Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Distributed CatBoost Approach with Weighted Aggregation for STI/HIV Risk Prediction

Zhaohui Tang

zhaohui.tang@unisq.edu.au

University of Southern Queensland

Yan Li

University of Southern Queensland

Abdulkadir Sengur

Fırat University

U. Rajendra Acharya

University of Southern Queensland

Research Article

Keywords: HIV/STI prediction, Distributed CatBoost Prediction, Leave-One-Country-Out validation

Posted Date: July 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10354103/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A Distributed CatBoost Approach with Weighted Aggregation for STI/HIV Risk Prediction

Zhaohui Tang[1*], Yan Li[1], Abdulkadir Sengur[2] and U. Rajendra Acharya[1]

[1]School of Science, Engineering and Digital Technologies, University of Southern Queensland, Toowoomba, Australia

[2]Electrical-Electronics Engineering Department, Technology Faculty, Firat University, Elazig, Turkey

[1]zhaohui.tang@unisq.edu.au, yan.li@unisq.edu.au, rajendra.acharya}@unisq.edu.au,

[2]ksengur@gmail.com

[*]Corresponding author

Abstract

STIs (Sexually Transmitted Infections) and HIV (Human Immunodeficiency Virus) remain significant health issues globally because of their widespread occurrence, often asymptomatic development, and social and medical implications. In this paper, we present an innovative approach for predicting STI and HIV risk with user privacy achieved, using structured tabular datasets from the Demographic and Health Surveys Program based on the distributed CatBoost algorithm. The presented dataset contains 168,459 observations collected across eight different countries, with variables related to demographics, behavior, and attitudes toward STI and HIV status. Instead of the classical centralized approach, where all training is done on the entire dataset, our framework utilizes multiple independent training sessions of CatBoost models on different subsets of data with subsequent aggregation of results with a reliability-weighted method. For increased robustness, the framework also employs local ensemble modeling, validation-based threshold adjustment, and prediction confidence calculation. Evaluation of our model is performed using a Leave-One-

Country-Out (LOCO) methodology. As observed from the experimental outcomes, the proposed approach shows high predictability in terms of classification of both HIV and STIs, yet with privacy-preservation achieved. For the prediction of HIV cases, the proposed framework obtained an average AUC score of 0.9850 and an average accuracy score of 0.9537. When it comes to STI prediction, the average AUC and accuracy scores obtained were 0.9396 and 0.9117, respectively. Further investigation through t-SNE analysis and feature correlation analysis highlighted heterogeneity in different countries and confirmed that the difference in performance at the level of each country is reasonable to be interpreted. In summary, the proposed framework can be considered as a viable and robust solution to the problem of prediction of HIV and STIs in a distributed environment, and thus a substitute for centralized ML models in the applications where user privacy is a concern.

Keywords: HIV/STI prediction, Distributed CatBoost Prediction, Leave-One-Country-Out validation.

1. Introduction

STIs and HIV remain significant public health problems across the world. According to World Health Organization (WHO), hundreds of millions of people acquire new curable STIs per year on a global scale [1]. In addition, in 2024, 1.3 million people acquired HIV infection around the globe [2]. The problem seems even more severe in low- and middle-income countries and Sub-Saharan Africa where HIV infection puts considerable strain on individuals, populations, and health care systems [2]. There are certain populations which face greater consequences related to these issues – for example, men who have sex with men, sex workers, transgender persons, and adolescent girls and young women [3]. Their increased vulnerability stems not just from risky behaviors but from wider structural barriers, stigmatization, discrimination, criminalization, and gendered disparities in access [3]. Untreated STIs can lead to health complications such as infertility or pelvic inflammatory disease. Some STIs can also make people more susceptible to acquiring HIV [1]. It can be difficult to diagnose and treat STIs. The main reason behind this difficulty lies in the fact that many infections are asymptomatic. As a result, they can continue spreading even when no symptoms present themselves [1]. Many countries lack the means for diagnosing infections using laboratories and

continue relying on syndromic management instead [4]. Although this approach remains practical in some contexts, it may miss asymptomatic infections and can contribute to over-treatment or under-treatment [4]. At the same time, people from marginalized populations often face additional barriers to screening and treatment, including stigma, fear of judgment, discrimination, and limited access to appropriate health services [3].

In response to these limitations, much recent scholarly attention has been drawn to the role that artificial intelligence (AI) and machine learning (ML) could play in the context of predicting STI and HIV risks. The potential advantages associated with these models are numerous, from the ability to analyze massive amounts of data, including sociodemographic factors, behaviors, and clinical data to more efficient ways of identifying patients who are at a high risk, as well as providing them with an opportunity for early intervention. A wide variety of supervised machine learning algorithms have been considered for this purpose, including Random Forest (RF), Gradient Boosting Machine (GBM), XGBoost, Support Vector Machines (SVM), and deep learning models. Recent reviews have reported strong overall predictive performance for HIV and STI risk modeling, and several individual studies have shown particularly high results for Random Forest and ensemble-based approaches in specific datasets [5]. At the same time, the growing number of studies is reflected in the creation of a range of predictive tools designed not only to predict but also to assess risks for patients in practice [5]. Latt et al. [6] performed a quasi-experimental pre-post study to assess anxiety levels and testing intentions linked to an AI-powered risk assessment test, MySTIRisk, relative to a typical sexual health webpage. According to their results, an AI-based solution induced clinically significant anxiety in patients; the odds ratio (OR) adjusted to be 3.12. Nevertheless, the tool maintained acceptable levels at 92.0%, and testing intentions remained high at 87.3%. Ji et al. [7] proposed an automated prediction model for determining global risk of STI/HIV in gay men. In this context, an ensemble model based on GBM, RF, XGBoost and ML algorithms was utilized, combined with the Synthetic Minority Oversampling Technique (SMOTE) to deal with the class imbalance problem. By applying Demographic and Health Surveys Program (DHSP) data for eight countries along with a 75% training set and 25% testing one supplemented with 8-fold cross-validation, they achieved accuracy at 99%

and AUC at 0.99; the highest performance was demonstrated by RF. Finally, Latt et al. [8] assessed the behaviors related to conducting an HIV test with the help of Decision Tree (DT), RF, SVM, and Logistic Regression (LR) as well as SMOTE to solve class imbalanced data problems. Their dataset consists of five cross-sectional waves of South African National HIV Prevalence, Incidence, Behavior and Communication Survey (SABSSM) surveys from 2002 to 2017. Using the 80/20 ratio in training and testing datasets and 5-fold cross-validation, the highest performance on the 2002 dataset was observed using RF with accuracy at 80.98%, precision at 81.51%, F1-score at 80.30%, and Area Under the Curve (AUC) at 88.31%. Nadarzynski et al. [9] introduced a culturally sensitive Chatbot-Assisted Self-Assessment (CASA) intervention to explore disclosure of health-related information. Based on data from 1,287 ethnically diverse survey respondents in the UK and 41 follow-up interviews, the authors found that previous chatbot experience and more positive attitudes toward chatbots were strong predictors of disclosing sensitive health information, and the predictive model achieved an R^2 value of 0.43. Nguyen et al. [10] proposed a RF Federated Learning (RFFL) framework for predicting STI/HIV risk while preserving privacy by avoiding centralized data sharing. Their method also incorporated SMOTE to handle class imbalance and Pearson correlation for feature selection. Using 168,459 records from 8 countries in the Demographic and Health Surveys (DHS) Program, with data distributed locally and aggregated dynamically at the server level, the model achieved 93% accuracy and an AUC of 0.97. Siyamayambo et al. [5] conducted a systematic review and meta-analysis to examine how supervised machine learning methods, including RF, XGBoost, and LR, have been used in the diagnosis and management of STIs and HIV. The present review is based on 20 primary studies carried out in 20 Sub-Saharan African countries during the period from 2015 to May 2024 and reveals the average F1-score of the examined models as 0.79, which means that they can provide promising results for sexual medicine. The comprehensive review provided by Smerina et al. [11] examines ways in which artificial intelligence together with wearable devices may be used to improve sexual medicine. The research focuses on applying machine learning algorithms and deep learning together with using wearable technologies including penile tumescence sensing systems and pelvic floor tracking devices for diagnostic, treatment, and monitoring purposes. The review is based on information obtained from scientific publications, web pages, as well as conference papers

and highlights the growing significance of the proposed methods paying special attention to their safety, feasibility, and increasing precision. Yu et al. [12] proposed a deep learning algorithm based on the Socio-ecological model (SEM) for the prediction of STI in both males and females. They compared five models – FT-Transformer, GatedTabTransformer, GANDALF, XGBoost, and Logistic Regression. With 10,159 records from the UK National Survey of Sexual Attitudes and Lifestyles (Natsal-3), split into 90% training and 10% testing sets, the FT-Transformer achieved the strongest results, reaching 87.0% accuracy and 0.843 AUC for males, and 87.5% accuracy and 0.879 AUC for females.

In this paper, we present a distributed CatBoost-based [15-17] framework for STI/HIV risk prediction, designed to make use of locally learned predictive patterns with user privacy preserved while still producing a single global risk estimate. Figure 1 shows the flowchart of the proposed method. Rather than training one pooled model on all available data, our approach learns separate local models on different data partitions and combines their outputs through a weighted probability aggregation strategy. In this way, the proposed framework aims to preserve local predictive structure and, at the same time, improves the robustness of the final decision. The proposed methodology includes data preprocessing, local models training, validation-based weighting, global probability aggregation, classification threshold calibration, and uncertainty estimation. After the preprocessing stage, CatBoost models are trained separately within each local partition. To avoid randomness, it is decided not to rely only on one model in each partition but rather to have an ensemble of several CatBoost models with different random seeds per partition. Output probabilities from all models inside local ensembles are averaged to produce a more stable output probability. Once probabilities for each sample from local partitions are produced, no assumption that they are equally valid is made. In particular, partitions that are supported by more data and that achieve stronger validation AUC values are given greater influence in the final prediction stage. The global risk score is then computed by combining the weighted local probabilities, resulting in a single probability estimate for each sample. To produce the final binary decision, the optimal threshold for each data point probability is found using a validation split instead of using the default threshold value. Such a strategy allows the decision rule to be calibrated according to the data distribution and helps balance false-positive and false-negative errors, which

is particularly important in imbalanced STI/HIV risk prediction tasks.. Finally, the uncertainty estimation module is applied to account for variability between outputs from local models in terms of confidence about final prediction. The proposed approach is tested on STI/HIV tabular dataset and proves itself to be more robust than a simple concatenation of data across partitions.

The main contributions of this work are as follows:

1. In our proposed framework for STI/HIV risk prediction, we utilize an ensemble of CatBoost that learns independently within each partition but aggregates results using probability averaging rather than combining the datasets into one.
2. We devise a partition-specific ensemble method, which uses several CatBoost algorithms, each trained using a different seed, to obtain a better partition-specific model.
3. To improve the reliability of global predictions from partitions, we introduce a validation-based weighted aggregation technique, where more weight is assigned to partitions that perform better and have more data available.
4. We consider the incorporation of threshold calibration and uncertainty estimation in the proposed model, which allows for a more robust decision rule.

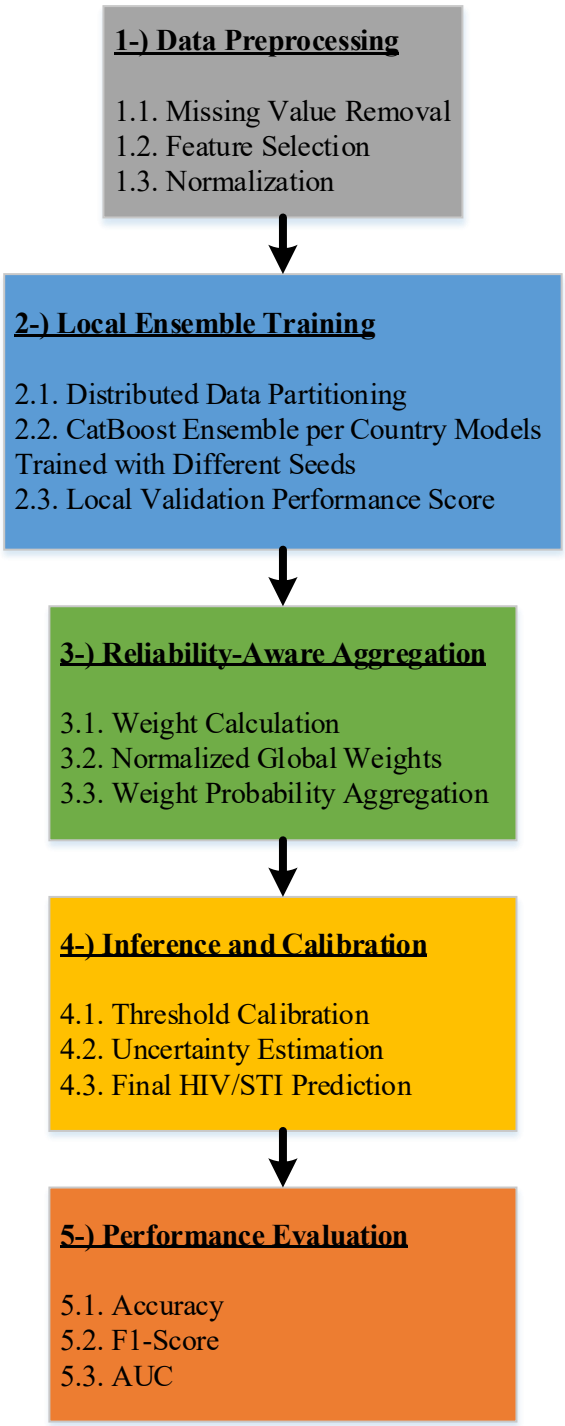


Figure 1 The illustration of the proposed method

These contributions are important in the STI/HIV prediction field because the underlying datasets are typically heterogeneous, distributed across different populations, and often subject to varying levels of

quality and sample size. Under such conditions, reliability-aware weighted aggregation helps ensure that stronger and better-supported partitions contribute more to the final prediction, improving robustness. Threshold calibration is also important because the preferred balance between sensitivity and specificity may differ depending on the screening or intervention context. In addition, predictive uncertainty estimation provides an interpretable measure of confidence, which is especially valuable in health-related applications where predictions may influence further testing, prevention strategies, or clinical attention.

2. Material and Methods

The rationale behind the proposed framework is grounded on a simple but essential principle that when data samples are distributed across several partitions, the same pooled model does not always reflect the predictive structure of the problem adequately [13, 14]. Combining all observations into one single training set simplifies computations but requires the assumption that there exists only one decision rule capable of capturing all significant patterns in the given dataset. In many real settings, however, different partitions may show differences in class balance, feature distributions, or the strength of associations between predictors and outcomes. When such variation is ignored, part of the useful local structure may be lost. As a result, the current methodology does not assume a priori that a single global model will provide the same description of all data samples. Instead, multiple learners are estimated locally, and their results are aggregated in a controlled way. The goal is not to sacrifice the process of global prediction but to achieve this by creating a global prediction based on local predictions. For this reason, the proposed methodology is particularly suitable for working with structured tabular health data, since heterogeneity in partitions may occur despite the similarity in the global prediction problem. The CatBoost algorithm is chosen as the base learner because of its relevance to the tabular classification task and capability of modeling nonlinear interactions [15]. At the same time, CatBoost is also robust enough to cope well in case of complex statistical dependencies in the data. Unlike this, in the suggested algorithm, CatBoost is not estimated once as a central predictor [16]. Instead, it is applied to each partition to ensure that each local estimator captures relevant statistical properties of the data.

2.1. CatBoost as a Local Predictor

Let M_k denote the CatBoost model trained on data partition k [17]. For an input sample x, the local probability estimate can be written as;

$$p_k(x) = M_k(x)$$

where $p_k(x)$ denotes the predicted probability that x belongs to the positive class. Output expressed in probability form is crucial since this allows comparison, combination, and scaling among local models. In this paradigm, local models generate their risk estimates depending on how the data was distributed when training them. Therefore, the technique treats local models as independent prediction mechanisms rather than subdivisions of one global classifier.

2.2. Local Ensemble Stabilization

Using one fitted model for each partition may create unnecessary variance, particularly if the local samples are small and there is slight variation in the fitted solution due to the stochastic nature of training [18]. In order to avoid the creation of this type of variance, the CatBoost ensemble method is used with multiple models in each partition [19].

If $M_{k,l}$ denotes the l-th model trained on partition k, its probability output for sample x can be expressed as

$$p_{k,l}(x) = M_{k,l}(x)$$

Assuming that L local models are trained for partition k, the partition-level prediction is obtained by averaging their outputs:

$$\bar{p}_k(x) = \frac{1}{L}\sum_{l=1}^{L} p_{k,l}(x)$$

The resulting quantity, $\bar{p}_k(x)$, serves as the stabilized probability estimate for partition k. The local averaging process fulfills an important theoretical role. The effect of each individual model is reduced, and this enables a prediction to be made based on a consensus of different local learners. This means that the subsequent aggregation step is more robust to local variability and more dependent on predictable structure.

2.3. Reliability-Aware Weighting of Local Predictions

Once the probabilities have been stabilized within each partition, the next issue that comes up concerns whether all partitions should have the same weight in contributing to the decision [20, 21]. According to the model, it would be wrong for all partitions to have the same contribution to the decision process. Some partitions contain more training cases, which allows for a better-fitted model, while others contain models that fit validation data more accurately.

Let n_k denote the number of training samples in partition k, and let q_k denote its validation quality score, measured by validation AUC. The unnormalized weight assigned to partition k is defined as

$$w_k = n_k {q_k}^2$$

This expression combines data support and predictive quality in a single quantity. The inclusion of $n_k$reflects the idea that larger training sets often yield more stable local estimators. The term ${q_k}^2$ reflects the idea that empirical validation performance should matter when deciding how much trust to place in a local predictor. Squaring the quality term makes the weighting more selective, so that stronger local models receive a clearer advantage over weaker ones.

To ensure that the local contributions form a proper weighted average, the weights are normalized as

$$\alpha_k = \frac{w_k}{\sum_{r=1}^{K} w_r}$$

where K is the number of partitions used in computing the final result. The normalized weights provide a measure of the importance of each local predictor towards the output variable. Consequently, the method does not fall into the trap of making all partitions equally important or considering only one optimal partition among all partitions.

2.4. Global Probability Estimation

After predicting probabilities for each partition and calculating the corresponding weights, a global probability is computed through weighted average [22]. The global predicted probability for a certain input x can be expressed as;

$$P(x) = \sum_{k=1}^{K} \alpha_k \, \bar{p}_k(x)$$

This formula forms the key prediction rule for the algorithm. It produces one overall risk score without losing the local information that was acquired earlier. Hence, the final value is neither an average nor a prediction produced by only one classifier that is pooled together. Rather, it is a combination of different probabilistic predictions obtained from various regions within the data. From a theoretical perspective, this model may be viewed as a mixture-of-experts model, whereby every partition provides a local view of the input data and the algorithm decides to what degree this viewpoint should affect the final outcome [23]. This method has the advantage of being applied to heterogeneous environments in which the goal is not fitting one particular function but combining several heterogeneous predictive models into one final output.

2.5. Threshold Selection and Final Classification

The result of the aggregation step provides a continuous probability value. To obtain a discrete decision, one needs to specify a classification threshold τ [24]. Although a hard-coded classification threshold of 0.5 is commonly used for ordinary cases, this approach is not always suitable, especially for cases where there is an imbalance in the classes of interest or where the cost of making false positives errors differs from false negatives errors [25]. Therefore, in this context, the optimal threshold is selected based on the results obtained during the validation process. The class prediction at a particular threshold τ is given by;

$$\hat{y}(x) = 1, if\ P(x) \geq \tau;\ 0, if\ P(x) < \tau$$

The optimal threshold is then chosen as

$$\tau^* = arg\ max\ F1(\tau), \tau \in T$$

where T denotes the range of candidate threshold values considered during validation. This calibration process adds an extra layer of flexibility to the model. Instead of taking the probability to label relationship as static, it ensures that the decision function is adjusted according to the predictive requirements of the task. Since F1-score combines both precision and recall, maximizing this value helps ensure that the classifier does not adopt overly cautious or overly liberal boundaries.

2.6. Predictive Uncertainty

Apart from providing a final probability estimate, the system provides a measure of the uncertainty involved in making predictions [26]. The uncertainty is intrinsically part of the prediction process due to the generation of the prediction from a combination of outputs of several models rather than a single score [27]. If the different models provide similar estimates of the probability for a certain input, then the prediction itself is considered stable. Otherwise, the uncertainty is considered high.

Let $\{p_m(x)\}_{m=1}^{P}$ denote the collection of all member-level probability predictions contributing to sample x, where R is the total number of model outputs involved in the calculation. The mean predictive probability is

$$\mu(x) = \frac{1}{R}\sum_{m=1}^{R} p_m(x)$$

and the corresponding predictive standard deviation is

$$\sigma(x) = \sqrt{[\frac{1}{R}\sum_{m=1}^{R} p_m(x) - \mu(x)^2]}$$

The quantity $\sigma(x)$ summarizes the degree of disagreement among model outputs. A higher number implies more spread and consequently more uncertainty. However, this does not change the ultimate conclusion about the class to which the data belongs. Nevertheless, it improves the understanding of the outcome. Two datasets can generate almost identical results for the prediction probability but have different levels of uncertainty.

3. Experimental works and results

In this study, for model evaluation purposes, a LOCO approach was considered. That is, in each fold of the experiment, one country served only for model testing, whereas other countries' data were used as training samples for the construction of the federated model. As can be seen, the proposed approach makes it possible to test the model on the unseen country and thus obtain reliable results that will reflect its capability of working with various populations. Besides, within the outer LOCO split, an inner split was conducted by

which 15% of each country's data in the training subset were allocated for model validation. In other words, the validation subset could not serve as a test set but was used only to control model functioning during training and find the optimal classification threshold. For features pre-processing, 13 features were extracted using the RF-based algorithm, while robust normalization was performed in order to mitigate the effect of outliers. Finally, unlike in the original problem setting where the decision threshold was equal to 0.5, it was varied in the range from 0.20 to 0.80 with the step equal to 0.01, and the optimal threshold resulting in the best F1 score on the validation data was chosen.

3.1. Dataset

The dataset employed in this research was sourced from the DHS Program database and comprised of structured tabulated data gathered from eight countries for the purpose of predicting the risk of HIV and STI [28]. From the reference paper, it is stated that there were 168,459 records in the dataset, which were mainly focused on demographic and behavioral attributes of men with binary outcome labels of HIV/STI [10]. Prior to developing the model, the data underwent preprocessing stage where unnecessary variables such as those that are highly incomplete or have low contribution in terms of informativeness were removed, while missing value records were excluded. Moreover, overlapping variables were determined via Pearson correlation analysis. Lastly, feature importance was ranked using RF algorithm to determine the best predictors. At the end of the preprocessing stage, the model was constructed based on the 13 most important predictors, which include demographic, behavioral, and attitudinal attributes such as age, educational attainment, wealth index, residence location, condom use behavior, marriage, age at first sexual intercourse, recent sexual activity, sexual partner count, exchange of goods for sex, and attitude towards condom usage during STIs. The HIV and STI records per country can be seen on Figure 2 (a) and (b), respectively.

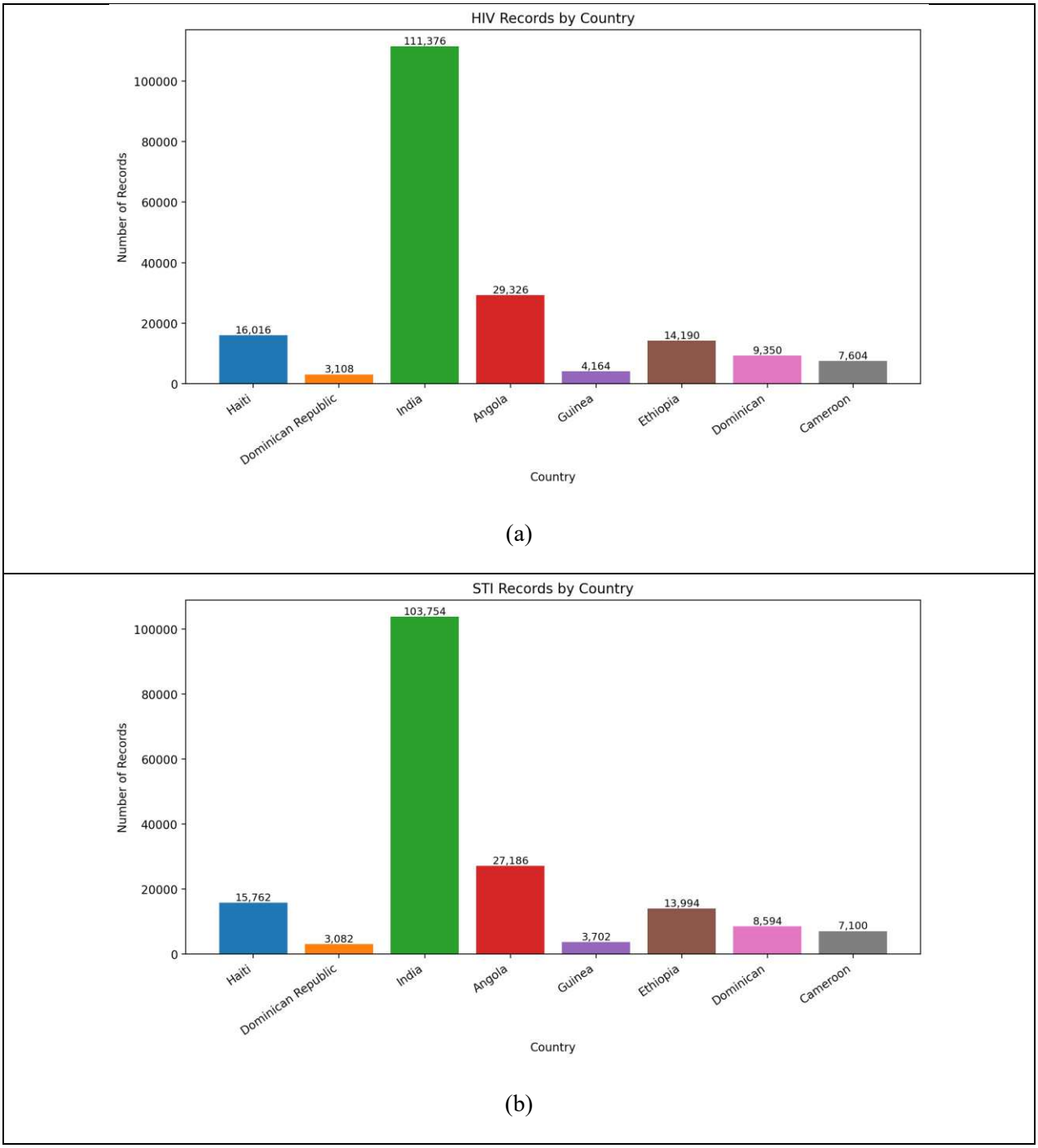
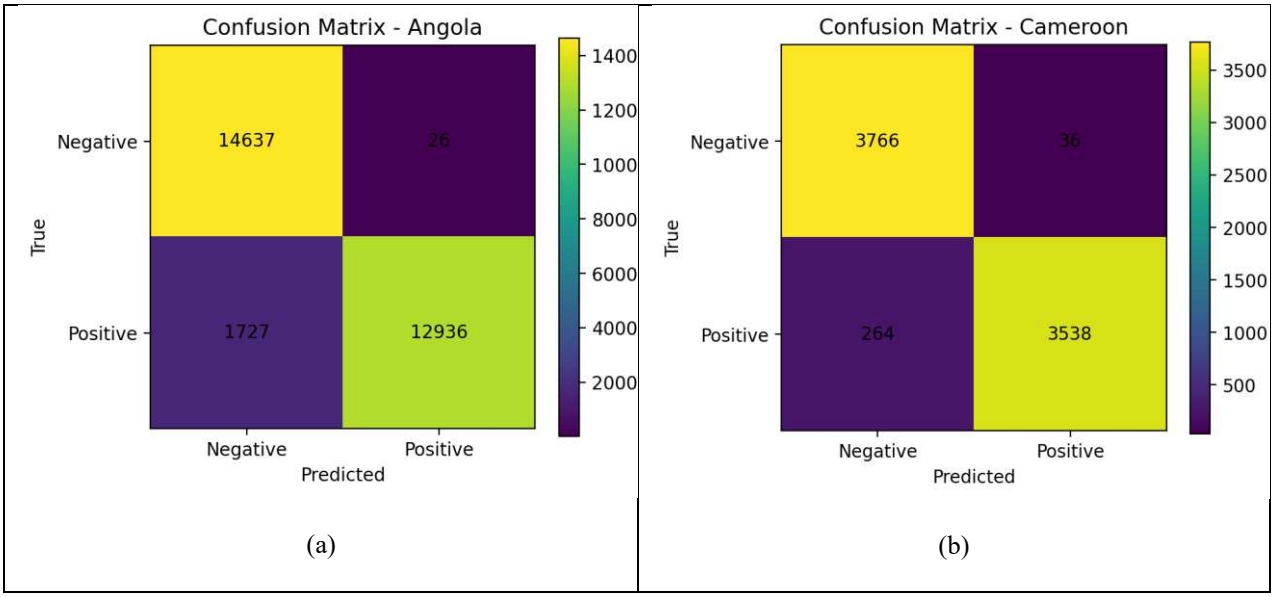


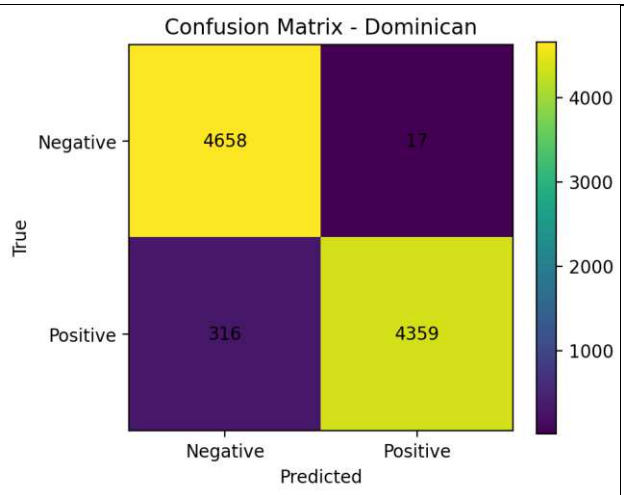
Figure 2 The number of records for each country, (a) The HIV records for each country, (b) The STI records for each country.

3.2. Results on HIV prediction

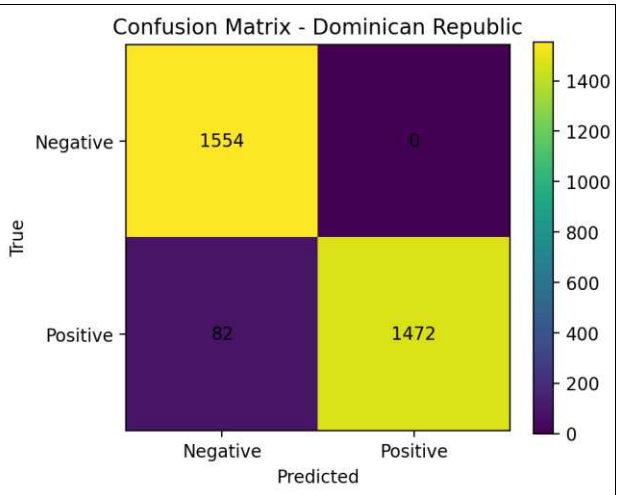
The initial experiment was conducted on HIV prediction and the obtained confusion matrixes, ROC, PR and Brier curves for each country were given in Figures 3, 4, 5 and 6, respectively. Table 1 also shows

the calculated performance evaluation metrics for each country. Figure 3 provides a country-level view of the classification results through the confusion matrices. Across all countries, most samples were classified correctly, as shown by the large values on the main diagonal. For example, in India, the model correctly identified 55,676 negative cases and 48,982 positive cases, while misclassifying only 12 negatives as positive and 6,706 positives as negative. A similar pattern was observed in Angola, where 14,637 negatives and 12,936 positives were correctly classified, compared with 26 false positives and 1,727 false negatives. The results for Ethiopia show that there were 7,095 negatives and 5,734 positives that the model accurately classified, with no false positives but with 1,361 false negatives. Similarly, in Cameroon, the algorithm was able to classify 3,766 negatives and 3,538 positives, although 36 positives were wrongly classified, as well as 264 negatives. The pattern continues for Haiti, where 7,995 negatives and 7,403 positives were correctly classified, although 13 positives were misclassified, as well as 605 negatives.

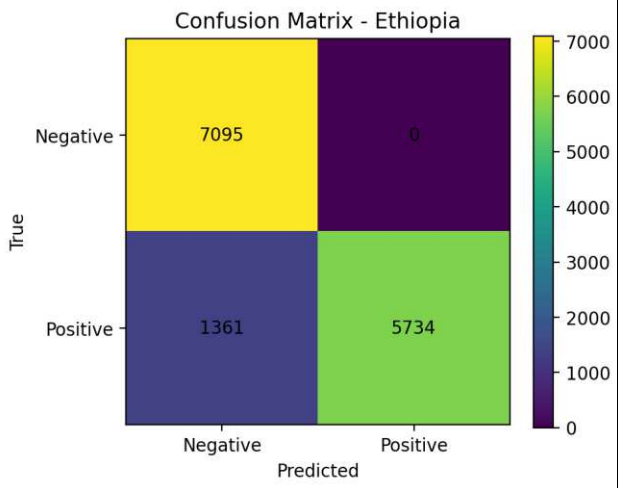



Confusion Matrix - Dominican
Negative
Positive
True
4658
17
316
4359
Negative
Positive
Predicted
4000
3000
2000
1000

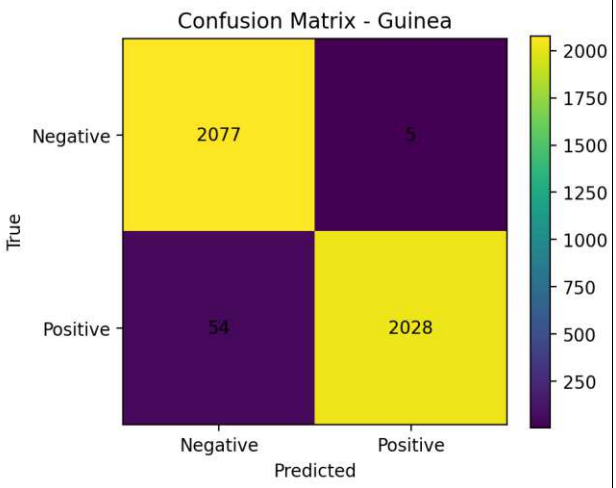
(c)


Confusion Matrix - Dominican Republic
Negative
Positive
True
1554
0
82
1472
Negative
Positive
Predicted
1400
1200
1000
800
600
400
200
0

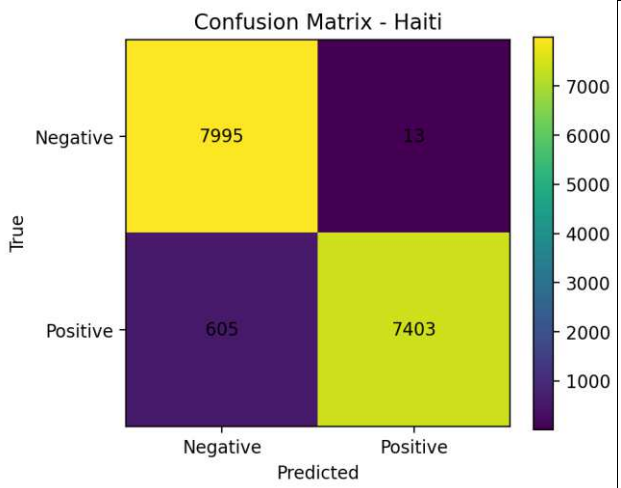
(d)


Confusion Matrix - Ethiopia
Negative
Positive
True
7095
0
1361
5734
Negative
Positive
Predicted
7000
6000
5000
4000
3000
2000
1000
0

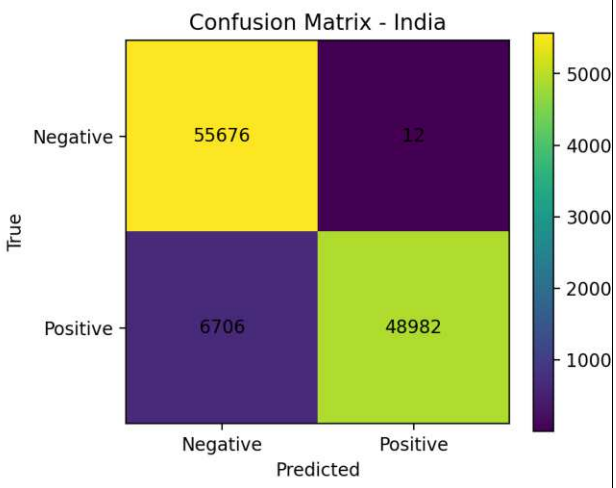
(e)


Confusion Matrix - Guinea
Negative
Positive
True
2077
5
54
2028
Negative
Positive
Predicted
2000
1750
1500
1250
1000
750
500
250

(f)


Confusion Matrix - Haiti
Negative
Positive
True
7995
13
605
7403
Negative
Positive
Predicted
7000
6000
5000
4000
3000
2000
1000

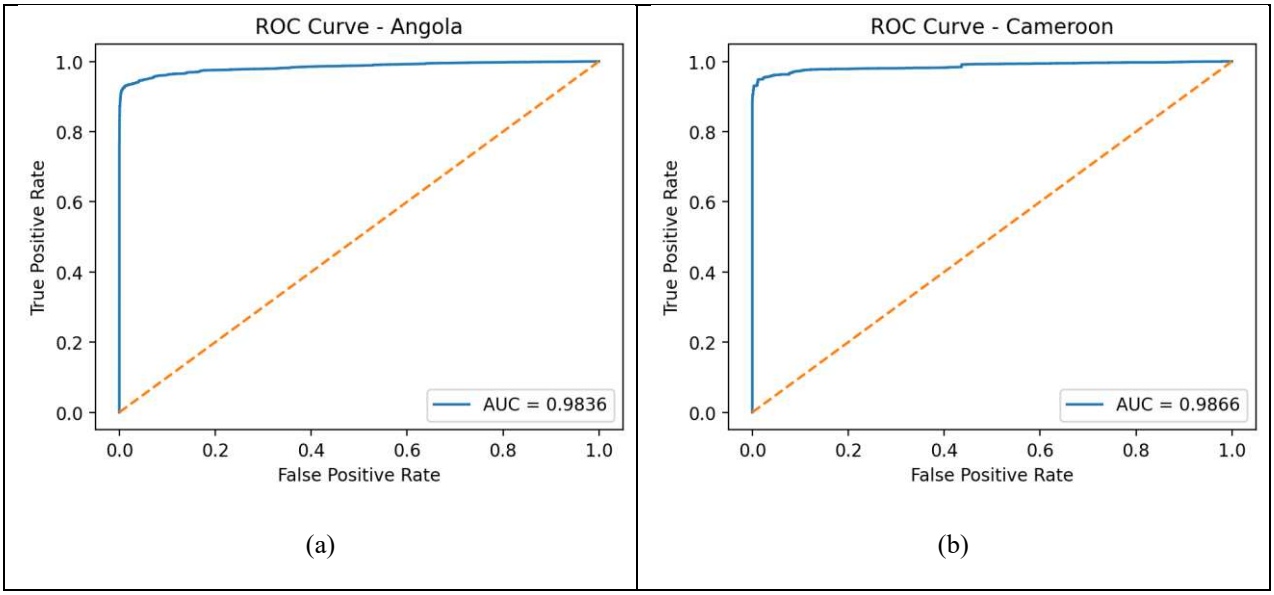
(g)

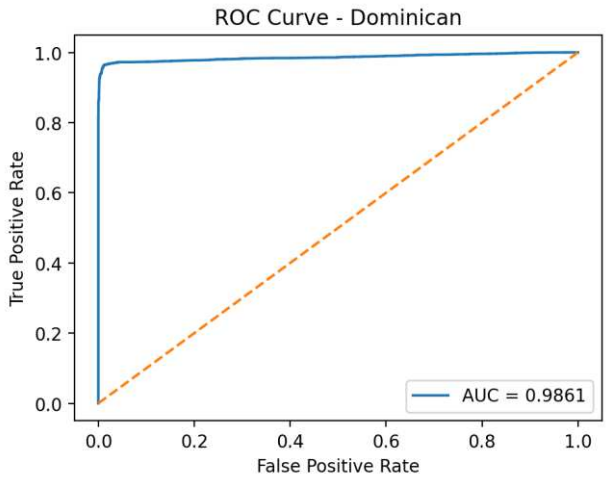

Confusion Matrix - India
Negative
Positive
True
55676
12
6706
48982
Negative
Positive
Predicted
5000
4000
3000
2000
1000

(h)

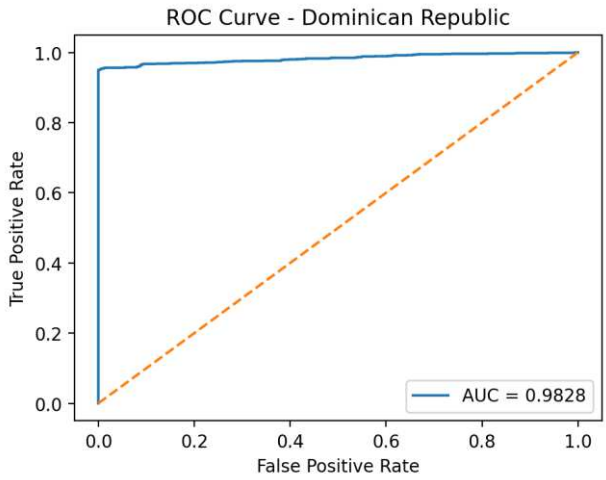
Figure 3 Obtained confusion matrixes for each country (a) Angola, (b) Cameroon, (c) Dominican, (d) Dominican Republic, (e) Ethiopia, (d) Guinea, (g) Haiti, and (h) India

Figure 4 shows the obtained ROC curves on HIV prediction for each country. As seen in the figure, the model performed strongly in all countries, but the results were not exactly the same across settings. Guinea produced the highest AUC at 0.9929, followed very closely by Haiti at 0.9910 and India at 0.9897. Cameroon, Dominican, Angola, and Dominican Republic also showed very strong discrimination, with AUC values of 0.9866, 0.9861, 0.9836, and 0.9828, respectively. Ethiopia had the lowest AUC among the eight countries at 0.9676, although this still reflects a high level of classification performance. When the countries are compared directly, the overall pattern suggests that the framework remained highly stable across different local datasets, while still reflecting some country-level variation. Guinea, Haiti, and India stood out as the strongest-performing cases, with ROC curves that stayed closest to the upper-left corner and therefore indicated near-ideal separation between classes. In contrast, Ethiopia showed a slightly weaker curve than the others, which suggests that classification was somewhat more difficult in that setting. The remaining countries fell into a relatively narrow middle range, indicating that the model generalized well across most partitions without any major drop in performance.

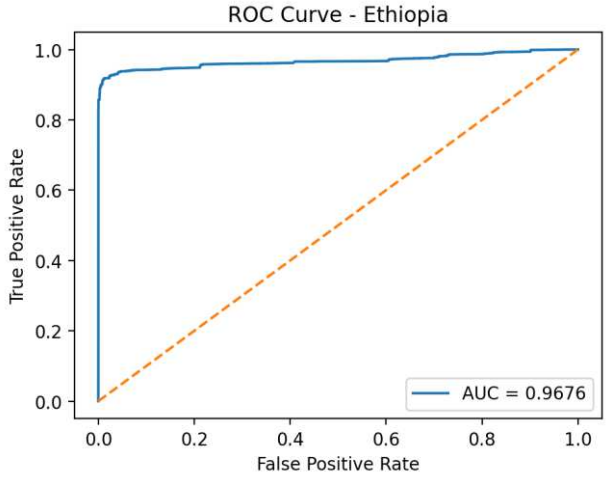



ROC Curve - Dominican
True Positive Rate
False Positive Rate
AUC = 0.9861

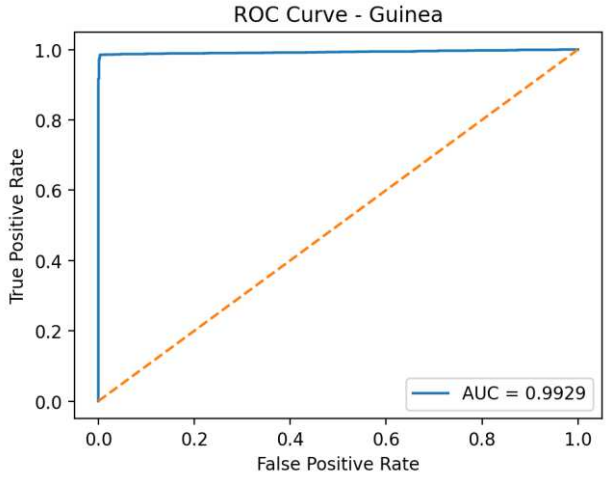
(c)


ROC Curve - Dominican Republic
True Positive Rate
False Positive Rate
AUC = 0.9828

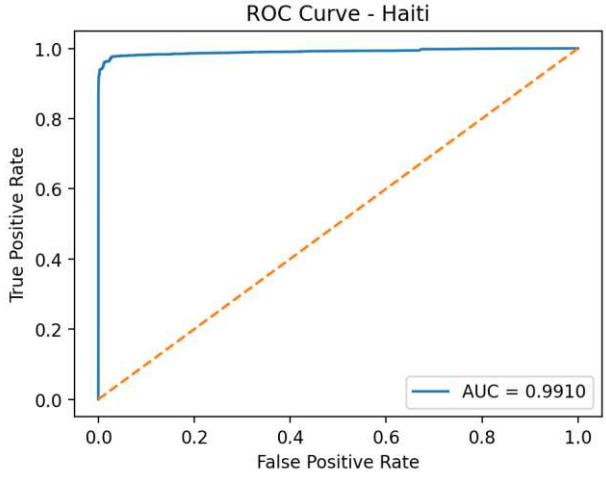
(d)


ROC Curve - Ethiopia
True Positive Rate
False Positive Rate
AUC = 0.9676

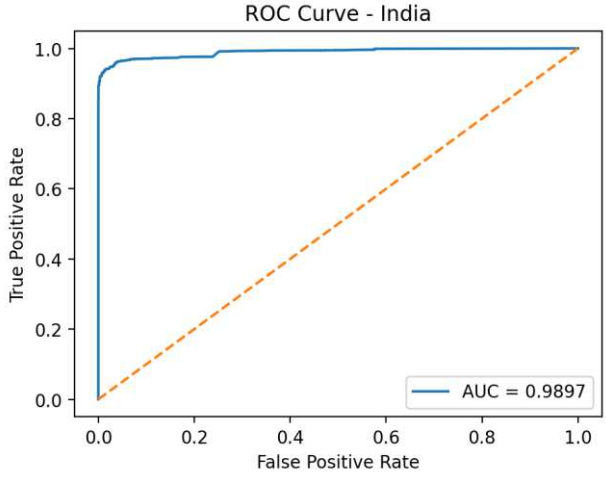
(e)


ROC Curve - Guinea
True Positive Rate
False Positive Rate
AUC = 0.9929

(f)


ROC Curve - Haiti
True Positive Rate
False Positive Rate
AUC = 0.9910

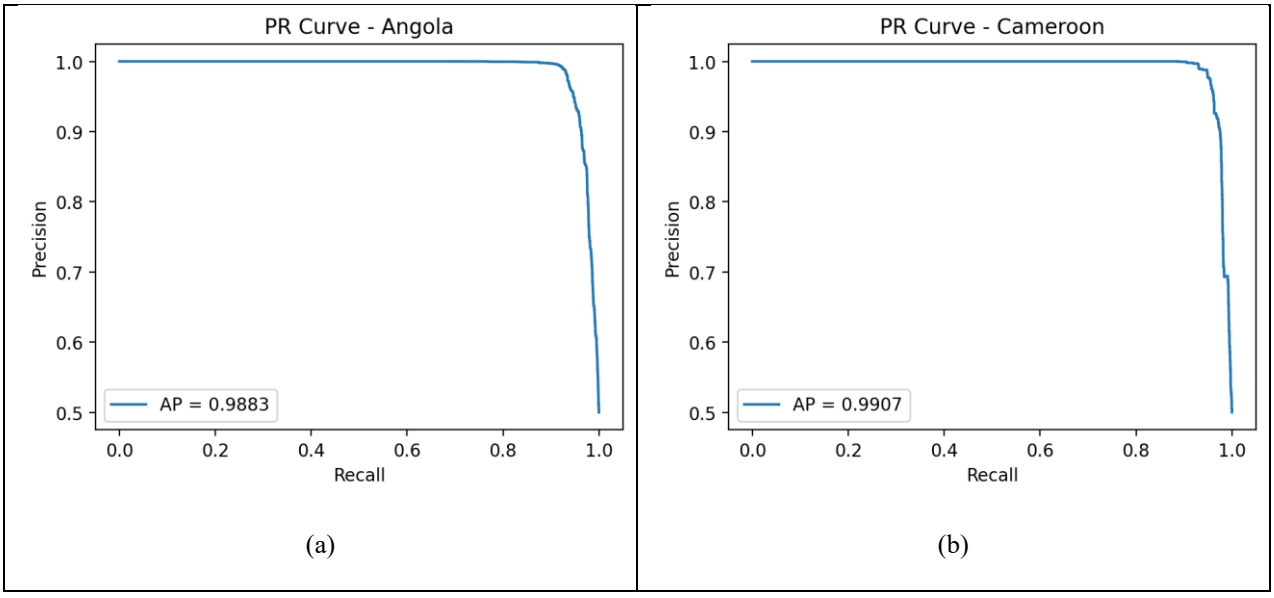
(g)

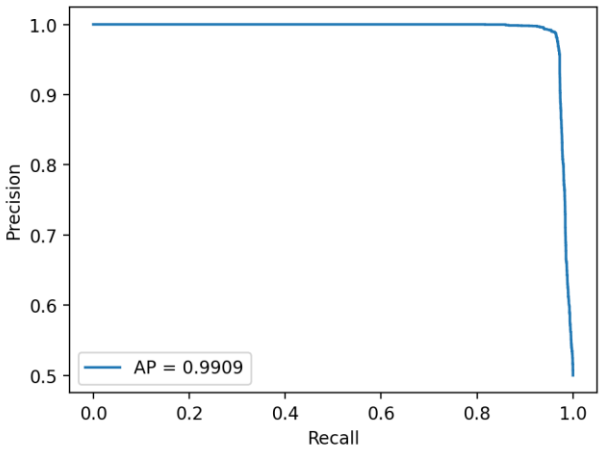

ROC Curve - India
True Positive Rate
False Positive Rate
AUC = 0.9897

(h)

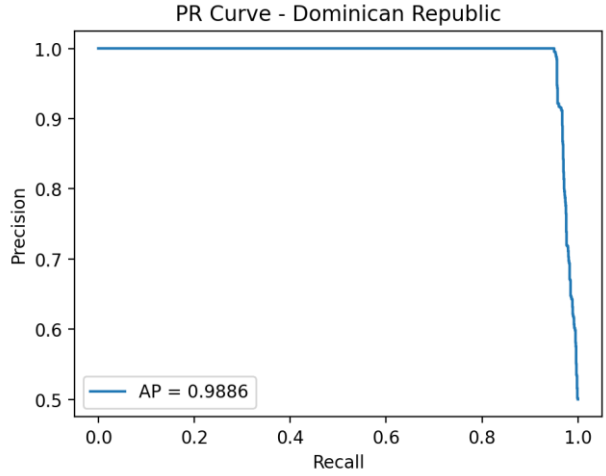
Figure 4 Obtained ROC curves for each country (a) Angola, (b) Cameroon, (c) Dominican, (d) Dominican Republic, (e) Ethiopia, (d) Guinea, (g) Haiti, and (h) India

As demonstrated in Figure 5, it can be noted that PR metrics also demonstrated good stability across countries; however, there were certain small discrepancies between curves. India achieved the best average precision with the result of 0.9919, while Cameroon reached the second place with 0.9907, while Dominican country and Dominican Republic got 0.9909 and 0.9886, respectively. Angola demonstrated impressive precision metrics with the value of 0.9883. Guinea maintained its top-level position at 0.9956; however, the lowest average precisions among the countries under consideration were observed for Ethiopia and Haiti, which equaled 0.9794. Taking into account the side-by-side comparison of all of the countries under consideration, one should note that the trend is almost the same as it was described for ROC analysis: performance levels were high everywhere, yet some countries appeared easier to process than others. Guinea and India can be regarded as the best in this test with precision staying as close as possible to the top-right corner of the plot. Cameroon, Dominican, and Dominican Republic also demonstrated similar tendencies, yet their results were very stable. It was noticed that Ethiopia and Haiti demonstrated more significant drops on the plots towards their end positions.

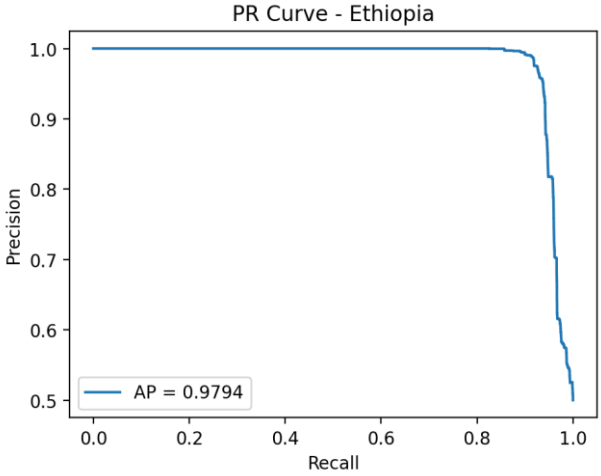



PR Curve - Dominican
Precision
Recall
AP = 0.9909

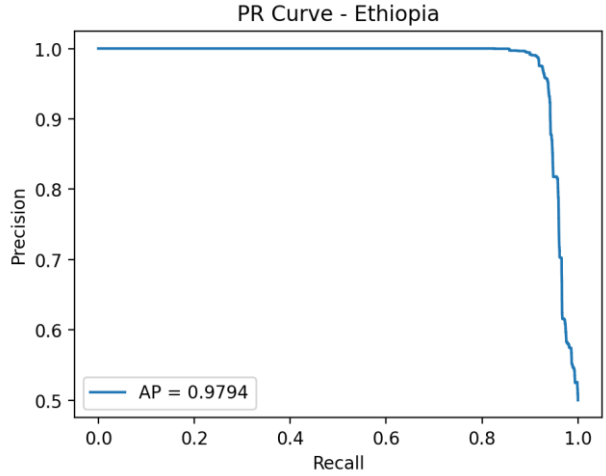
(c)


PR Curve - Dominican Republic
Precision
Recall
AP = 0.9886

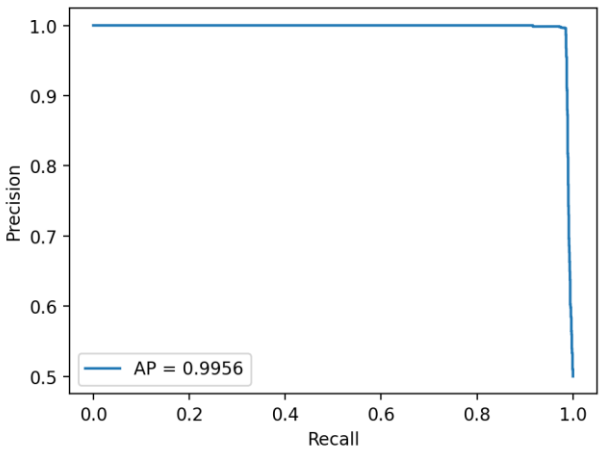
(d)


PR Curve - Ethiopia
Precision
Recall
AP = 0.9794

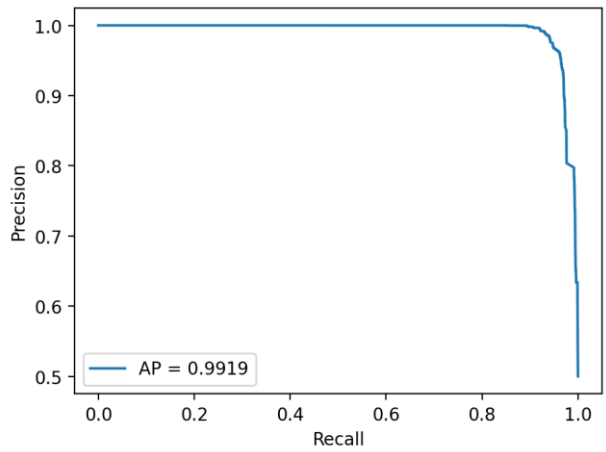
(e)


PR Curve - Ethiopia
Precision
Recall
AP = 0.9794

(f)


PR Curve - Guinea
Precision
Recall
AP = 0.9956

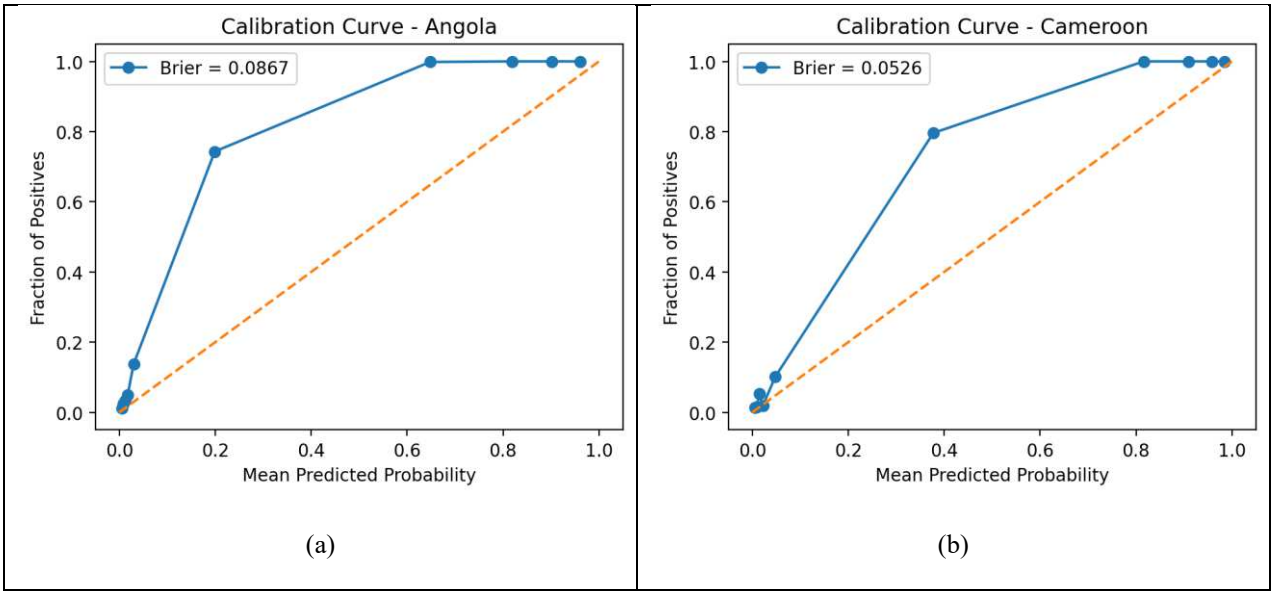
(g)

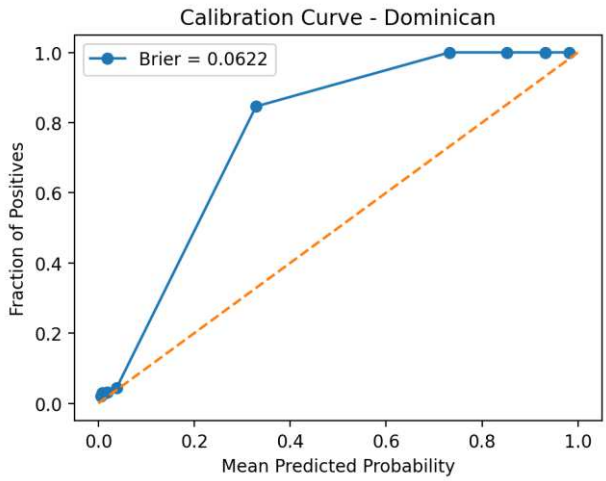

PR Curve - India
Precision
Recall
AP = 0.9919

(h)

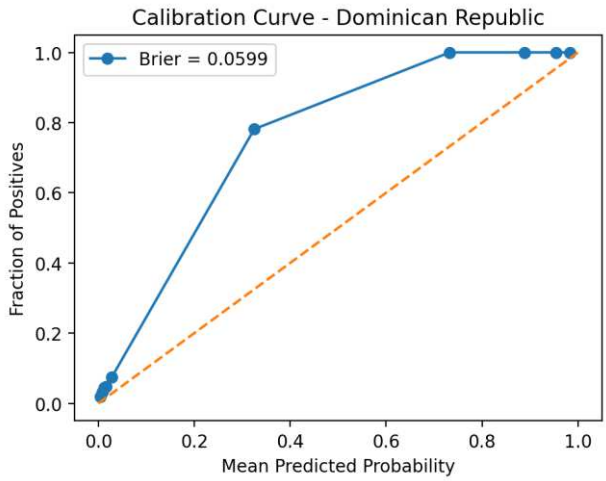
Figure 5 Obtained PR curves for each country (a) Angola, (b) Cameroon, (c) Dominican, (d) Dominican Republic, (e) Ethiopia, (d) Guinea, (g) Haiti, and (h) India

Figure 6 suggests that the model probabilities were generally well aligned with the observed outcomes, although the level of calibration was not identical across countries. The country that was able to show the best results in terms of calibration was Guinea, with the Brier score of 0.0328, then Cameroon with 0.0526, Dominican Republic with 0.0599, Dominican with 0.0622, and Angola with 0.0867. On the other hand, the highest Brier score was found for India, which amounted to 0.1219, whereas the maximum values could be observed for Ethiopia and Haiti, with the value of 0.1309. If to compare the countries among each other, it becomes obvious that certain differences exist. Thus, Guinea, Cameroon, Dominican Republic, and Dominican can be considered as the ones with the most accurate probability estimates due to their curves being close to the diagonal reference line and low Brier scores. However, Angola can still be regarded as the one with good calibration performance despite the curve deviating significantly from the diagonal reference line.

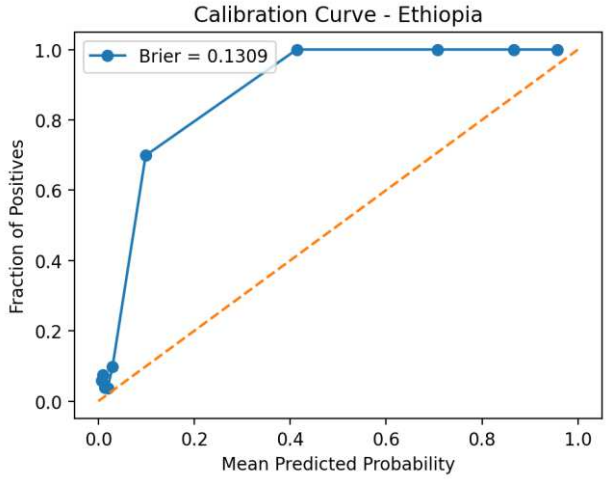



Calibration Curve - Dominican
Brier = 0.0622
Fraction of Positives
Mean Predicted Probability

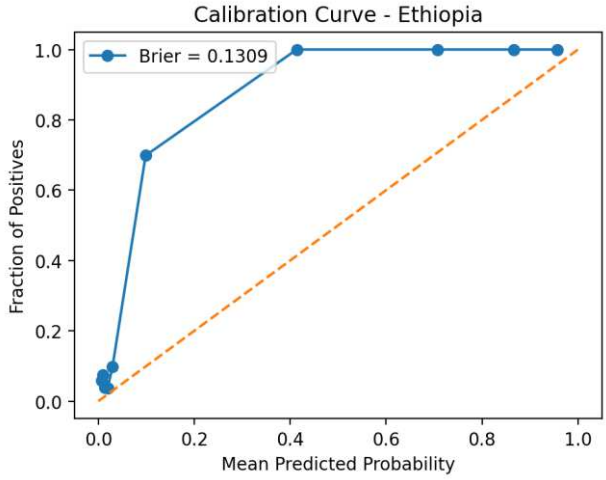
(c)


Calibration Curve - Dominican Republic
Brier = 0.0599
Fraction of Positives
Mean Predicted Probability

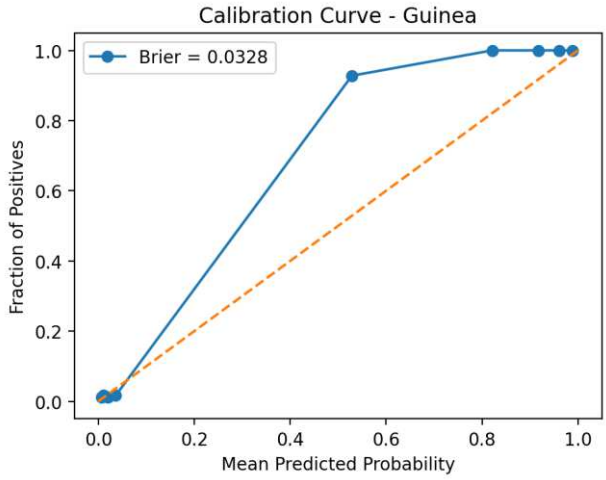
(d)


Calibration Curve - Ethiopia
Brier = 0.1309
Fraction of Positives
Mean Predicted Probability

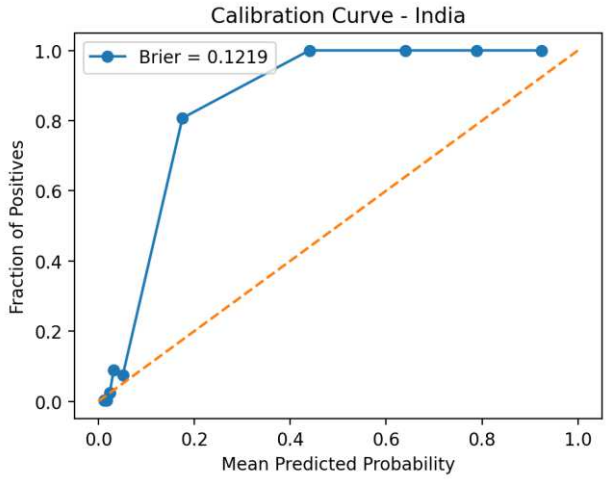
(e)


Calibration Curve - Ethiopia
Brier = 0.1309
Fraction of Positives
Mean Predicted Probability

(f)


Calibration Curve - Guinea
Brier = 0.0328
Fraction of Positives
Mean Predicted Probability

(g)


Calibration Curve - India
Brier = 0.1219
Fraction of Positives
Mean Predicted Probability

(h)

Figure 6 Obtained calibration curves for each country (a) Angola, (b) Cameroon, (c) Dominican, (d) Dominican Republic, (e) Ethiopia, (d) Guinea, (g) Haiti, and (h) India

The results of the performance evaluation metrics that have been calculated using the LOCO strategy are presented in Table 1 below. From Table 1, it is evident that the proposed methodology exhibited outstanding performance in each of the country-specific partition tests. The framework generated relatively high results in terms of AUC, accuracy, F1-score, and average precision values in each of the country-specific partitions. From Table 1, Guinea attained the highest scores of 0.9929 in terms of AUC, 0.9861 in terms of accuracy, 0.9861 in terms of F1-score, and 0.9956 in terms of average precision. Additionally, Guinea scored the lowest Brier score value of 0.0328 compared to other countries. These results indicate that the developed model had excellent discriminative performance in Guinea. The framework attained similar performance in Haiti, Cameroon, and Dominican, where each country scored above 0.9860 in terms of AUC value and around 0.96 in terms of accuracy. India also had high discriminative performance with an AUC value of 0.9897 and an average precision value of 0.9919; however, the framework's accuracy and calibration in India were less accurate than those in the top-performing countries. Lastly, the lowest performance was recorded in Ethiopia, where the framework exhibited an AUC value of 0.9676, accuracy value of 0.9034, and F1-score value of 0.8939. The proposed framework achieved a mean AUC of 0.9850 ± 0.0074, a mean accuracy of 0.9537 ± 0.0239, a mean F1 score of 0.9511 ± 0.0268, an average precision of 0.9899 ± 0.0046, and a Brier score of 0.0762 ± 0.0322, indicating consistently strong performance under the LOCO evaluation setting. Even so, these values still indicate a relatively strong level of performance. When the countries are compared more closely, the table suggests that discrimination remained strong across all partitions, whereas calibration varied more noticeably. For example, Guinea, Cameroon, and Dominican Republic had comparatively low Brier scores, indicating more reliable probability estimates, while India and Ethiopia had higher Brier scores, suggesting that probability calibration was more challenging in those settings.

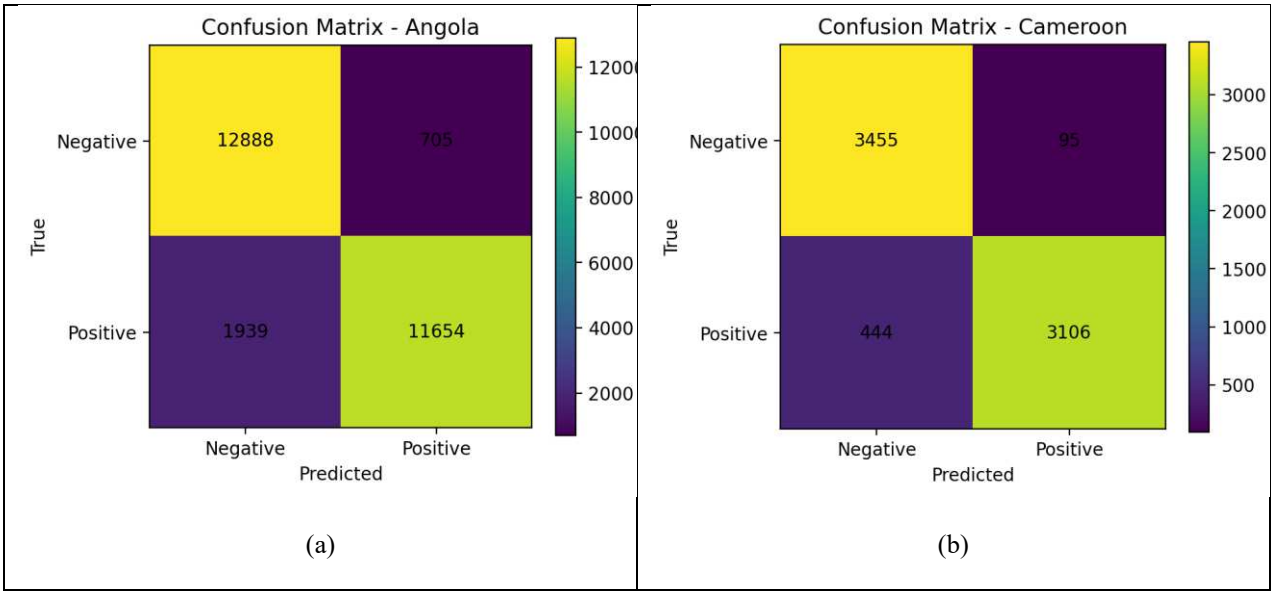
Table 1 Country-level performance summary of the proposed framework for HIV prediction.

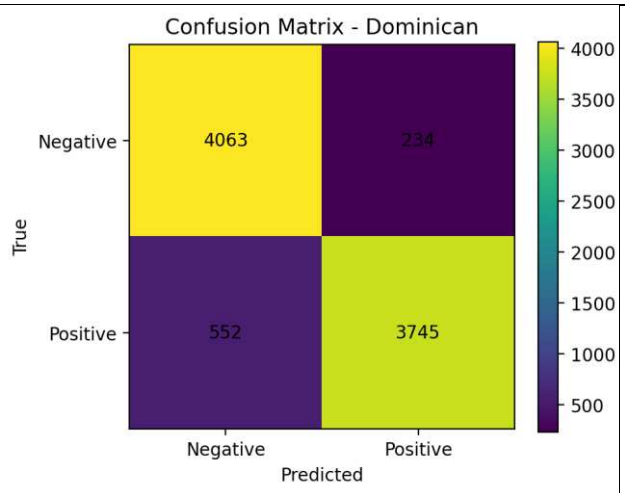
Country	**AUC**	**Accuracy**	**F1 Score**	**Average Precision**	**Brier Score**

Haiti	0.9910	0.9614	0.9605	0.9938	0.0622
Dominican Republic	0.9828	0.9736	0.9729	0.9886	0.0599
India	0.9897	0.9398	0.9359	0.9919	0.1219
Angola	0.9836	0.9402	0.9366	0.9883	0.0867
Guinea	0.9929	0.9861	0.9861	0.9956	0.0328
Ethiopia	0.9676	0.9034	0.8939	0.9794	0.1309
Dominican	0.9861	0.9644	0.9638	0.9909	0.0622
Cameroon	0.9866	0.9605	0.9594	0.9907	0.0526
Mean ± Std	0.9850 ± 0.0074	0.9537 ± 0.0239	0.9511 ± 0.0268	0.9899 ± 0.0046	0.0762 ± 0.0322

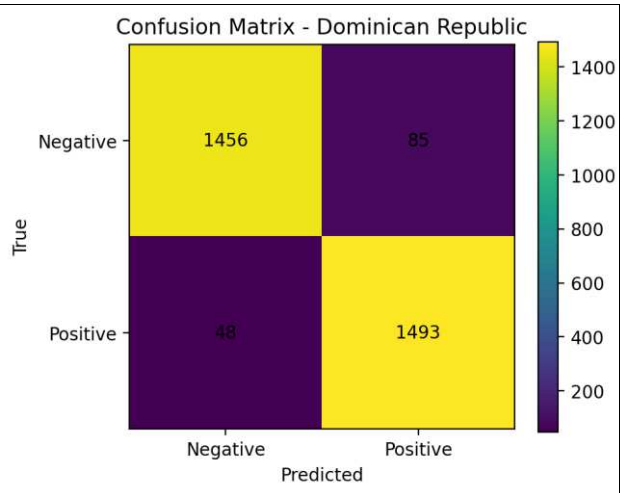
3.3. Results on STI prediction

The second experiment was conducted on STI prediction and the obtained visual and numerical results were given in Figures 7, 8, 9 and 10, respectively. Table 2 also shows the calculated performance evaluation metrics for each country.

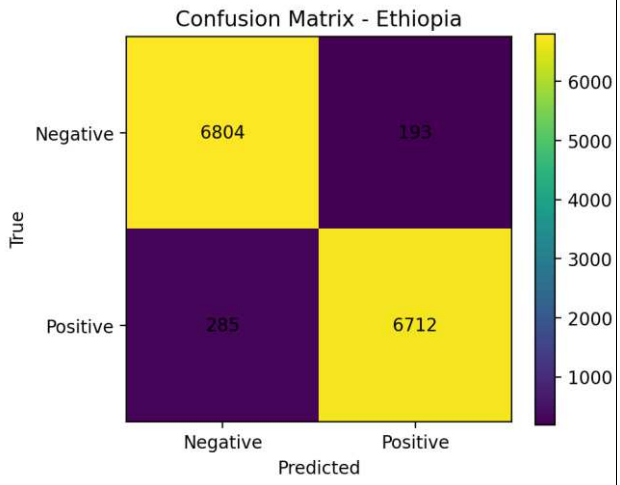



Confusion Matrix - Dominican
Negative
Positive
True
4063
234
552
3745
Negative
Positive
Predicted
4000
3500
3000
2500
2000
1500
1000
500

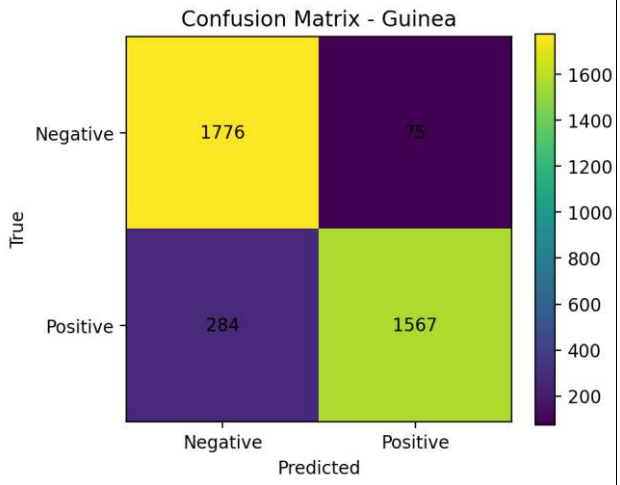
(c)


Confusion Matrix - Dominican Republic
Negative
Positive
True
1456
85
48
1493
Negative
Positive
Predicted
1400
1200
1000
800
600
400
200

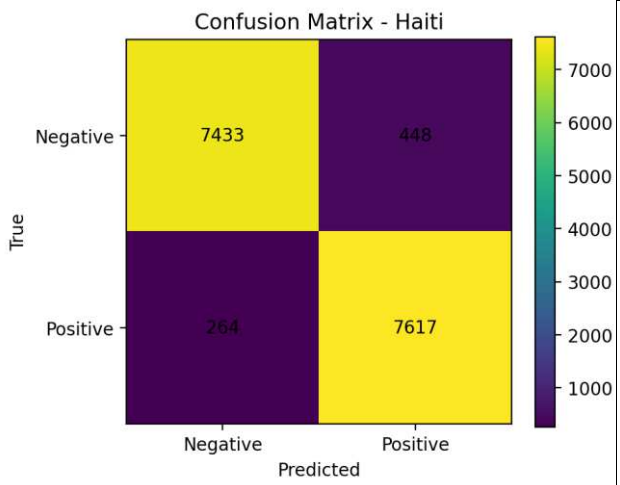
(d)


Confusion Matrix - Ethiopia
Negative
Positive
True
6804
193
285
6712
Negative
Positive
Predicted
6000
5000
4000
3000
2000
1000

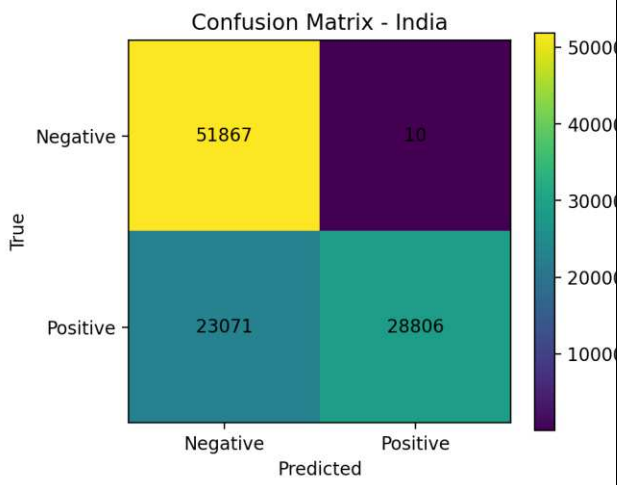
(e)


Confusion Matrix - Guinea
Negative
Positive
True
1776
75
284
1567
Negative
Positive
Predicted
1600
1400
1200
1000
800
600
400
200

(f)


Confusion Matrix - Haiti
Negative
Positive
True
7433
448
264
7617
Negative
Positive
Predicted
7000
6000
5000
4000
3000
2000
1000

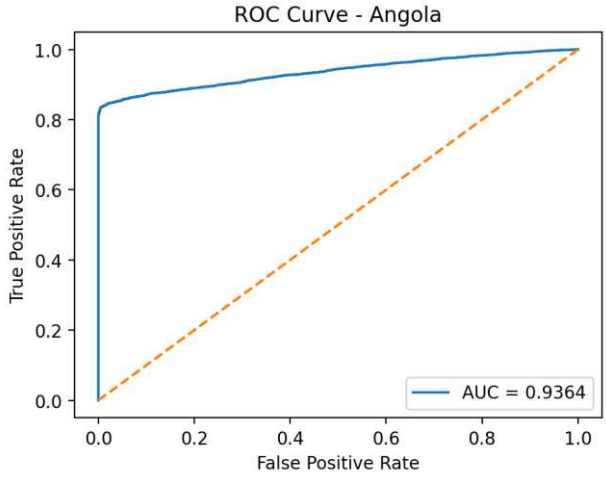
(g)


Confusion Matrix - India
Negative
Positive
True
51867
10
23071
28806
Negative
Positive
Predicted

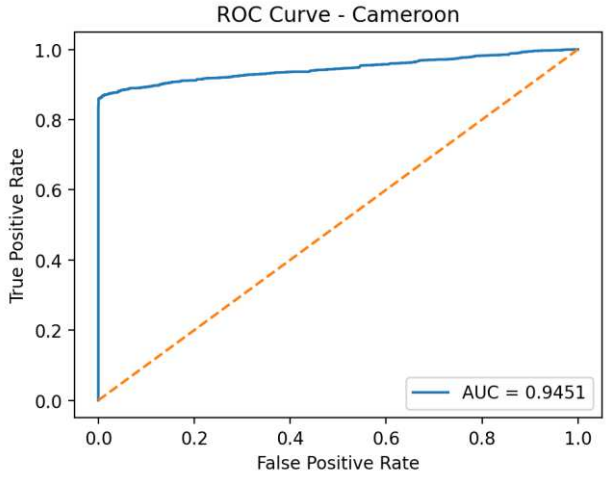
(h)

Figure 7 Obtained confusion matrixes on STI classification for each country (a) Angola, (b) Cameroon, (c) Dominican, (d) Dominican Republic, (e) Ethiopia, (d) Guinea, (g) Haiti, and (h) India

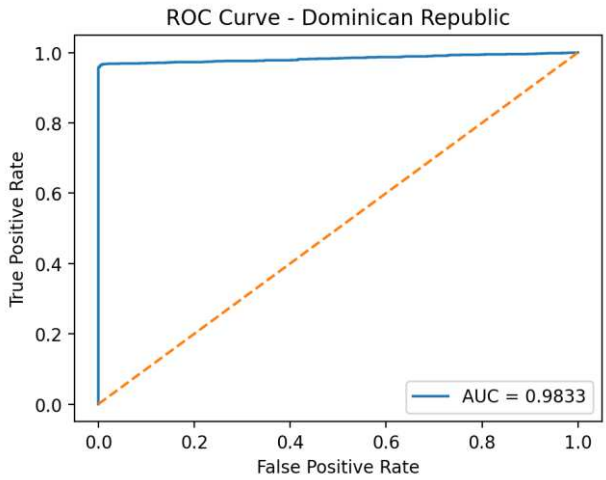
Figure 7 presents the confusion matrices for STI classification across the different countries. Overall, the results show that the model made considerably more correct predictions than incorrect ones in most country-specific partitions. In Angola, for example, the model correctly classified 12,888 negative samples and 11,654 positive samples, while misclassifying 705 negative samples as positive and 1,939 positive samples as negative. A similar trend appeared in Cameroon, where 3,455 negative and 3,106 positive cases were identified correctly, compared with 95 false positives and 444 false negatives. In the Dominican dataset, there were 4,063 true negative predictions and 3,745 true positive predictions, while there were 234 false negative predictions and 552 false positive predictions. The Dominican Republic generated one of the best results in the confusion matrix, with 1,456 true negative predictions and 1,493 true positive predictions, while there were only 85 false negative predictions and 48 false positive predictions. In Ethiopia, the confusion matrix results were also quite remarkable. There were 6,804 true negative predictions and 6,712 true positive predictions, while only 193 were false negative predictions and 285 were false positive predictions. In the Guinea dataset, there were 1,776 true negative predictions and 1,567 true positive predictions, while 75 were false negative predictions and 284 were false positive predictions. Similar to Ethiopia and Guinea, Haiti had excellent results in the confusion matrix. There were 7,433 true negative predictions and 7,617 true positive predictions, while there were only 448 false negative predictions and 264 false positive predictions. Finally, the Indian dataset presented a unique situation. Although there were 51,867 true negative predictions and 28,806 true positive predictions, there were 23,071 false negative predictions, which is a significantly higher figure than what was recorded for all other datasets. Therefore, it can be argued that in the Indian partition, the algorithm performed exceptionally well in detecting negative STIs but had significant difficulty detecting positive STI.


ROC Curve - Angola
True Positive Rate
False Positive Rate
AUC = 0.9364

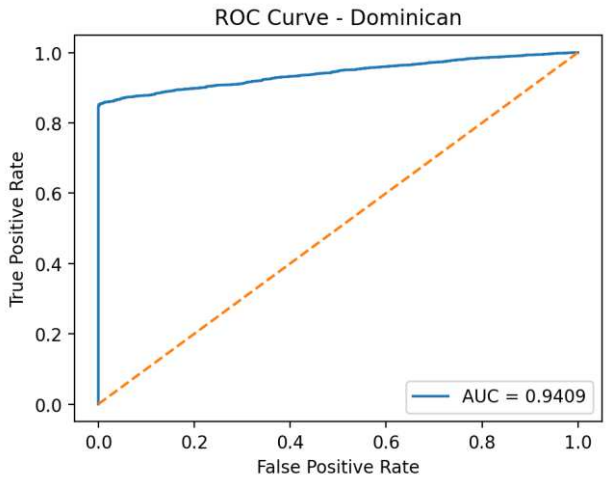
(a)


ROC Curve - Cameroon
True Positive Rate
False Positive Rate
AUC = 0.9451

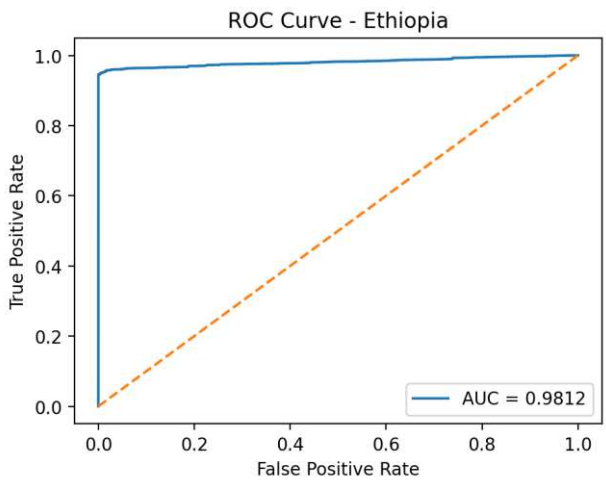
(b)


ROC Curve - Dominican Republic
True Positive Rate
False Positive Rate
AUC = 0.9833

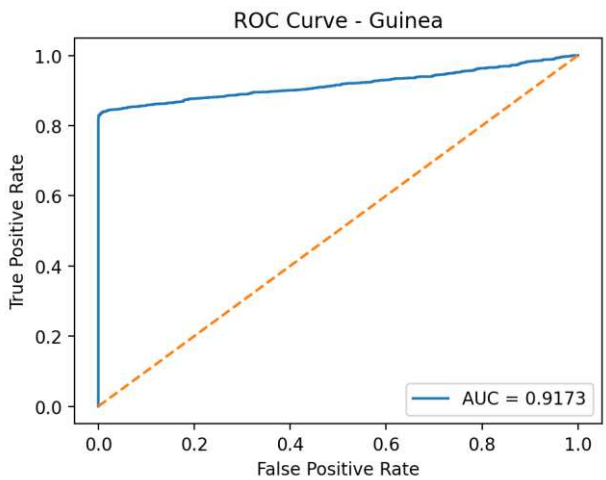
(c)


ROC Curve - Dominican
True Positive Rate
False Positive Rate
AUC = 0.9409

(d)


ROC Curve - Ethiopia
True Positive Rate
False Positive Rate
AUC = 0.9812

(e)


ROC Curve - Guinea
True Positive Rate
False Positive Rate
AUC = 0.9173

(f)

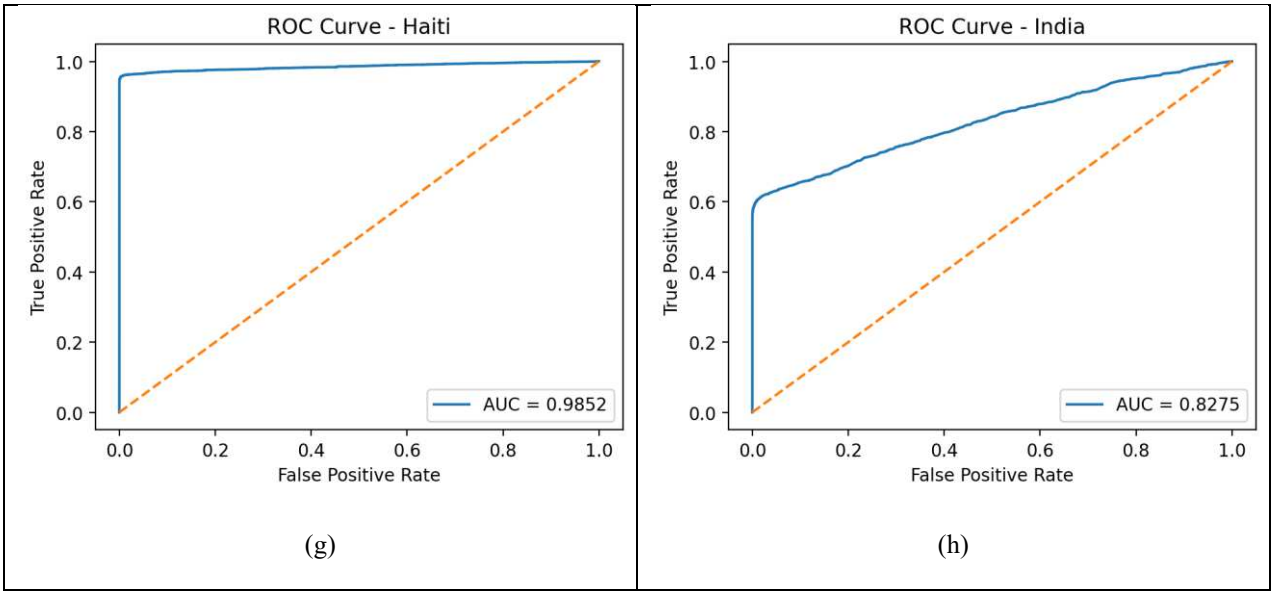
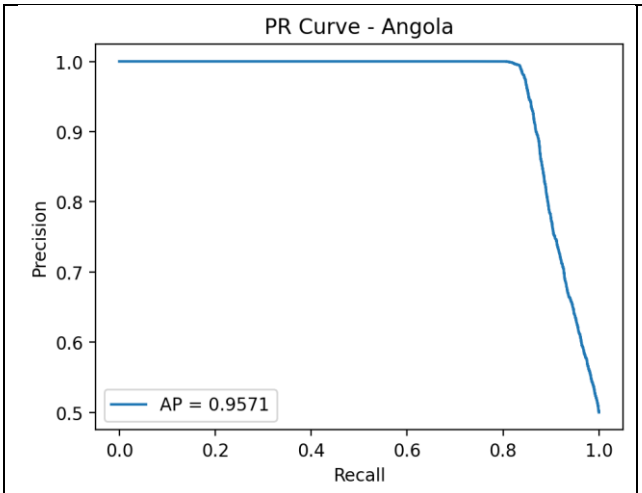
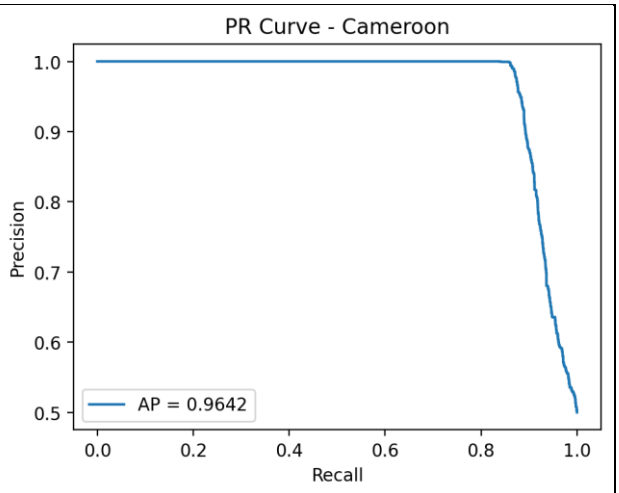


Figure 8 Obtained ROC curves for each country (a) Angola, (b) Cameroon, (c) Dominican, (d) Dominican Republic, (e) Ethiopia, (d) Guinea, (g) Haiti, and (h) India

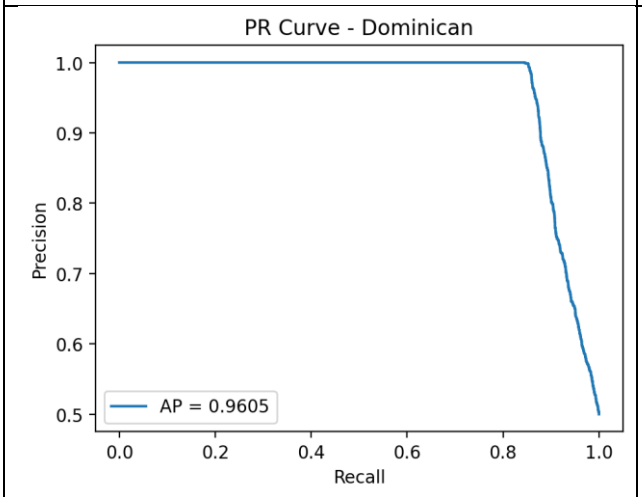
Figure 8 shows ROC curves for STI classification on the given eight country datasets. As seen in Figure 8, the proposed model produced excellently in most of the eight country datasets, showing high values of the AUC metric in almost all partitions. As such, this means that the model managed to correctly separate positive and negative STI samples, but, at the same time, the results varied depending on the country of origin of the data. Particularly impressive results were demonstrated by the country of Angola with AUC = 0.9364 and Cameroon with AUC = 0.9451, when the ROC curve did not drop below the dashed diagonal reference line and, hence, possessed strong discriminability. Even more pronounced results were obtained in the Dominican Republic with AUC = 0.9833, Ethiopia with AUC = 0.9812, and Haiti with AUC = 0.9852, as the ROC curves in these cases were close to the upper-left corner, which shows that the classification process was effective and the model was able to make accurate predictions of whether a patient suffers from STI or not. The data set for the Dominican Republic also demonstrated a satisfactory level of the model's performance with AUC = 0.9409, and Guinea with AUC = 0.9173 also showed promising results, even though the class separation was not quite clear. Finally, the Indian data set is worth mentioning because of the least favorable AUC score = 0.8275 and rather gentle ROC curves.


PR Curve - Angola
Precision
Recall
AP = 0.9571

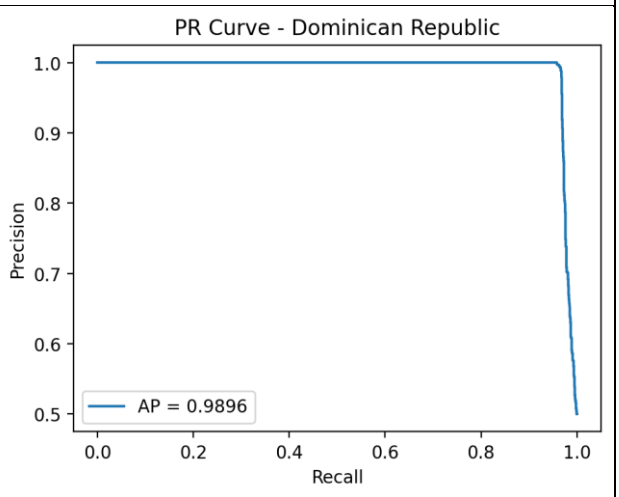
(a)


PR Curve - Cameroon
Precision
Recall
AP = 0.9642

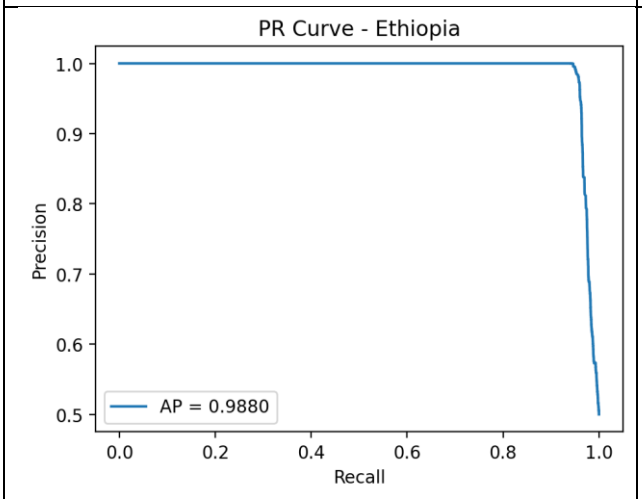
(b)


PR Curve - Dominican
Precision
Recall
AP = 0.9605

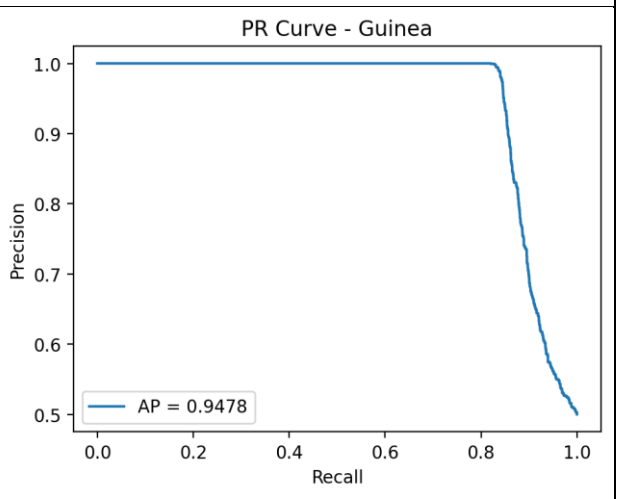
(c)


PR Curve - Dominican Republic
Precision
Recall
AP = 0.9896

(d)


PR Curve - Ethiopia
Precision
Recall
AP = 0.9880

(e)


PR Curve - Guinea
Precision
Recall
AP = 0.9478

(f)

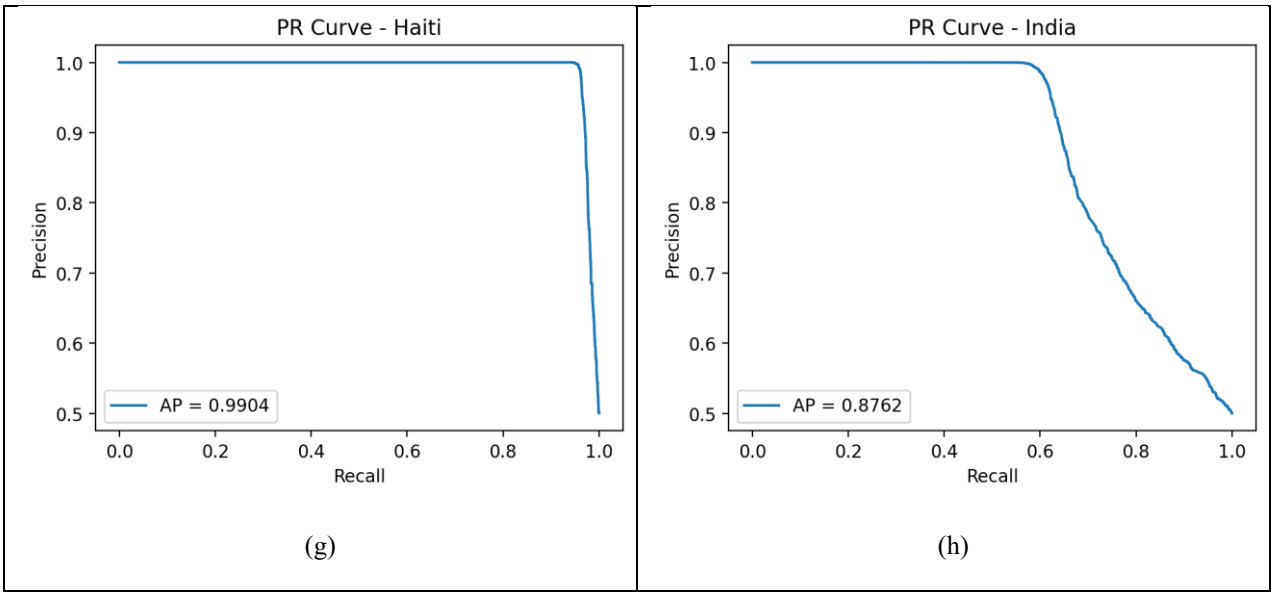
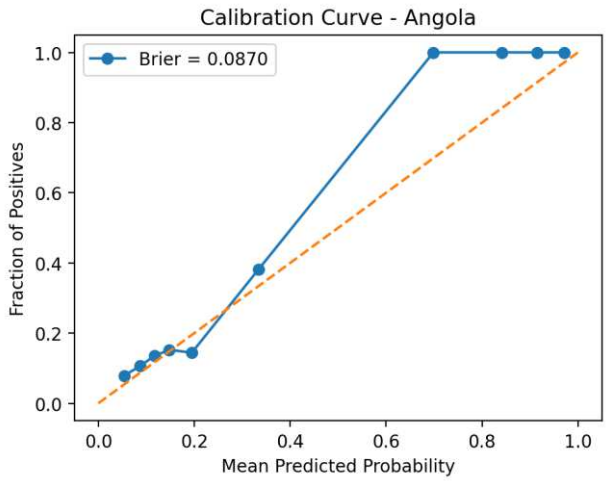
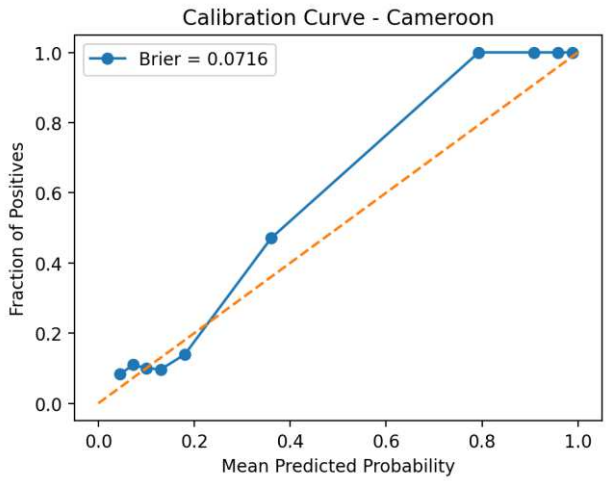


Figure 9 Obtained PR curves for each country (a) Angola, (b) Cameroon, (c) Dominican, (d) Dominican Republic, (e) Ethiopia, (d) Guinea, (g) Haiti, and (h) India

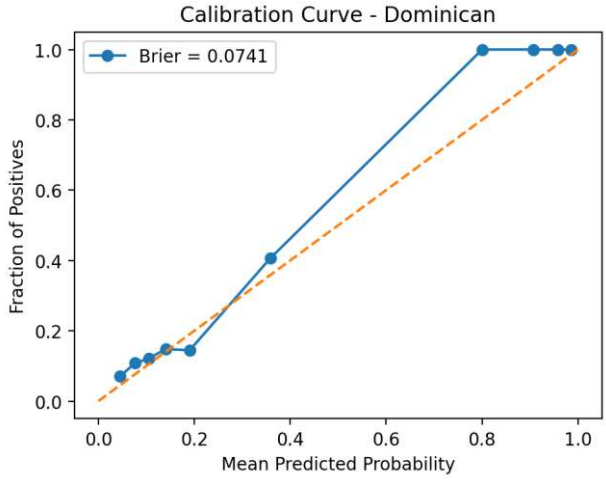
Figure 9 shows PR curves for STI detection from each of the eight country-based datasets. It can be observed from the curve that the precision of the algorithm remained fairly consistent despite changes in recall rate in most of the countries, implying that the positives have been identified effectively while maintaining good predictability as well. It is particularly evident that the algorithm performed exceptionally well for Haiti (Average Precision (AP) = 0.9904), the Dominican Republic (AP = 0.9896), and Ethiopia (AP = 0.9880) since their PR curves stayed close to the top right corner of the graph throughout. Moreover, Cameroon (AP = 0.9642), the Dominican dataset (AP = 0.9605), and Angola (AP = 0.9571) are also notable as all three of them had performed very well. Guinea produced a slightly lower but still solid result with an AP of 0.9478, indicating good precision-recall performance despite a somewhat earlier decline compared with the strongest country-specific curves. India, however, again stood out as the most challenging dataset, with the lowest AP value of 0.8762 and a more noticeable drop in precision as recall increased, suggesting that the model had greater difficulty maintaining the same level of confidence and consistency when identifying positive STI cases in this partition.


Calibration Curve - Angola
Brier = 0.0870
Fraction of Positives
Mean Predicted Probability
0.0
0.2
0.4
0.6
0.8
1.0

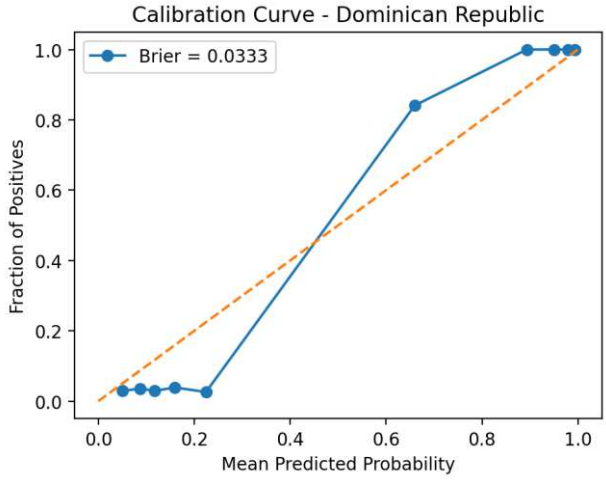
(a)


Calibration Curve - Cameroon
Brier = 0.0716
Fraction of Positives
Mean Predicted Probability
0.0
0.2
0.4
0.6
0.8
1.0

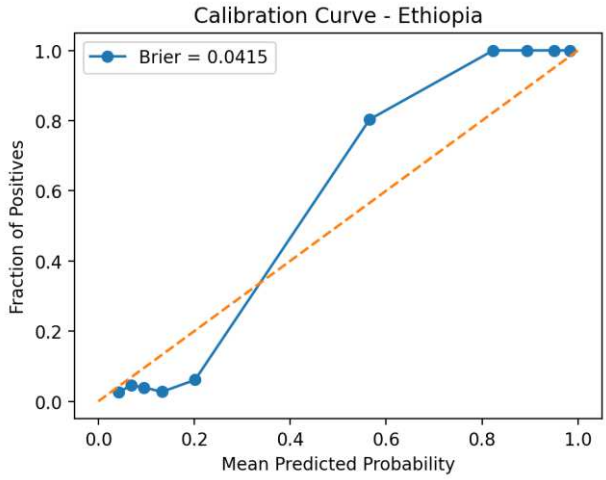
(b)


Calibration Curve - Dominican
Brier = 0.0741
Fraction of Positives
Mean Predicted Probability
0.0
0.2
0.4
0.6
0.8
1.0

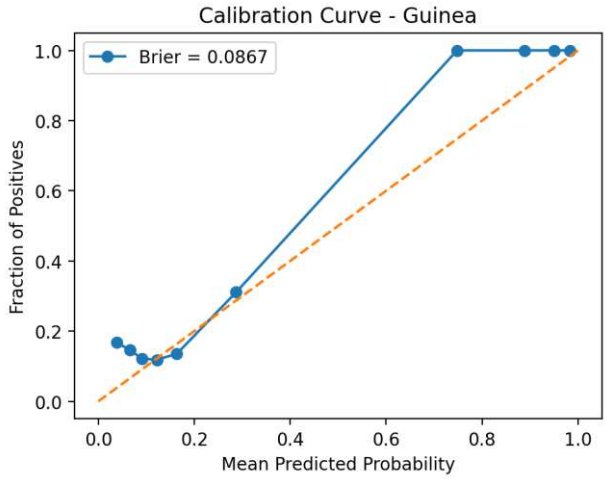
(c)


Calibration Curve - Dominican Republic
Brier = 0.0333
Fraction of Positives
Mean Predicted Probability
0.0
0.2
0.4
0.6
0.8
1.0

(d)


Calibration Curve - Ethiopia
Brier = 0.0415
Fraction of Positives
Mean Predicted Probability
0.0
0.2
0.4
0.6
0.8
1.0

(e)


Calibration Curve - Guinea
Brier = 0.0867
Fraction of Positives
Mean Predicted Probability
0.0
0.2
0.4
0.6
0.8
1.0

(f)

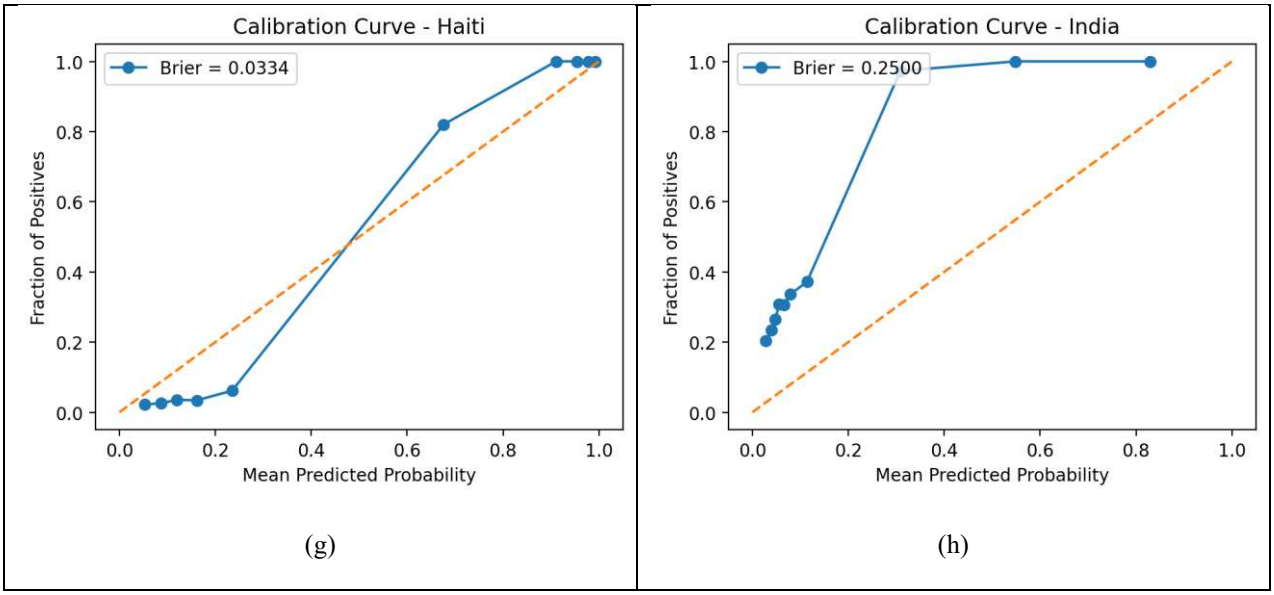


Figure 10. Obtained calibration curves for each country (a) Angola, (b) Cameroon, (c) Dominican, (d) Dominican Republic, (e) Ethiopia, (d) Guinea, (g) Haiti, and (h) India

The calibration curves for STI classification for the given sets of data are presented in Figure 10, which shows that calibration was more heterogeneous across countries than discrimination capacity was. The calibration curves of the Dominican Republic, Haiti, and Ethiopia are closest to the ideal straight line with Brier scores of 0.0333, 0.0334, and 0.0415, respectively. Thus, in these datasets, the probabilities predicted by the model were relatively close to the probabilities obtained based on empirical observations. Calibration is quite satisfactory in Cameroon and Dominican, with Brier scores of 0.0716 and 0.0741, and somewhat poor in Angola and Guinea, with Brier scores of 0.0870 and 0.0867. In those countries, the prediction line was fairly consistent with the reference one, but still showed some discrepancies, which indicates that predicted risks of classes did not match the actual ones in the studied probability ranges. India proved to be the least successful country in terms of model calibration, with the lowest Brier score of 0.2500 and the calibration curve being very distant from the straight line. It means that the probabilities predicted for the Indian dataset did not correspond to the probabilities of occurrence of classes very well compared to other countries' datasets.

Table 2 Country-level performance summary of the proposed framework for STI prediction.

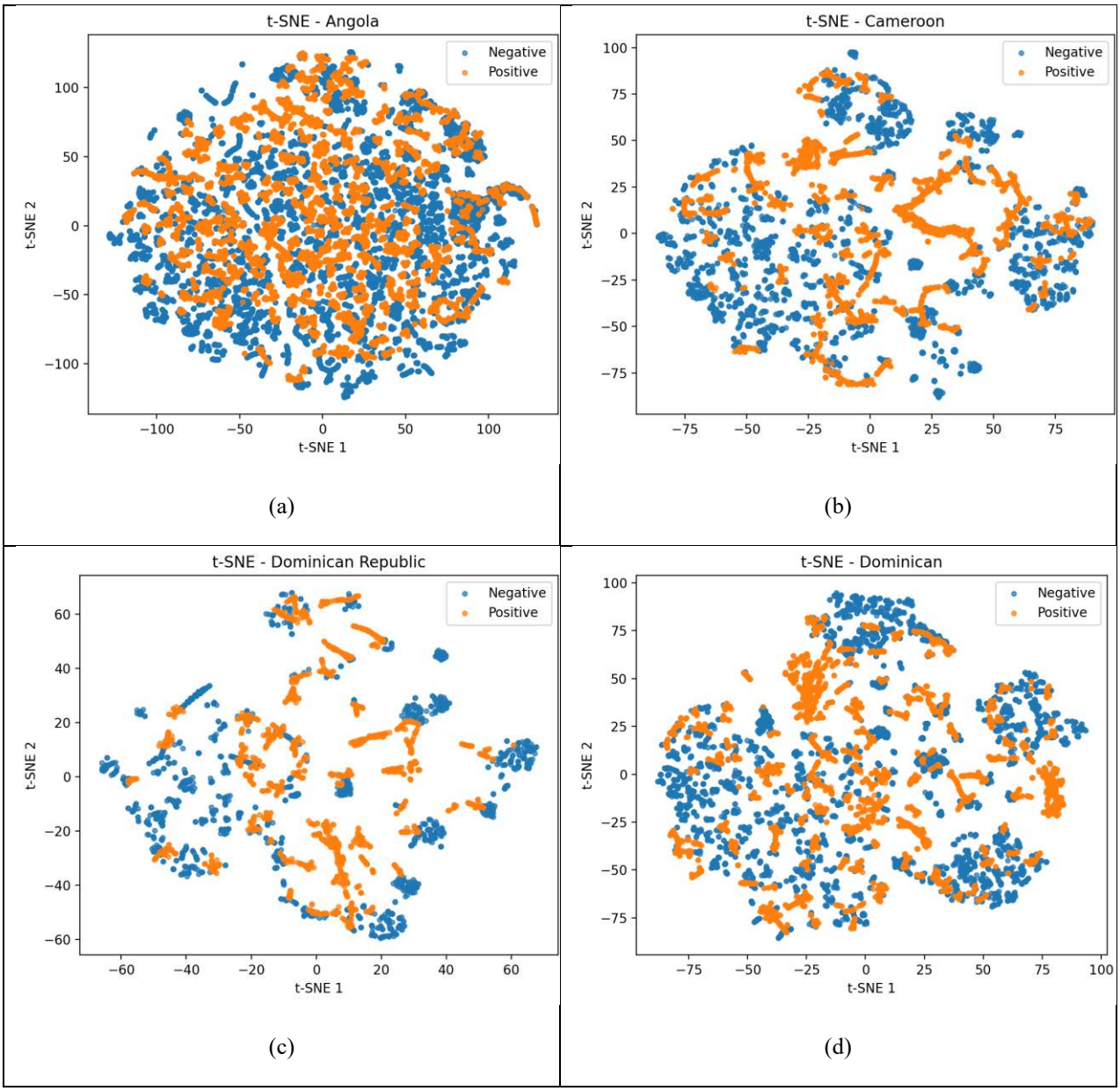
Country	AUC	Accuracy	F1 Score	Average Precision	Brier Score
Haiti	0.9852	0.9548	0.9553	0.9904	0.0334
Dominican Republic	0.9833	0.9568	0.9573	0.9896	0.0333
India	0.8275	0.7775	0.7140	0.8762	0.2500
Angola	0.9364	0.9027	0.8981	0.9571	0.0870
Guinea	0.9173	0.9030	0.8971	0.9478	0.0867
Ethiopia	0.9812	0.9658	0.9656	0.9880	0.0415
Dominican	0.9409	0.9085	0.9050	0.9605	0.0741
Cameroon	0.9451	0.9241	0.9209	0.9642	0.0716
Mean ± Std	0.9396 ± 0.0485	0.9117 ± 0.0561	0.9016 ± 0.0756	0.9592 ± 0.0350	0.0847 ± 0.0659

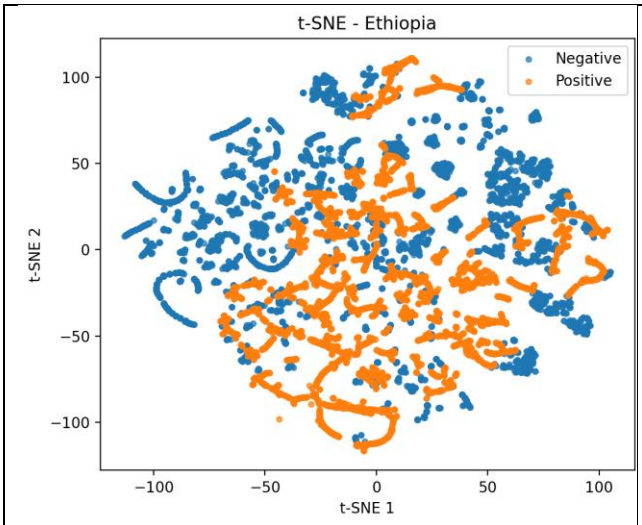
Table 2 illustrates the summary of the proposed STI prediction framework on a per-country basis and shows remarkable performance for all countries with different levels of performance. For example, Haiti has achieved excellent results with an AUC of 0.9852, accuracy of 0.9548, F1 score of 0.9553, average precision of 0.9904, and Brier score of 0.0334. These figures indicate superior discriminative ability and excellent calibrated probability estimations. Similarly, the Dominican Republic has shown outstanding results with an AUC of 0.9833, accuracy of 0.9568, F1 score of 0.9573, average precision of 0.9896, and Brier score of 0.0333. Hence, this country ranks amongst the top-performing countries due to highly reliable probability estimates. Similarly, Ethiopia has achieved excellent results with an AUC of 0.9812, accuracy of 0.9658, F1 score of 0.9656, average precision of 0.9880, and Brier score of 0.0415. Thus, this country ranks as one of the best-performing countries due to high-quality results. Cameroon, Angola, Dominican, and Guinea have shown promising results with an AUC of 0.9173 to 0.9451, accuracy of 0.9030 to 0.9241, F1 score of 0.8971 to 0.9209, and average precision of 0.9478 to 0.9642. However, the calibration performance is slightly inferior to the top-performing countries. In comparison, India was found to be the toughest test case because of poor results in all metrics: AUC of 0.8275, accuracy of 0.7775, F1 of 0.7140, and average precision of 0.8762, besides having a considerably higher Brier score of 0.2500. On the other hand, the overall scores

affirm the effectiveness of the system because of its means of 0.9396 ± 0.0485 for AUC, 0.9117 ± 0.0561 for accuracy, 0.9016 ± 0.0756 for F1, 0.9592 ± 0.0350 for average precision, and 0.0847 ± 0.0659 for Brier.

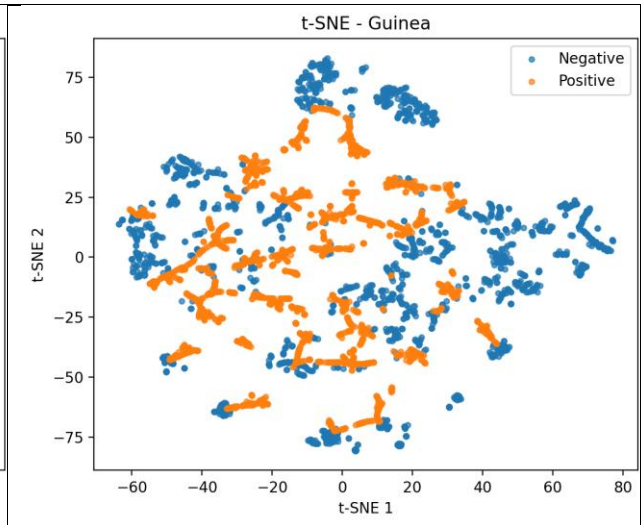
3.4. Statistical Analysis

The statistical analysis using t-SNE visualization technique was used in order to analyze the latent structure of the data set and investigate how positive and negative samples were distributed in the space formed by selected features [29]. The main purpose of this method was to provide a visual description of class separability and data heterogeneity and supplement traditional quantitative analysis with information on data structure. More specifically, the use of t-SNE visualization helped to check whether there was any particular tendency to cluster country-partitions, large overlaps between classes or differences in structural heterogeneity that might impact predictive performance. Based on visualized results, it can be concluded that class separability is quite different among various country-partitions and may lead to differences in the accuracy of predictive models. Figure 11 presents the t-SNE plots generated for HIV-positive samples' prediction in the country-specific and aggregated data sets. From those plots, it is clear that there are noticeable differences in sample separation among countries. For Angola, however, extensive intermingling between the classes was evident, pointing to poor visual distinction between the two classes in the small number of dimensions. With respect to Cameroon and Dominican Republic, clear signs of local clustering were observed, along with the presence of multiple local clusters where the two classes were partly segregated although overlaps were still present. The Dominican data was also seen to contain local class-dependent clustering, albeit with fuzzy boundaries. The Ethiopia and Guinea classes produced broad clusters, which implied class separation with some extent of mixing between the negative and positive classes. For Haiti, a similar result was found, with clusters existing but without any class separation. The Indian data set had the closest clusters, with the negative and positive classes located in a shared central region, indicating low discriminative ability between the classes after the t-SNE transformation. Based on the country-specific combined t-SNE plots, samples clustered into regions with partial overlap but still showed some effect from the geographical location of the data set.

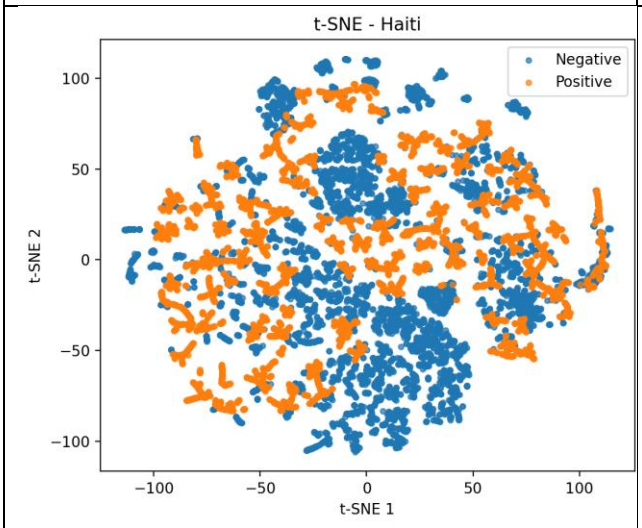

t-SNE - Angola
Negative
Positive
t-SNE 1
t-SNE 2
(a)
t-SNE - Cameroon
Negative
Positive
t-SNE 1
t-SNE 2
(b)
t-SNE - Dominican Republic
Negative
Positive
t-SNE 1
t-SNE 2
(c)
t-SNE - Dominican
Negative
Positive
t-SNE 1
t-SNE 2
(d)


t-SNE - Ethiopia
Negative
Positive
t-SNE 1
t-SNE 2

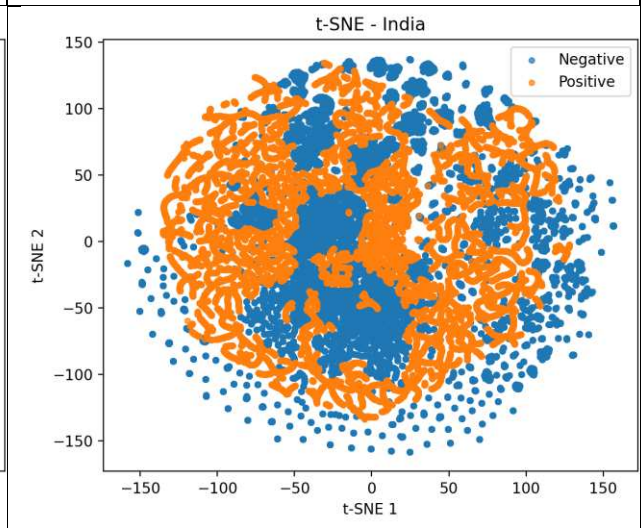
(e)


t-SNE - Guinea
Negative
Positive
t-SNE 1
t-SNE 2

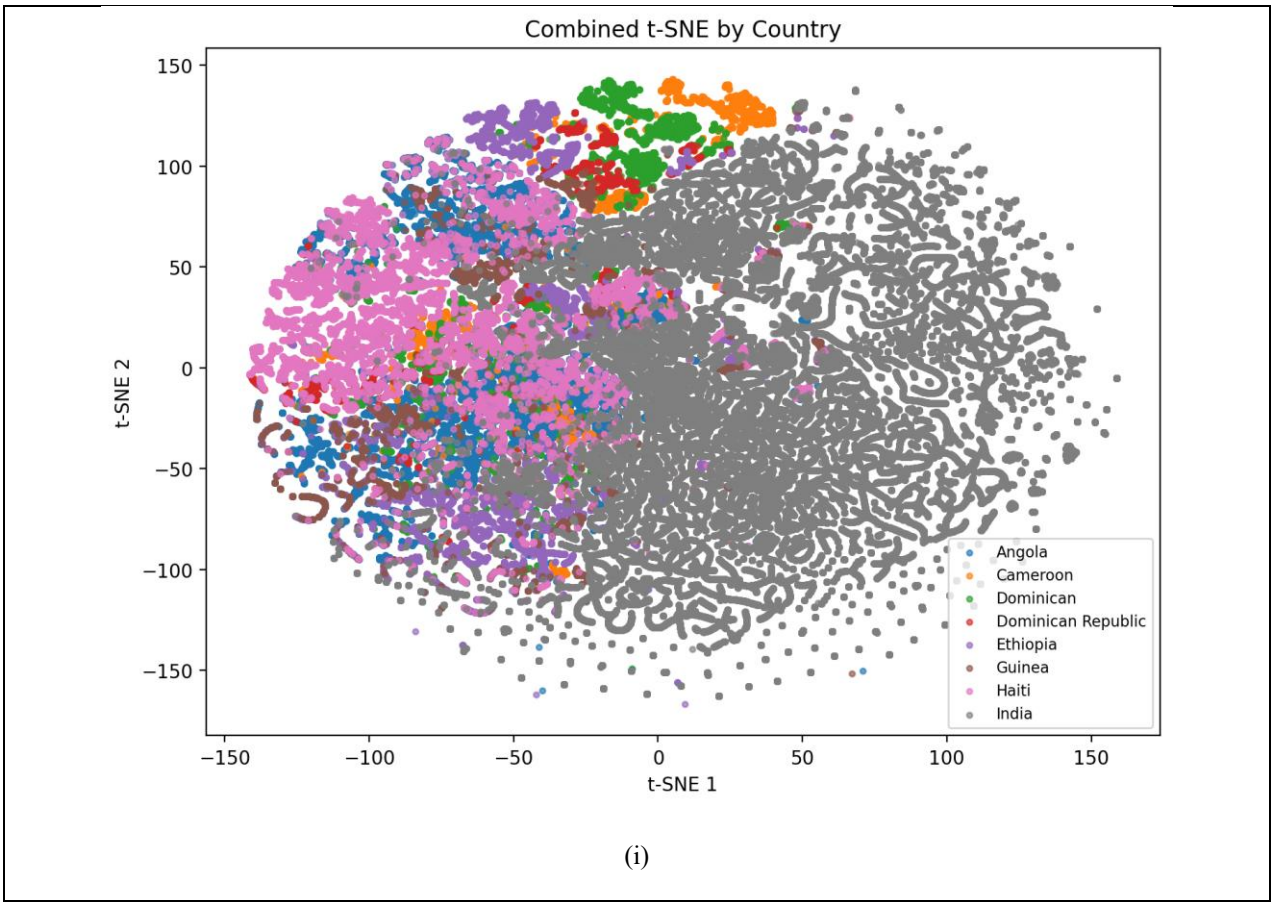
(f)


t-SNE - Haiti
Negative
Positive
t-SNE 1
t-SNE 2

(g)


t-SNE - India
Negative
Positive
t-SNE 1
t-SNE 2

(h)

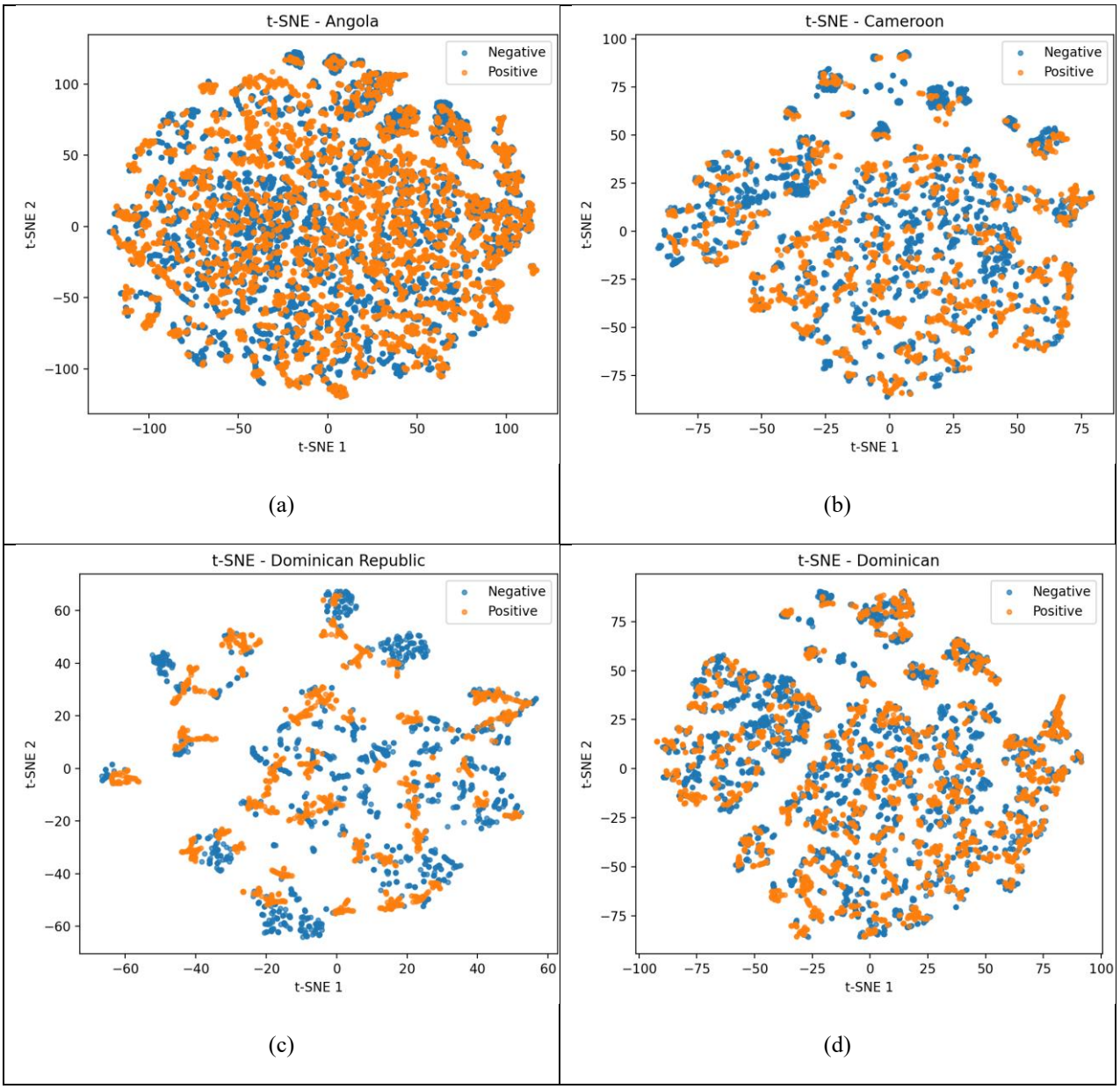


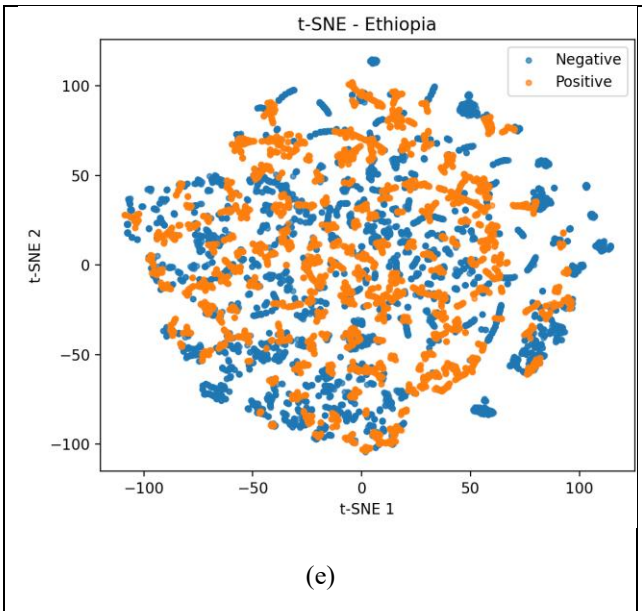
(i)

Figure 11 Obtained t-SNE graphs on HIV prediction for each country (a) Angola, (b) Cameroon, (c) Dominican, (d) Dominican Republic, (e) Ethiopia, (d) Guinea, (g) Haiti, (h) India, and (I) all countries

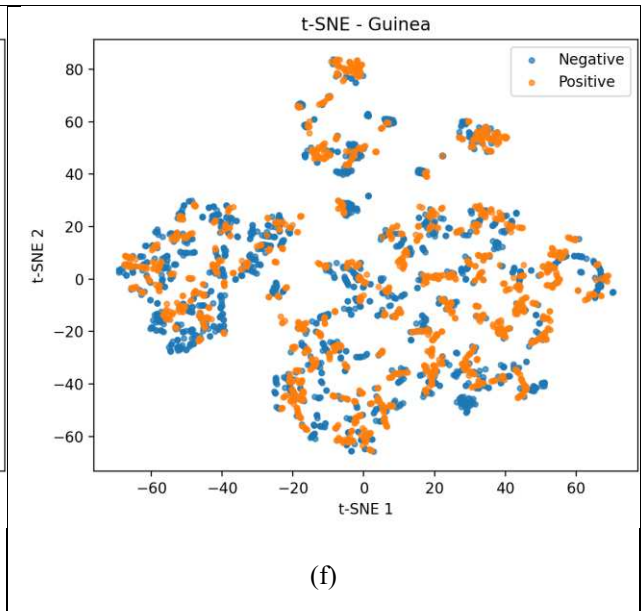
The t-SNE visualizations of STI prediction for country-specific datasets and aggregate datasets are shown in Figure 12. It can be seen from the graphs that the extent of separation depends on the country and negative and positive samples partially intersect each other. In Angola, the classes are evenly distributed throughout the entire area of the visualization, which suggests that there is considerable intersection, and the samples have no clear boundaries. In Cameroon, the samples are represented by multiple local clusters, but negative and positive samples are not separated from each other. For the Dominican Republic, the samples are relatively dispersed and contain many local clusters, while the Dominican dataset consists of tightly packed samples with incomplete mixing. The data for Ethiopia and Guinea show irregular clusters with no clear borders, meaning that the data have weak clustering properties and no distinct border between negative and positive samples. Haiti demonstrates a similar picture with some locally isolated pockets, but still, the negative and positive samples intersect each other. The India dataset seems to be the most mixed one, with

positive samples forming the majority of the visualization and class separability being very weak. The aggregated t-SNE plot by country also shows that the samples keep some clustering by country and considerable overlapping between the datasets. Overall, it can be stated that STIs demonstrate different patterns in different countries and low separability in feature space.

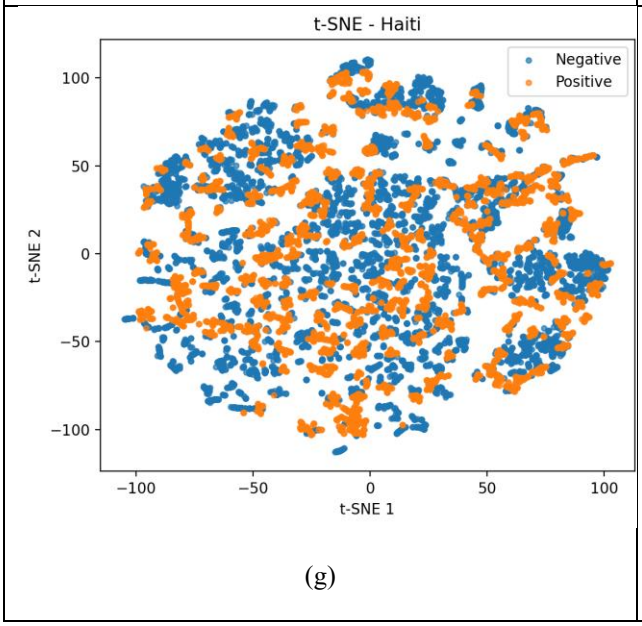



t-SNE - Ethiopia
Negative
Positive
t-SNE 1
t-SNE 2

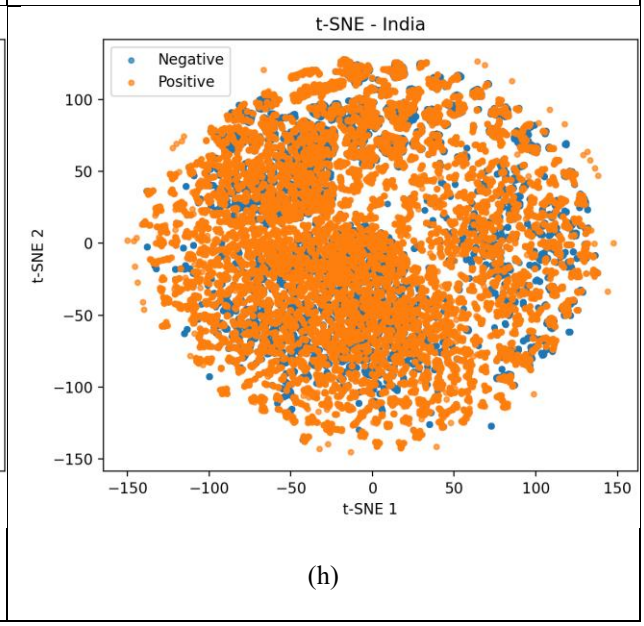
(e)


t-SNE - Guinea
Negative
Positive
t-SNE 1
t-SNE 2

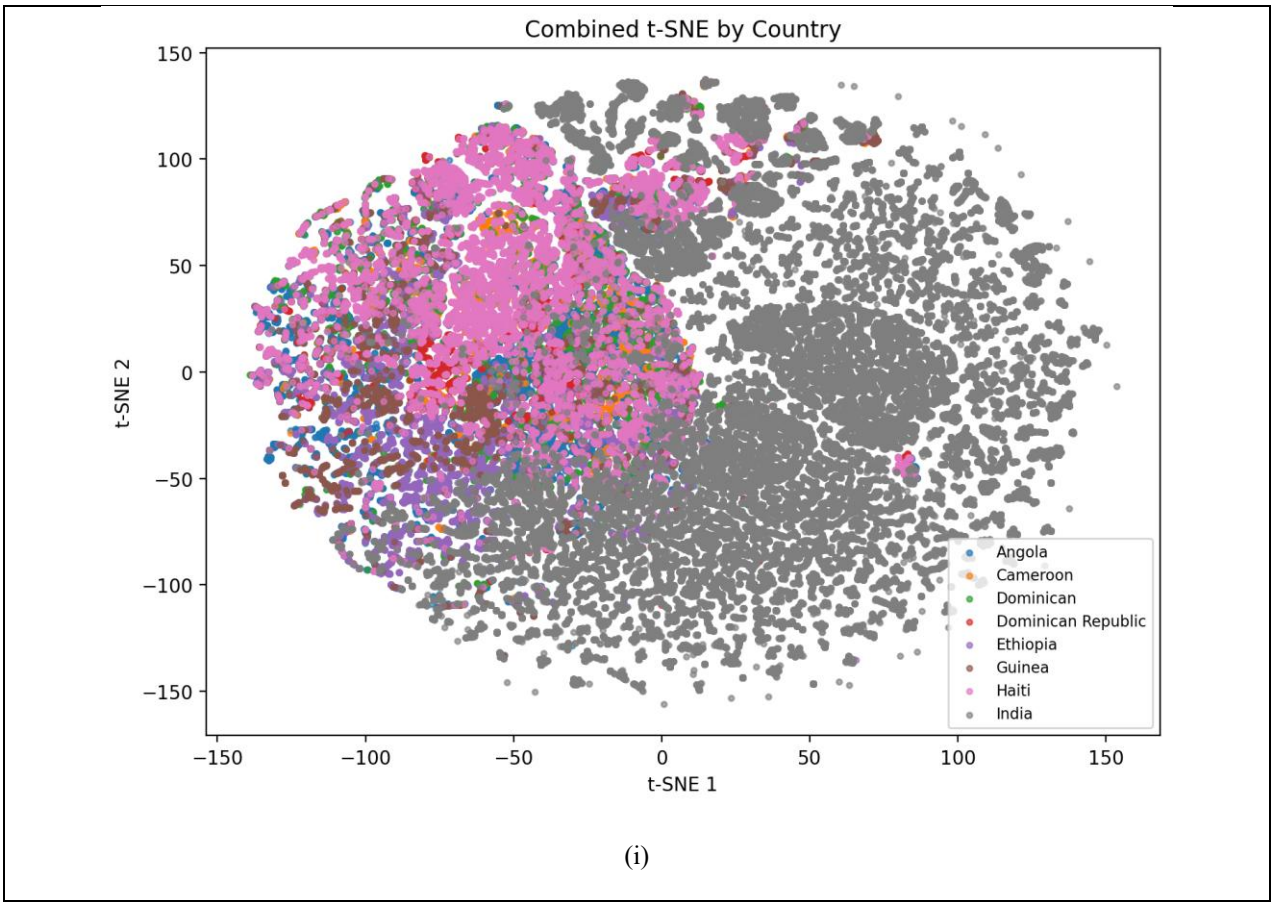
(f)


t-SNE - Haiti
Negative
Positive
t-SNE 1
t-SNE 2

(g)


t-SNE - India
Negative
Positive
t-SNE 1
t-SNE 2

(h)



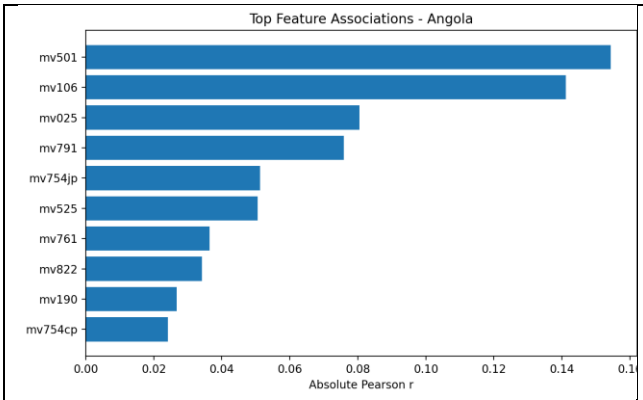
(i)

Figure 12 Obtained t-SNE graphs on STI prediction for each country (a) Angola, (b) Cameroon, (c) Dominican, (d) Dominican Republic, (e) Ethiopia, (d) Guinea, (g) Haiti, (h) India, and (i) all countries

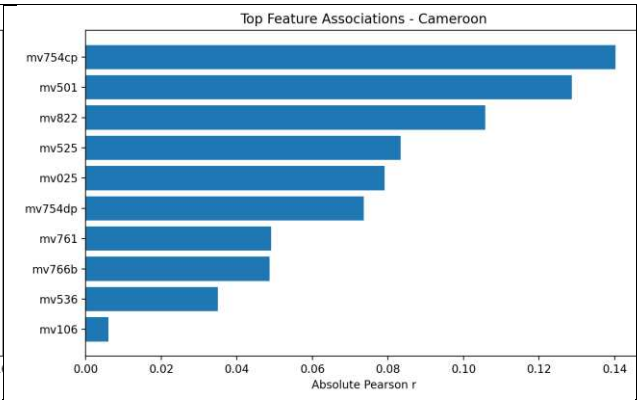
3.5. Explainability Analysis

To improve the interpretability of the proposed framework, a statistical feature association analysis was carried out for each country-specific partition. For the selected predictors, class-wise mean values and several statistical measures were computed, including Pearson correlation, t-test, Mann–Whitney U test, and Cohen's d [30, 31, 32]. The resulting explainability plots were ranked according to the absolute Pearson correlation values, highlighting the features most strongly associated with STI outcome in each country [33]. Figures 13 and 14 present the feature association based explainability analysis for HIV and STI prediction across the eight country-specific datasets, respectively. As illustrated in Figure 13, it can be noted that there was variation in the key contributing factors among different countries, meaning that the risks for HIV infection were based on unique demographic, behavioral, and attitudinal characteristics of each specific setting. For example, in Angola, the current marital status mv501 and the educational level mv106 were

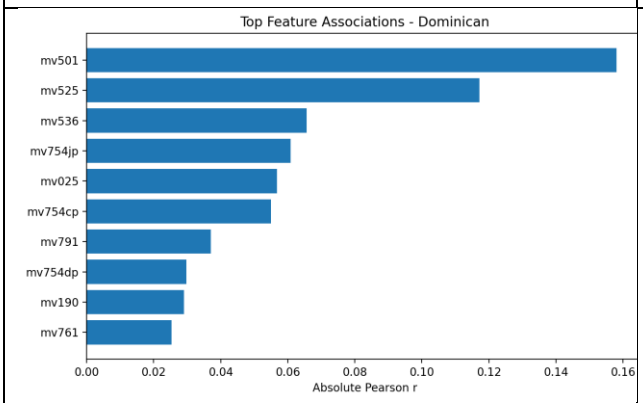
shown to have been the key contributors, followed by age mv012 and condom-related behaviors. This means that the impact of the socio-demographic situation and sexual behavior had affected the model's decisions. For Cameroon, the factor mv754cp, which means "having only one sexual partner," was considered the strongest indicator, together with the current marital status mv501, wife-justified condom negotiation mv822, and sexual activity during the past month mv536. The data from the Dominican Republic show that the current marital status mv501 is also the strongest predictor, with education level mv106 and recent sexual activity mv536 also having significant weight. As far as the partition of the Dominican Republic is concerned, the strongest indicator proved to be education level mv106, followed by condom use at last sex mv761 and wealth index mv190. The Ethiopian case had a unique structure in that regionality mv025 was identified as the most important predictor, significantly outperforming all other variables in explaining STI risk. It seems that regional or contextual characteristics have been particularly important in distinguishing between individuals based on their STI risk. Guinea had a prominent feature, in which condom use at last sex mv761 played the most significant role, followed by current marital status mv501, the wealth index mv190, and recent sexual activity mv536. The case of Haiti presented an interesting structure in which there was no single predictor dominating the analysis, meaning that each of the selected three variables—education level mv106, current marital status mv501, and recent sexual activity mv536—played equally strong roles in defining the dependent variable. Finally, the Indian case presented a relatively concentrated structure, in which regionality mv025 proved to be the dominant variable, while other variables such as recent sexual activity mv536, condom-related behavior mv754cp and mv754p, the wealth index mv190, and education level mv106 made relatively modest contributions. This suggests that with respect to the Indian partition, the model had relied more on contextual/ regional differences rather than on individual behavior related variables. In summary, the findings suggest that the analyzability study revealed that there was indeed a lot of heterogeneity between countries with some data sets more dependent on demographics/socioeconomic characteristics while others more dependent on sexual behavior/condom use.


Top Feature Associations - Angola
mv501
mv106
mv025
mv791
mv754jp
mv525
mv761
mv822
mv190
mv754cp
0.00 0.02 0.04 0.06 0.08 0.10 0.12 0.14
Absolute Pearson r

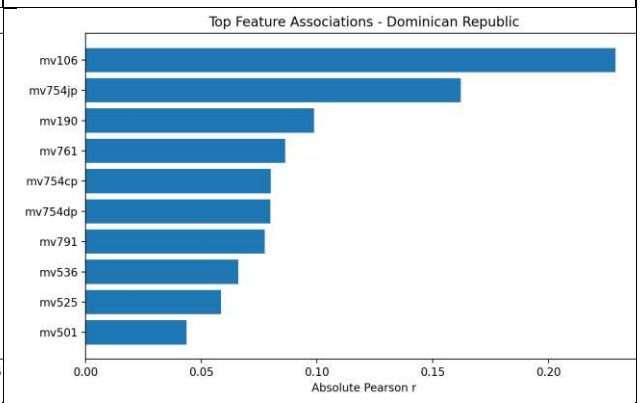
(a)


Top Feature Associations - Cameroon
mv754cp
mv501
mv822
mv525
mv025
mv754dp
mv761
mv766b
mv536
mv106
0.00 0.02 0.04 0.06 0.08 0.10 0.12 0.14
Absolute Pearson r

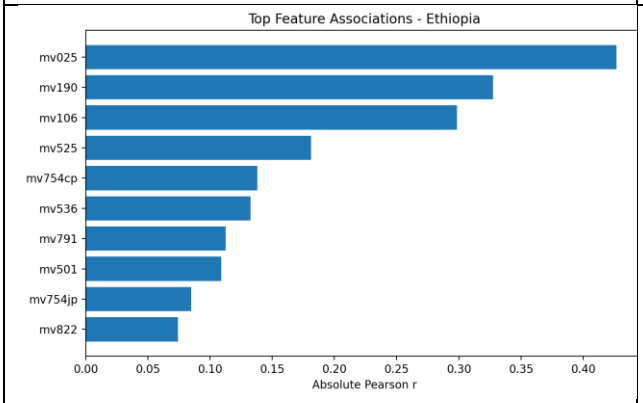
(b)


Top Feature Associations - Dominican
mv501
mv525
mv536
mv754jp
mv025
mv754cp
mv791
mv754dp
mv190
mv761
0.00 0.02 0.04 0.06 0.08 0.10 0.12 0.14 0.16
Absolute Pearson r

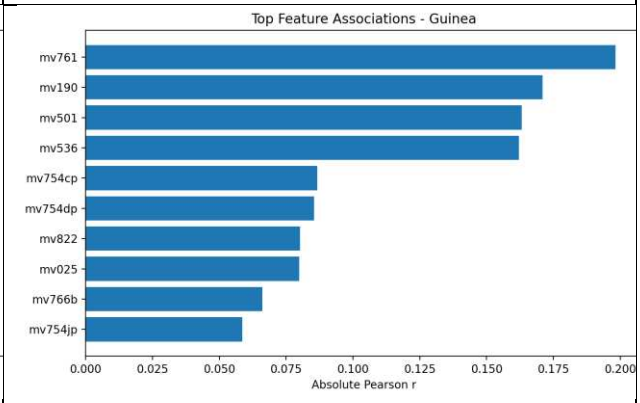
(c)


Top Feature Associations - Dominican Republic
mv106
mv754jp
mv190
mv761
mv754cp
mv754dp
mv791
mv536
mv525
mv501
0.00 0.05 0.10 0.15 0.20
Absolute Pearson r

(d)


Top Feature Associations - Ethiopia
mv025
mv190
mv106
mv525
mv754cp
mv536
mv791
mv501
mv754jp
mv822
0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40
Absolute Pearson r

(e)


Top Feature Associations - Guinea
mv761
mv190
mv501
mv536
mv754cp
mv754dp
mv822
mv025
mv766b
mv754jp
0.000 0.025 0.050 0.075 0.100 0.125 0.150 0.175 0.200
Absolute Pearson r

(f)

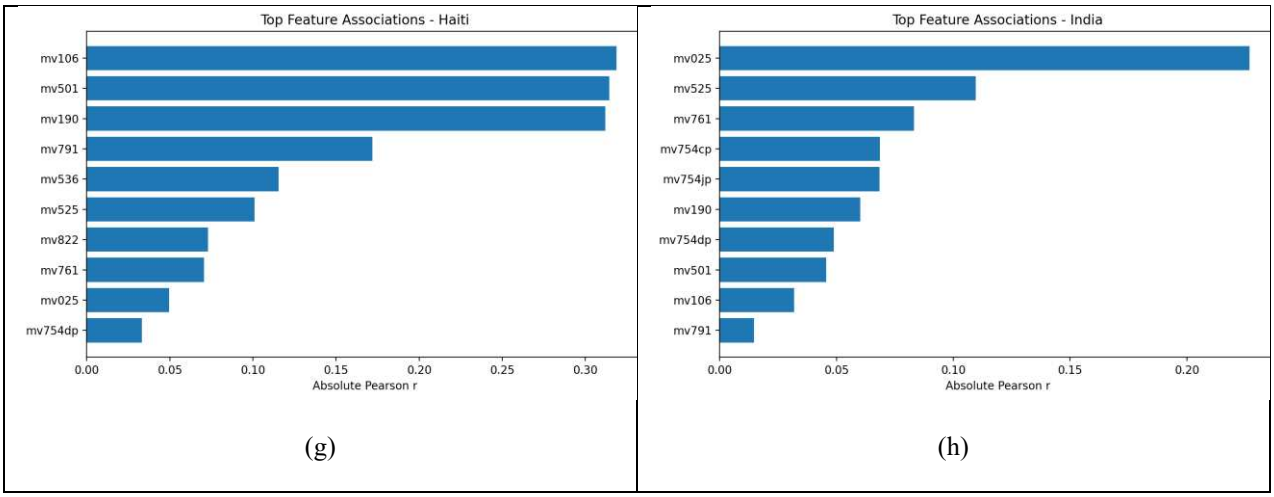
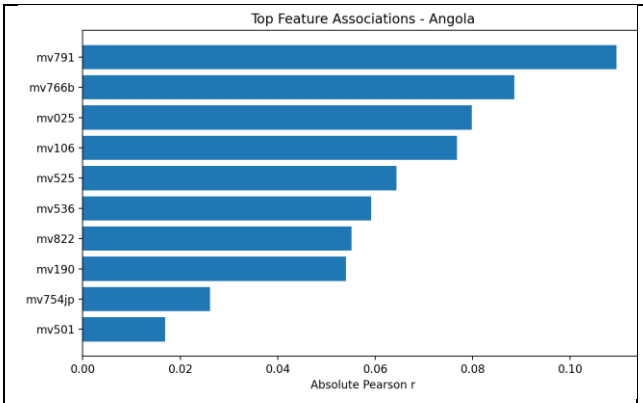
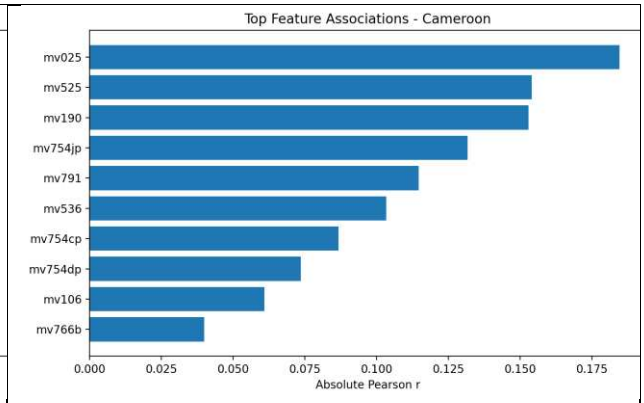


Figure 13 Explainability analysis for HIV prediction for each country (a) Angola, (b) Cameroon, (c) Dominican, (d) Dominican Republic, (e) Ethiopia, (d) Guinea, (g) Haiti, and (h) India

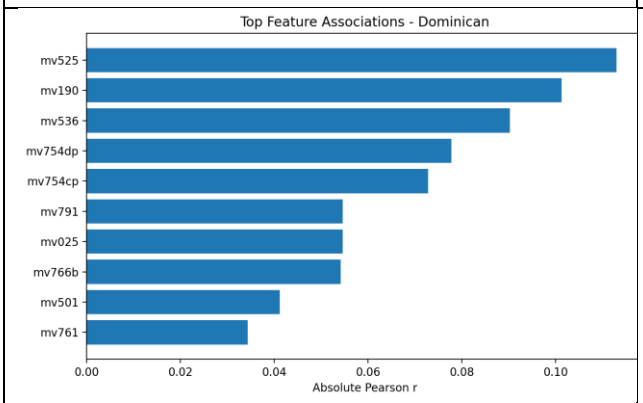
Figure 14 presents the explainability analysis for STI prediction across the country-specific datasets and shows that the most influential features differed noticeably from one country to another. In Angola, mv791 and mv766b had the strongest associations, followed by mv025, mv106, and mv525, indicating that attitudinal, behavioral, and contextual factors all contributed to the prediction pattern. In Cameroon, mv025 emerged as the leading feature, with mv525, mv190, mv754jp, and mv791 also showing relatively strong influence, suggesting that regional and socioeconomic differences played an important role. The Dominican dataset was characterized mainly by mv525, mv190, and mv536, while the Dominican Republic showed a slightly different structure in which mv190 and mv754jp were the most prominent predictors. In Ethiopia, the top-ranked features were mv754jp, mv754cp, mv822, mv525, and mv754dp, pointing to the importance of partner-related and behavioral variables. Guinea showed its strongest association for mv536, followed by mv501, mv754cp, and mv822, whereas Haiti was dominated by mv190, mv791, mv766b, and mv525. India displayed a weaker overall association pattern than the other countries, with mv525 showing the highest value, followed by mv501 and mv791. These findings indicate that STI-related prediction patterns were country-specific, with different demographic, behavioral, and attitudinal variables becoming more influential depending on the local dataset.


Top Feature Associations - Angola
mv791
mv766b
mv025
mv106
mv525
mv536
mv822
mv190
mv754jp
mv501
0.00
0.02
0.04
0.06
0.08
0.10
Absolute Pearson r

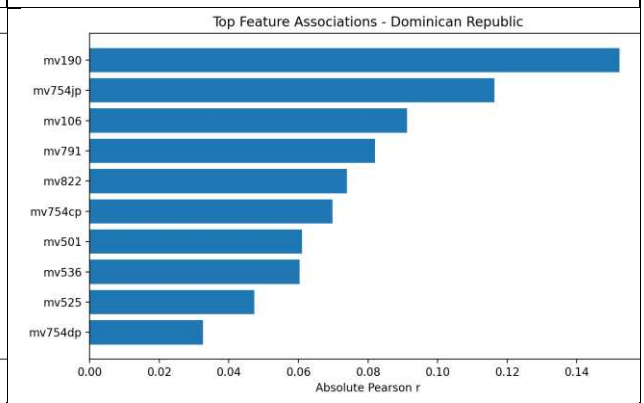
(a)


Top Feature Associations - Cameroon
mv025
mv525
mv190
mv754jp
mv791
mv536
mv754cp
mv754dp
mv106
mv766b
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
Absolute Pearson r

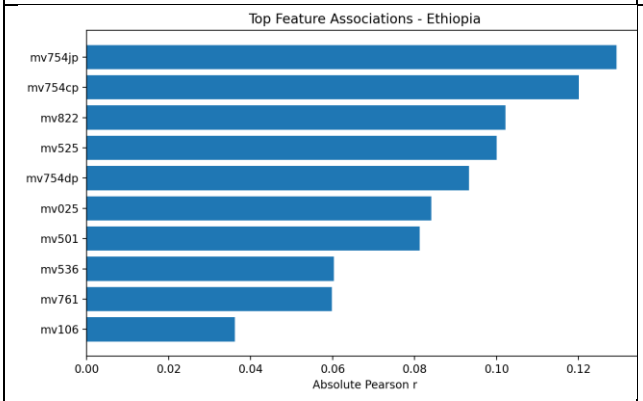
(b)


Top Feature Associations - Dominican
mv525
mv190
mv536
mv754dp
mv754cp
mv791
mv025
mv766b
mv501
mv761
0.00
0.02
0.04
0.06
0.08
0.10
Absolute Pearson r

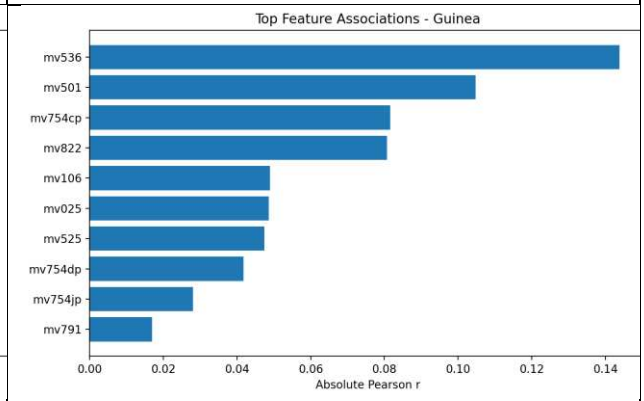
(c)


Top Feature Associations - Dominican Republic
mv190
mv754jp
mv106
mv791
mv822
mv754cp
mv501
mv536
mv525
mv754dp
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
Absolute Pearson r

(d)


Top Feature Associations - Ethiopia
mv754jp
mv754cp
mv822
mv525
mv754dp
mv025
mv501
mv536
mv761
mv106
0.00
0.02
0.04
0.06
0.08
0.10
0.12
Absolute Pearson r

(e)


Top Feature Associations - Guinea
mv536
mv501
mv754cp
mv822
mv106
mv025
mv525
mv754dp
mv754jp
mv791
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
Absolute Pearson r

(f)

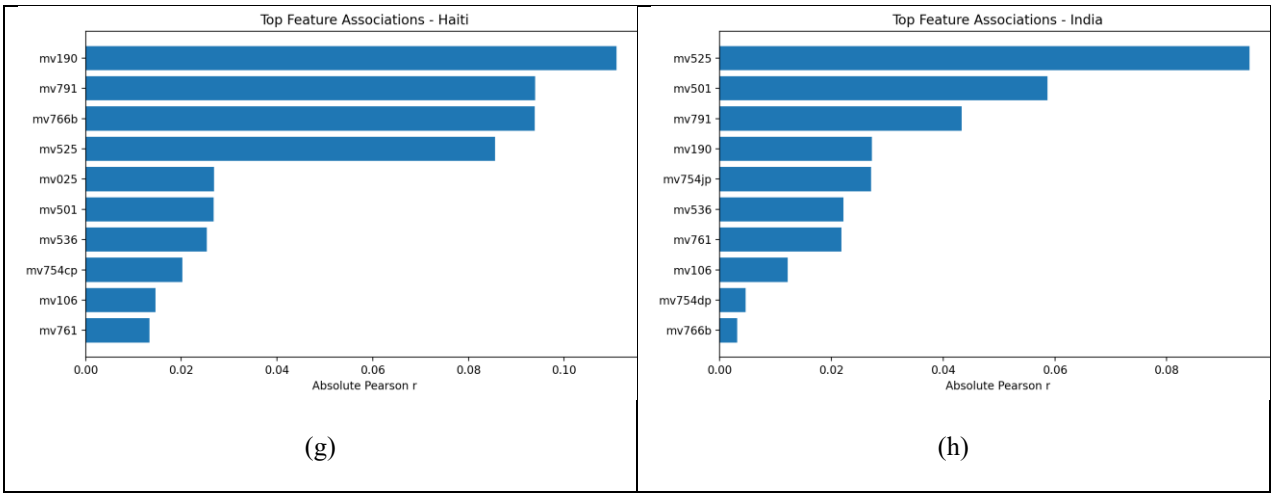


Figure 14 Explainability analysis for STI prediction for each country (a) Angola, (b) Cameroon, (c) Dominican, (d) Dominican Republic, (e) Ethiopia, (d) Guinea, (g) Haiti, and (h) India

4. Discussions

In this paper, we introduce a novel distributed framework of risk assessment for STIs/HIV using the CatBoost algorithm with techniques, such as local ensemble learning, reliability-aware weighted ensemble, threshold calibration, and uncertainty estimation applied to the unified prediction pipeline. The framework has been tested on structured tabular data provided by the DHS Program, where the total amount of 168,459 instances gathered from eight different countries includes demographic, behavioral, and attitude features related to the risk of having an HIV/STI infection. According to our experimental evaluation, carried out via the LOCO protocol, the suggested method is able to demonstrate high-quality performance consistently in the majority of country-specific scenarios. In particular, in case of HIV risk prediction, the framework reached a mean AUC score of 0.9850 and mean accuracy 0.9537, while in STI case - having a mean AUC of 0.9396 and a mean accuracy of 0.9117. These results suggest that combining local predictive structure with weighted global aggregation can provide a robust and interpretable strategy for STI/HIV risk assessment in distributed health data environments.

Table 3 Performance comparison of the proposed method with other results from literature

Paper	Year	Max AUC	Max accuracy	Highlights contribution	Type of learning	Reliability-aware	Predictive uncertainty

							estimation provided?
Ahlström et al. [34]	2019	0.89	N/A	This study demonstrates the feasibility of utilising machine learning techniques, and centralised data to accurately predict an individual's HIV status based on electronic registry data.	Centralized learning	No	No
Xu et al. [35]	2022	0.75	N/A	Authors developed a machine learning-based risk-prediction tool for STI/HIV. This tool can be integrated with digital platforms to increase STI/HIV testing.	Centralized learning	No	No
Bao et al. [36]	2021	0.76	82%	This research demonstrated the advantage of machine learning over the logistic regression model on Australian Men Who has sex with men group.	Centralized learning	No	No
Balzer et al. [37]	2020	0.73	N/A	They used super learner—a type of	Centralized learning	No	No

				machine learning model to demonstrate the potential of using AI in evaluating HIV risk score.			
Nguyen et al. [10]	2023	0.97	93%	We proposed a RFFL approach to assess the STIs/HIV risk. The framework allows for an adaptable aggregation process that can align with the throughput of the communication system.	~~Federated learning~~ Distributed learning	No	No
The proposed	2026	0.99	0.95	A distributed CatBoost framework with local ensemble learning, weighted aggregation, threshold tuning, and uncertainty estimation for STI/HIV risk prediction.	Distributed learning	Yes	Yes

Table 3 summarizes the main characteristics and predictive results of previously published studies together with those of the proposed framework. Earlier studies used a variety of machine learning approaches for STI and HIV risk prediction, but most of them were based on centralized learning settings. Ahlström et al. [34] reported a maximum AUC of 0.89 and demonstrated that machine learning could successfully predict HIV status from electronic registry data. Xu et al. [35] achieved a maximum AUC of 0.75 and developed a

machine learning-based risk prediction tool intended to support digital STI and HIV testing platforms. Bao et al. [36] reported a maximum AUC of 0.76 and an accuracy of 82 percent, showing that machine learning outperformed logistic regression in Australian men who have sex with men cohort. The maximum AUC of 0.73 was obtained by Balzer et al. [37], while they highlighted the usefulness of the super learner technique for assessing the risk score of HIV. On the other hand, Nguyen et al. [10] developed a federated learning methodology and reached the maximum AUC of 0.97 with 93% accuracy, demonstrating that privacy-preserving distributed learning is capable of producing good results. In contrast to these methods, the proposed framework achieved the best performance by reaching the maximum AUC of 0.99 and 95% accuracy while developing a distributed CatBoost architecture based on local ensemble learning, reliable weighted aggregation, threshold selection, and uncertainty estimation. The above comparisons demonstrate that the proposed model not only outperforms some of the previous models in terms of prediction performance but also provides a more flexible framework for STIs and HIV cases.

The advantages of the proposed method are;

- It preserves local predictive structure and user privacy by training models separately on different data partitions instead of forcing all observations into a single pooled model.
- It improves prediction stability through local ensemble learning based on multiple CatBoost models trained with different random seeds.
- It uses reliability-aware weighted aggregation, allowing partitions with stronger validation performance and larger training support to contribute more to the final prediction.
- It provides a more flexible decision mechanism through threshold calibration rather than relying on a fixed default cutoff.
- It includes predictive uncertainty estimation, which adds an interpretable measure of confidence to the final model outputs.
- It demonstrates strong predictive performance for both HIV and STI prediction across multiple country-specific settings.

- It is more suitable than conventional centralized approaches for heterogeneous and distributed health datasets.

The disadvantages of the proposed method are~~;~~:

- The framework is more computationally demanding than a single centralized model because multiple local models and ensemble members must be trained and aggregated. Although the method improves interpretability through uncertainty estimation and feature association analysis, it is still more complex than standard single-model approaches

- Calibration performance was not equally strong across all country-specific partitions, suggesting that probability estimates may still require further refinement in some settings.

5. Conclusions

In this paper, we propose a distributed machine learning algorithm for predicting the risk of sexually transmitted infections (STIs) and HIV infection based on DHS datasets from eight different countries. By applying local models, weighted decision making with consideration of reliability measures, and calibrated thresholds for uncertainty estimation, we demonstrate the efficiency of the proposed method under the LOCO evaluation methodology. Our results show that our framework exhibits high levels of discrimination capabilities when predicting both STIs and HIV infection risks and is relevant even for different partitions of the dataset. We also approve that the possibility of maintaining the local predictive information and its integration via a weighted decision process are a good choice compared to using centralized machine learning methods alone.

Future work should focus on extending the proposed framework to broader and more diverse real-world settings. In particular, it is worthwhile to test the framework using multiple datasets that span more

631 countries and involve different target populations. It is also worth exploring advanced approaches to data
632 aggregation and calibration and implementing fairness-related metrics in order to improve the predictive and
633 robust performance of the framework. Lastly, future studies might include comparisons between the proposed
634 model and other federated and distributed learning frameworks currently in use.

635

636 **Funding Declaration**

637 No funding was received for conducting this study

638 **Ethics and Consent to Participate Declarations**

639 This work does not need any Ethics or Consent to Participant statement since only public databases are used
640 in this work.

641 **References**

642 [1] World Health Organization. (2025, September 10). Sexually transmitted infections (STIs).
643 https://www.who.int/news-room/fact-sheets/detail/sexually-transmitted-infections-%28stis%29

644 [2] UNAIDS. (2025). Global HIV & AIDS statistics — Fact sheet. https://www.unaids.org/en/resources/fact-
645 sheet

646 [3] World Health Organization. (2014). HIV prevention, diagnosis, treatment and care for key populations.
647 World Health Organization. https://iris.who.int/bitstream/handle/10665/128048/9789241507431_eng.pdf

648 [4] World Health Organization. (2021). Guidelines for the management of symptomatic sexually transmitted
649 infections. https://www.who.int/publications/i/item/9789240024168

650 [5] Siyamayambo, C., Phalane, E., & Phaswana-Mafuya, R. N. (2025). Diagnosis and Management of
651 Sexually Transmitted Infections Using Artificial Intelligence Applications Among Key and General

652 Populations in Sub-Saharan Africa: A Systematic Review and Meta-Analysis. Algorithms, 18(3), 151.
653 https://doi.org/10.3390/a18030151.

654 [6] Jaiteh, M., Phalane, E., Shiferaw, Y. A., Jallow, H., & Phaswana-Mafuya, R. N. (2025). The Application
655 of Machine Learning Algorithms to Predict HIV Testing Using Evidence from the 2002–2017 South African
656 Adult Population-Based Surveys: An HIV Testing Predictive Model. Tropical Medicine and Infectious
657 Disease, 10(6), 167. https://doi.org/10.3390/tropicalmed10060167.

658 [7] Ji, X., Tang, Z., Osborne, S. R., Van Nguyen, T. P., Mullens, A. B., Dean, J. A., & Li, Y. (2025). STI/HIV
659 risk prediction model development—A novel use of public data to forecast STIs/HIV risk for men who have
660 sex with men. Frontiers in Public Health, 12, 1511689. https://doi.org/10.3389/fpubh.2024.1511689.

661 [8] Latt, P. M., Soe, N. N., Xu, X., Bao, Y., Lee, D., Ong, J. J., Chow, E. P. F., Zhang, L., & Fairley, C. K.
662 (2025). Anxiety responses and testing intentions among gay and bisexual men using an AI-powered HIV/STI
663 risk assessment tool: a quasi-experimental study. BMC Public Health, 25, 4028.
664 https://doi.org/10.1186/s12889-025-25064-2.

665 [9] Nadarzynski, T., Knights, N., Husbands, D., Graham, C., Llewellyn, C. D., Buchanan, T., Montgomery,
666 I., Soruco Rodriguez, A., Ogueri, C., Singh, N., Rouse, E., Oyebode, O., Das, A., Paydon, G., Lall, G.,
667 Bulukungu, A., Yanyali, N., Stefan, A., & Ridge, D. (2025). Chatbot-assisted self-assessment (CASA): Co-
668 designing an AI-powered behaviour change intervention for ethnic minorities. PLOS Digital Health, 4(2),
669 e0000724. https://doi.org/10.1371/journal.pdig.0000724.

670 [10] Nguyen, T. P. V., Yang, W., Tang, Z., Xia, X., Mullens, A. B., Dean, J. A., & Li, Y. (2024). Lightweight
671 federated learning for STIs/HIV prediction. Scientific Reports, 14, 6560. https://doi.org/10.1038/s41598-
672 024-56115-0.

673 [11] Smerina, D. R., & Pearlman, A. M. (2025). The Intersection of Artificial Intelligence, Wearable Devices,
674 and Sexual Medicine. Current Urology Reports, 26, 14. https://doi.org/10.1007/s11934-024-01244-2.

[12] Yu, W., Chen, Y., Liu, B., Huang, E., & Guo, F. (2026). Findings from a transformer-based prediction model: social norms and confidentiality are associated with STIs risk in British men and women. BMC Public Health, 26, 1097. https://doi.org/10.1186/s12889-026-26602-2.

[13] Zhong, S., & Ghosh, J. (2003). A unified framework for model-based clustering. Journal of machine learning research, 4(Nov), 1001-1037.

[14] Ando, R. K., Zhang, T., & Bartlett, P. (2005). A framework for learning predictive structures from multiple tasks and unlabeled data. Journal of machine learning research, 6(11).

[15] Zhou, F., Pan, H., Gao, Z., Huang, X., Qian, G., Zhu, Y., & Xiao, F. (2021). Fire prediction based on catboost algorithm. Mathematical Problems in Engineering, 2021(1), 1929137.

[16] Hussain, S., Mustafa, M. W., Jumani, T. A., Baloch, S. K., Alotaibi, H., Khan, I., & Khan, A. (2021). A novel feature engineered-CatBoost-based supervised machine learning framework for electricity theft detection. Energy Reports, 7, 4425-4436.

[17] Huang, G., Wu, L., Ma, X., Zhang, W., Fan, J., Yu, X., ... & Zhou, H. (2019). Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions. Journal of Hydrology, 574, 1029-1041.

[18] Miyoshi, T., & Yamane, S. (2007). Local ensemble transform Kalman filtering with an AGCM at a T159/L48 resolution. Monthly Weather Review, 135(11), 3841-3861.

[19] Joshi, A., Saggar, P., Jain, R., Sharma, M., Gupta, D., & Khanna, A. (2021). CatBoost—An ensemble machine learning model for prediction and classification of student academic performance. Advances in Data Science and Adaptive Analysis, 13(03n04), 2141002.

[20] Salazar, J. C., Weber, P., Nejjari, F., Sarrate, R., & Theilliol, D. (2017). System reliability aware model predictive control framework. Reliability Engineering & System Safety, 167, 663-672.

[21] Dong, L., Wu, W., Guo, Q., Satpute, M. N., Znati, T., & Du, D. Z. (2019). Reliability-aware offloading and allocation in multilevel edge computing system. IEEE Transactions on Reliability, 70(1), 200-211.

[22] Kang, Z., Zhao, X., Peng, C., Zhu, H., Zhou, J. T., Peng, X., ... & Xu, Z. (2020). Partition level multiview subspace clustering. Neural Networks, 122, 279-288.

[23] Masoudnia, S., & Ebrahimpour, R. (2014). Mixture of experts: a literature survey. Artificial Intelligence Review, 42(2), 275-293.

[24] Freeman, E. A., & Moisen, G. G. (2008). A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. Ecological modelling, 217(1-2), 48-58.

[25] Zou, Q., Xie, S., Lin, Z., Wu, M., & Ju, Y. (2016). Finding the best classification threshold in imbalanced classification. Big Data Research, 5, 2-8.

[26] Malinin, A., & Gales, M. (2018). Predictive uncertainty estimation via prior networks. Advances in neural information processing systems, 31.

[27] Draper, D. (1995). Assessment and propagation of model uncertainty. Journal of the Royal Statistical Society Series B: Statistical Methodology, 57(1), 45-70.

[28] Ji, X., Tang, Z., Osborne, S. R., Van Nguyen, T. P., Mullens, A. B., Dean, J. A., & Li, Y. (2025). STI/HIV risk prediction model development—A novel use of public data to forecast STIs/HIV risk for men who have sex with men. Frontiers in Public Health, 12, 1511689.

[29] Cai, T. T., & Ma, R. (2022). Theoretical foundations of t-sne for visualizing high-dimensional clustered data. Journal of Machine Learning Research, 23(301), 1-54.

[30] Nachar, N. (2008). The Mann-Whitney U: A test for assessing whether two independent samples come from the same distribution. Tutorials in quantitative Methods for Psychology, 4(1), 13-20.

[31] Wall Emerson, R. (2023). Mann-Whitney U test and t-test. Journal of Visual Impairment & Blindness, 117(1), 99-100.

[32] Bishara, A. J., & Hittner, J. B. (2012). Testing the significance of a correlation with nonnormal data: comparison of Pearson, Spearman, transformation, and resampling approaches. Psychological methods, 17(3), 399.

[33] Ghanem, V. G. (2025). Socio-Behavioral and Spatial Determinants of HIV/AIDS Incidence in Ghana: An Ecological Cross-Sectional Study with Explainable Machine Learning.

[34] Ahlström, M. G., Ronit, A., Omland, L. H., Vedel, S., & Obel, N. (2019). Algorithmic prediction of HIV status using nation-wide electronic registry data. EClinicalMedicine, 17.

[35] Xu, X., Ge, Z., Chow, E. P., Yu, Z., Lee, D., Wu, J., ... & Zhang, L. (2022). A machine-learning-based risk-prediction tool for HIV and sexually transmitted infections acquisition over the next 12 months. Journal of clinical medicine, 11(7), 1818.

[36] Bao, Y., Medland, N. A., Fairley, C. K., Wu, J., Shang, X., Chow, E. P., ... & Zhang, L. (2021). Predicting the diagnosis of HIV and sexually transmitted infections among men who have sex with men using machine learning approaches. Journal of Infection, 82(1), 48-59.

[37] Balzer, L. B., Havlir, D. V., Kamya, M. R., Chamie, G., Charlebois, E. D., Clark, T. D., ... & Petersen, M. L. (2020). Machine learning to identify persons at high-risk of human immunodeficiency virus acquisition in rural Kenya and Uganda. Clinical Infectious Diseases, 71(9), 2326-2333.