.
23 Mb AW, Pluskiewicz W. The study of preprocessing methods' utility in analysis of multidimensional and highly im. *ResearchGate* 2016.
24 SMLaS-I L. A Unified Approach to Interpreting Model Predictions. *Neural Information Processing Systems* 2017.
25 Nethi AK, Karam AG, Alvarez KS, *et al*. Using Machine Learning to Identify Patients at Risk of Acquiring HIV in an Urban Health System. *J Acquir Immune Defic Syndr* 2024;97:40–7.
26 Saito T, Rehmsmeier M. Preprocessing for imbalanced data: The weighted average F1 score. *Int J Data Sci Anal* 2015;3:1–18.
27 Marcus JL, Sewell WC, Balzer LB, *et al*. Artificial Intelligence and Machine Learning for HIV Prevention: Emerging Approaches to Ending the Epidemic. *Curr HIV/AIDS Rep* 2020;17:171–9.
28 Nakamura M, Kajiwara Y, Otsuka A, *et al*. LVQ-SMOTE - Learning Vector Quantization based Synthetic Minority Over-sampling Technique for biomedical data. *BioData Min* 2013;6:16.
29 Department of Disease Control MoPH. Thai national aids program review. Thailand, 2022.
30 Pattanasin S, van Griensven F, Mock PA, *et al*. Recent declines in HIV infections at Silom Community Clinic Bangkok, Thailand corresponding to HIV prevention scale up: An open cohort assessment 2005-2018. *Int J Infect Dis* 2020;99:131–7.
31 Zhou J, Yang L, Ma J, *et al*. Factors Associated With HIV Testing Among MSM in Guilin, China: Results From a Cross-Sectional Study. *Int J Public Health* 2022;67:1604612.
32 Teeraananchai S, Boettiger DC, Lertpiriyasuwat C, *et al*. The impact of same-day and rapid ART initiation under the Universal Health Coverage programme on HIV outcomes in Thailand: a retrospective real-life cohort study. *J Int AIDS Soc* 2025;28:e26406.