Article

A machine learning model for early detection of sexually transmitted infections

Juma Shija, **Judith Leo**, **Elizabeth Mkoba**
The School of Computational and Communication Science and Engineering, The Nelson Mandela African Institution of Science and Technology, Arusha, Tanzania
E-mail: shijaj@nm-aist.ac.tz, judith.leo@nm-aist.ac.tz, elizabeth.mkoba@nm-aist.ac.tz

Received 18 September 2024; *Accepted 25 October 2024*; *Published online 25 November 2024*; *Published 1 June 2025*

Abstract
Sexually transmitted infections (STIs) are diseases transmitted mostly through unprotected sex with an infected partner. STIs can be transmitted to an infant before or during childbirth. More than one million sexually transmitted infections (STIs) are acquired every day worldwide. In the most recent years, the prevalence of STIs reached approximately 20% among Tanzanian older adults living in metropolitan areas. If not treated properly and on time, STIs can have severe consequences, including infertility, sterility, increased susceptibility to more serious diseases such as the Human Immunodeficiency Virus (HIV), and even death. However, stigma and shame associated with STIs remain significant barriers to proper diagnosis and timely treatment, leading many patients to face increased risks. The purpose of this paper is to present a machine-learning model for early detection of sexually transmitted infections that was developed. The developed model can be deployed into health systems for self-diagnosis to remove communication barriers between sexual health clinics and STI patients. The study used a quantitative research method and got its dataset of 13,335 records from the Government of Tanzania Health Operations Management Information System (GoT-HoMIS) in areas with many STI cases. This was done by using surveys and questionnaires to get the data. The dataset was split into a 70%:15%:15% ratio for training, testing, and validation, respectively, and five machine learning algorithms were evaluated: AdaBoost, Support Vector Machine, Random Forest, Decision Tree, and Stochastic Gradient Descent. Based on evaluation metrics, the AdaBoost model was identified as the best-performing model, achieving an accuracy of 97.45%, an F1 score of 97.7%, and the Receiver Operating Characteristics Area Under the Curve (ROC-AUC) with a higher true positive rate and a lower false positive rate. The study recommends integrating a machine learning model into healthcare systems to detect STIs early, improve medical care, reduce disease progression, and remove stigmatisation barriers. Also, it can provide insights into infection patterns, allowing practitioners to adapt their responses. Machine learning-based solutions in mobile apps and telemedicine systems promote early testing and treatment.

Keywords machine learning; sexually transmitted infections; artificial intelligence; stigmatisation.

Computational Ecology and Software
ISSN 2220-721X
URL: http://www.iaees.org/publications/journals/ces/online-version.asp
RSS: http://www.iaees.org/publications/journals/ces/rss.xml
E-mail: ces@iaees.org
Editor-in-Chief: WenJun Zhang
Publisher: International Academy of Ecology and Environmental Sciences

1 Introduction

According to the World Health Organization (WHO), more than one million sexually transmitted infections (STI) are acquired every day worldwide. Each year, there are an estimated 376 million new infections with 1 of 4 STIs such as chlamydia, gonorrhoea, syphilis, and trichomoniasis (WHO, 2019). According to the literature, it is believed that the same trend or at a higher rate is happening in Tanzania; however, it is difficult to track the data because most patients feel ashamed and stigmatised from attending clinics and sharing their symptoms and STI status. For example, the following Fig. 1 briefly describes the monthly attendance for patients with STIs who visited the hospital in one of the hospitals in Tanzania (Naomi C.A. Juliana et al., 2020). The following are the long forms of the abbreviated words in Fig. 1, whereby CT means Chlamydia trachomatis, NG means Neisseria gonorrhoea, TV means Trichomonas vaginalis and Syph means Syphilis.

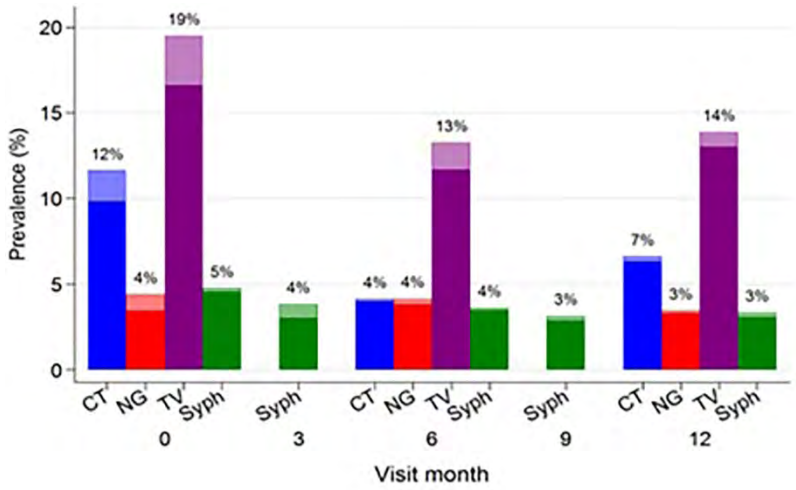

Fig. 1 Attendance per month for patients with sexually transmitted infections.

Therefore, due to the challenge of STI data in terms of its collection, several researchers have tried to develop ICT-based solutions; however, most existing studies have focused on developing systems for predicting STIs and other diseases without considering that STI patients often face stigma, making it difficult for them to report or share their symptoms, disclose their status, or even visit hospitals for diagnosis (Patoli, 2007). Hence, most of the existing systems are working with poor-quality data, and therefore it is difficult to provide effective predictions and inform policymakers and decision-makers.

By definition, stigma is a psychosocial barrier in the treatment of STIs (Lichtenstein, 2003). Due to stigmatisation, most STI patients tend to buy some medicine without any prescription from medical experts (Naomi C.A. Juliana et al., 2020). The tendency of keeping quiet and treating STIs with the wrong medicine or without a doctor's prescription, most of the time, has resulted in chronic diseases such as cancer, infertility, pelvic inflammatory disease, ectopic pregnancy, and adverse outcomes of pregnancy, including pre-term delivery and low birth weight for most of these patients (Aral, 2001). As a result, the impact of the disease becomes severe (Gerbase and Zemouri, 2020). Therefore, there is a need for an automated machine learning model-equipped system that can interact with the patients on a real-time basis without the need to visit the hospital to enable STI patients to communicate their symptoms and receive required recommendations on time. This study, therefore, proposed to develop a tool to detect STIs through the use of an automated machine

learning system. Hence the tool enables patients to interact with medical doctors without the need for physical meetups; the patients can share their symptoms and interact with the system. Then, the patients can be advised on what to do based on their symptoms and explanations. As a result, the patients are not required to explain themselves again to the doctor, and hence, the feeling of stigmatisation and discrimination can be reduced among patients with STIs.

2 Related Works

Machine learning is useful nowadays in detecting diseases like cancer, hence the early prevention of the disease (Kumar et al., 2019). For example, the study done by Radhika and her colleagues used a convolutional neural network (CNN) for the classification of malignancy in mammograms, and the model had an accuracy of 93%. Whereby another study also used Machine Learning (ML) algorithms such as decision trees to analyse the symptoms of diseases and predict them (Radhika et al., 2020). The model had an accuracy score of 90%, which was the best performance for their model. In addition, Mihaylov and his colleagues (Mihaylov et al., 2019) applied ML models to understand survival prognosis in breast cancer. Five classifiers were used in this study to find the most accurate survival prognosis. These were Support Vector Regression (SVR) linear, Lasso, kernel Ridge Decision Tree Regression (DTR), and Multilayer Perceptron (MLP) regressor. Furthermore, in developing an expert system for men and genital problems diagnosis and treatment (Naser and Al-Hanjori, 2016), ML techniques were applied as well.

Moreover, other researchers have developed expert systems to give awareness on STIs only and then connected their applications with the doctors' phone numbers or email addresses; as a result, patients are afraid of meeting with the doctors or talking to them.

Thompson et al. in their research developed a system that diagnoses patients and gives out the type of STI a patient is suffering from. In their work, a patient was required to answer some questions related to the symptoms; then finally, a system gave the final answer to the patient, such as "You have been diagnosed; you are suffering from syphilis." The system ends by giving some predicted diseases to the patients; as a result, the patients can also decide to find some false medical treatment from the pharmacy or be afraid of visiting the health centres or hospitals for further treatment.

In addition, Thompson and his colleagues developed a system framework in which there was no external link, such as a hospital where the patient could get assistance.

In their study, Książek et al. (2019) conducted a study using ten machine learning algorithms: K-Nearest neighbours (KNN), RegLog, Gaussian Naïve Bayes (gaussNB), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Random Forest, MLP1, and Support Vector Machine (SVM). The proposed model achieved the best accuracy and F-1 score values of 0.8849 and 0.8762, respectively.

3 Methods and Materials

3.1 Data collection

The datasets were collected in Tanzania from Mbeya and Dar es Salaam regions with high STI prevalence, as shown in Fig. 2, with the help of the GoT-HoMIS. Surveys and questionnaires were conducted to acquire facts on the presence and types of STIs that occur frequently in the areas. The dataset comprises patients who attended specific hospitals from 2020 to 2021. In total, the study included 13,335 datasets for patients who were diagnosed with and without STI symptoms.

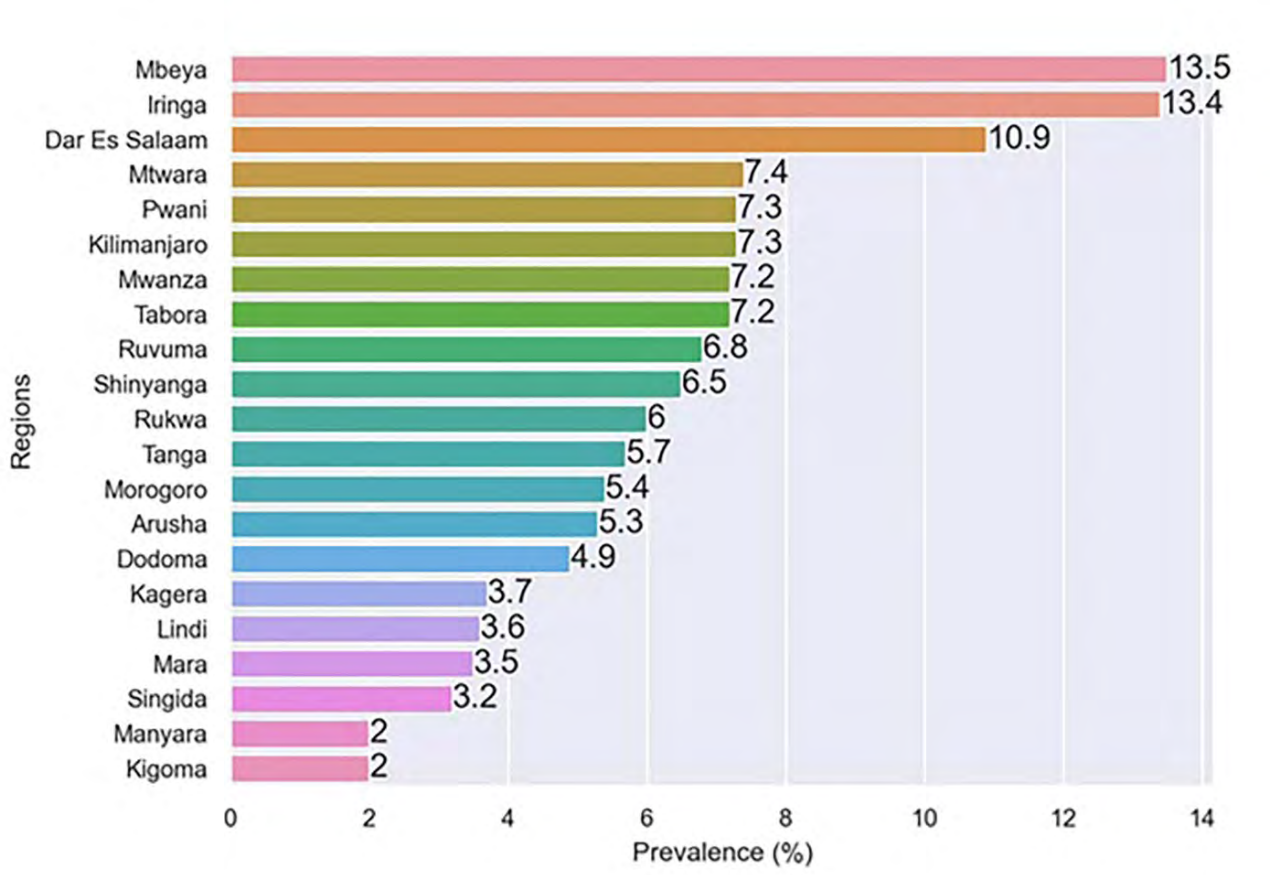

Fig. 2 HIV and STD Prevalence in Tanzania based on the 2021 dataset.

3.2 Data cleaning and analysis

With the help of TensorFlow Data Validation (TFDV), missing data were observed in the dataset, as shown in Fig. 3.

Fig. 3 Missing data in the dataset.

The dataset was cleaned by imputing the missing data and removing other diseases that were not STIs. Some features were removed due to the large number of missing data, such as symptom 6 and symptom 7, as shown in Fig. 3. Therefore, the study remained with 8,900 cleaned datasets that were related to STI diagnosis. Then, the dataset was converted to a decision table format for ML algorithms to be implemented. Table 1 illustrates the sample of the dataset in a decision table format.

Table 1 Decision table.

Patient	Sex	Age	Symptom1	Symptom2	Symptom3	Symptom4	Symptom-n	Lab Test Result
P1	F	21	YES	YES	NO	YES …	NO	POSITIVE
P2	M	15	NO	NO	NO	NO …	NO	NEGATIVE
P3 . . .	M	32	YES	NO	YES	NO …	NO	POSITIVE
P_n	F	18	YES	NO	NO	NO …	YES	POSITIVE

The decision table was reduced by removing the unnecessary columns such as patient name, sex, and age. From the dataset, it was observed that most of the patients who attended the hospitals during that period and tested STIs positive were teenagers, and few were at the middle age of twenty to thirty years old, but very few were above thirty years old as shown in Fig. 4 and Table 2. Additionally, most patients who attended the STI clinic were pregnant women who must be tested for STIs to prevent unborn children from infections.

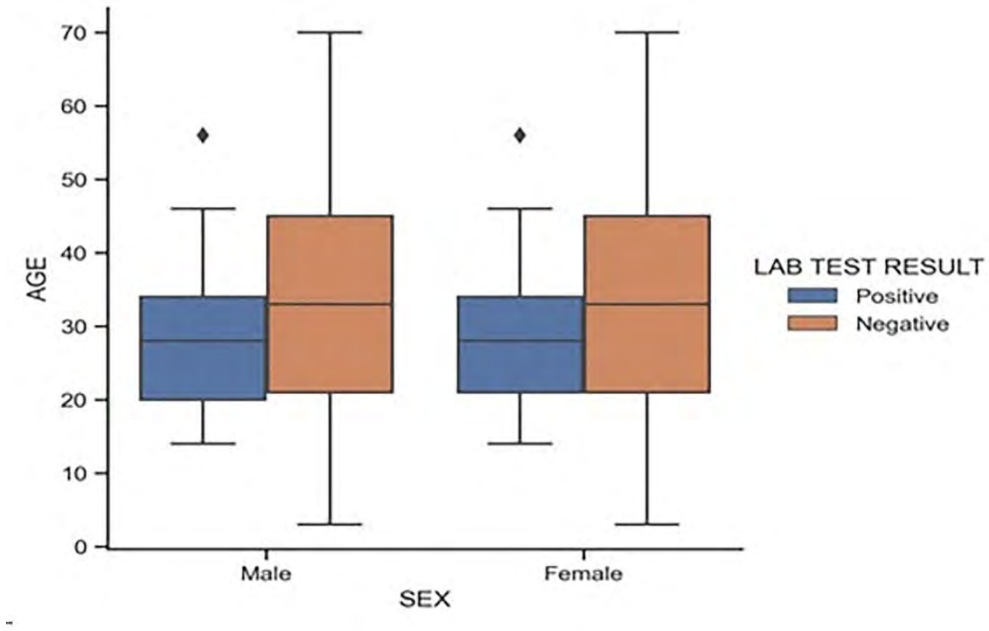

Fig. 4 Distribution of STI patients according to their age and sex.

Based on the data collected, it was observed that there was no difference in the number of males and females attending hospitals. It was also observed that most females attended hospitals after their male partners were positively diagnosed with STIs and inform them on the infections.

Table 2 Distribution of patients according to their age and sex.

	Male Positive Blue (Years)	**Male Negative Orange (Years)**	**Female Positive Blue (Years)**	**Female Negative Orange (Years)**
Median	30	35	30	35
Interquartile Range	20 - 35	22 - 45	22 - 35	22 - 45
Whiskers	15 - 45	9 - 70	15 - 45	9 - 70
Outliers	60	Nil	60	Nil

3.3 Development of the machine learning model

For the development part of the model, five ML algorithms were used to predict the STIs' infection status. These five ML algorithms included AdaBoost, Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), and Stochastic Gradient Descent (SGD).

3.3.1 AdaBoost algorithm

Ada Boost combines many weak classifiers to form strong classifiers (Beja-Battais, 2023) as shown in Equation 1.

$$\mathbf{H(x) = sign\left(\sum_{t=1}^{T} \alpha_t . h_t(x)\right)} \quad (1)$$

where h_t= weak classifier at iteration t

α_t = weight of weak classifier

$H(x)$ = weighted sum of the weak classifiers and the sign function determines the final class label.

3.3.2 Random Forest

For classification, the random forest uses the majority vote of each tree (Louppe, 2014). Using M trees, the final prediction $\hat{y}$ for input x is provided by equation 2:

$$\mathbf{\hat{y} = mode\{h_i(x): i = 1, 2, \dots, M\}} \quad (2)$$

where $\boldsymbol{h_i(x)}$ is the prediction of the i tree and M is the total number of trees.

3.3.3 Support Vector Machine (SVM)

Support Vector Machines (SVMs) are the type of supervised learning model used in classification and regression. They are well-known for their performance in high-dimensional environments and when the number of dimensions exceeds the number of samples (Deng et al., 2012). Here are the key mathematical equations and concepts that underpin SVMs.

For this study, the nature of the dataset was binary, hence the linear SVM was used.

The purpose of a training dataset of N points is to discover the hyperplane that best separates the classes

$y_i \in \{-1, 1\}$ and $x_i \in \{-1, 1\}$

The decision function is used to reflect how far the input data point is from the hyperplane, which is the optimal operator between the two classes. The decision function is given by equation 3:

$$\mathbf{f(x) = w^T x + b} \qquad (3)$$

where

w = weight vector (normal to the hyperplane).

x = feature vector of the input data point.

b = bias term (intercept).

The hyperplane which is the decision boundary is given by Equation 4

$$\mathbf{w^T x + b = 0} \qquad (4)$$

Points on each side of the hyperplane satisfy the inequalities $w^T x + b > 0$ and $w^T x + b < 0$ respectively.The SVM aims to find the hyperplane that maximises the margin, which is given by Equation 5.

$$\mathbf{Margin} = \frac{\mathbf{2}}{\|\mathbf{w}\|} \qquad (5)$$

Then for the hard margin, the optimisation problem for linearly separable data is given by $\min_{w,b} \frac{1}{2}\|w\|^2$ subject to the constraints $y_i(w^T x_i + b) \geq 1 \; \forall i = 1,2,\ldots,N$.

3.3.4 Decision Tree

A decision tree is a flowchart-like tree structure in which each internal node represents an attribute test, each branch reflects the test's conclusion, and each leaf node (terminal node) stores a class label (Myles et al., 2004).

3.3.5 Stochastic Gradient Descent

Stochastic Gradient Descent (SGD) is an optimisation approach that minimises a function by iteratively updating the model's parameters. It is notably popular for training machine learning models on huge datasets and deep learning. Here is a breakdown of the mathematical concepts underpinning SGD.

The gradient descent aims to minimise the objective function $f(\theta)$, where θ is the model parameter. To minimise the objective function, it is required to move in the opposite direction of the gradient ($\nabla f(\theta)$), where the gradient is the partial derivative of the objective function.

In gradient descent, always the parameters are required to be updated using the rule depicted in Equation 6, where η is the learning rate.

$$\boldsymbol{\theta_{t+1} = \theta_t - \eta \nabla f(\theta_t)} \qquad (6)$$

SGD computes the gradient from a single data point (or a small set of data points). This adds stochasticity (randomness) to the updates, which might help avoid local minima and accelerate convergence.

The updating rule of SGD is the same as that of gradient, as shown in Equation 6, but gradient uses only one data point x or a small set at each iteration. For the SGD, the input $x^{(i)}$and output $y^{(i)}$ are all considered from the training set (Equation 7).

$$\boldsymbol{\theta_{t+1} = \theta_t - \eta \nabla f(\theta_t; x^{(i)}, y^{(i)})} \qquad (7)$$

After selecting the five ML algorithms, the dataset was separated into training, testing, and validation data at a ratio of 70%, 15%, and 15%, respectively. The training dataset was fitted to the models for training, and then after training the model was tested using the testing dataset. Then, the model produces the efficacy for the

whole dataset.

For the hardware, an Intel Core i7-powered computer with a memory of 8GB was used during model training. The environment used was an anaconda where the Jupyter Notebook was used. Additionally, TensorFlow and Scikit-learn libraries for ML model development were used.

3.4 Evaluation criteria

Models evaluation was done by finding the accuracy and confusion matrix of the models as well as the Receiver Operating Characteristics Area Under the Curve (ROC-AUC). The confusion matrix was applied at this stage to visualise the performance of algorithms. Table 3 briefly describes features involved in the confusion matrix.

Whereby, by definition, True Positive (TP) means that predicted positive and it is true; for example, a patient is positively predicted and is positive, and True Negative (TN) means that predicted negative and it is true, for example, a patient is negatively predicted and is negative. A False Positive (FP) sometimes called type I error, means that a predicted positive is false; for example, a patient is positively predicted but is negative, and a False Negative (FN) sometimes called type II error, means that a predicted negative and is false, for example, a patient is negatively predicted but is positive (Valero-Carreras et al., 2023).

Table 3 Confusion Matrix (Susmaga, 2004).

	Predicted STIs Positive	Predicted STIs Negative
Actual STIs Positive	True Positive (TN)	False Negative (FN)
Actual STIs Negative	False Positive (FN)	True Negative (TN)

Additionally, the evaluation criteria considered calculations of the accuracy, precision, and F1-score as illustrated in the following equations 8, 9, and 10. By definition, the F_1-score is a tool for assessing both recall and precision at the same time. It also employs harmonic mean instead of arithmetic mean in its computation processes (Valero-Carreras et al., 2023).

- Accuracy was calculated using the following formula as shown in equation 8:

$$\mathbf{Accuracy} = \frac{\mathbf{TP+TN}}{\mathbf{TP+TN+FP+FN}} \quad (8)$$

- Precision was calculated using the following formula as shown in equation 9:

$$\mathbf{Precision} = \frac{\mathbf{TP}}{\mathbf{TP+FP}} \quad (9)$$

- F1-score was calculated using the following formula as shown in equation 10:

$$\mathbf{f_1 - score} = \frac{\mathbf{2TP}}{\mathbf{2TP+FP+FN}} \quad (10)$$

3.5 Validation criteria

Although the models were validated by testing using the test data, to overcome overfitting, it was not enough

to test the model on a single portion of data, hence the concept of cross-validation cut across. K-Fold Cross Validation was used to validate the models, where the k factor selected was 10.

4 Results and Discussion

4.1 Results of model performance

Table 4 shows the results of the five models in terms of their accuracy, precision, and F1 score. Then the confusion matrix was applied to easily visualise the best-performing model in terms of accuracy, and as a result, AdaBoost was selected to be the best-performing model (Table 4 and Fig. 5).

Table 4 Models' accuracy.

Algorithms	Accuracy	Precision	F1-Score
AdaBoost	97.45	97.53	97.7
Decision Tree	97.39	98.56	97.6
Random Forest	97.25	97.64	97.53
Stochastic Gradient Descent	96.95	98.05	97.26
Support Vector Machine	96.95	98.09	97.63

Using the confusion matrix method, Fig. 5 shows a short explanation of the math that was used to find the AdaBoost model's accuracy, precision, and F1 score.

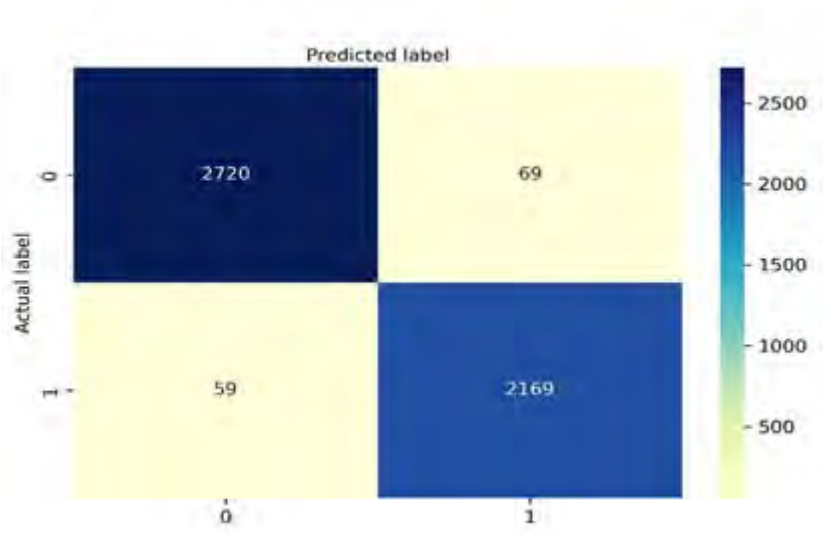

Fig. 5 Confusion Matrix Heat Map for AdaBoost Model.

Based on the obtained values, such as TP = 2720, FP = 69, FN = 59, and TN = 2169, as shown in Fig. 4. The 97.45% accuracy, 97.53% precision, and 98.35% specificity were calculated as shown in equations 11, 12, and 13.

- The result of accuracy was calculated using the following formula, as shown in Equation 11:

$$\text{Accuracy} = \frac{\mathbf{4889}}{\mathbf{5017}} = 0.9745=97.45\% \quad (11)$$

- The result of precision was calculated using the following formula, as shown in Equation 12:

$$\text{Precision} = \frac{\mathbf{2720}}{\mathbf{2789}} = 0.9753=97.53\% \quad (12)$$

- The results of F1-score was calculated using the following formula, as shown in Equation 13:

$$\text{F1 Score} = \frac{\mathbf{5440}}{\mathbf{5568}} = 0.9770 = 97.70\% \quad (13)$$

Based on the performance results, Fig. 6 shows the performance analysis of the AdaBoost Model.

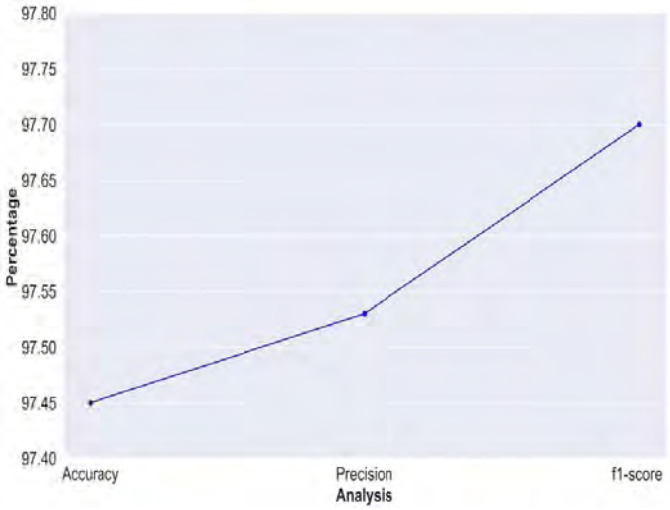

Fig. 6 Performance Analysis of AdaBoost Model in terms of its Accuracy, Precision, and F_1-Score.

Then, the selected model based on its accuracy was validated using a 10-fold cross-validation method as shown in Fig. 7 and Table 5. The average performance for the 10-fold cross-validation method was 97.53%, which outperformed the initial performance of the AdaBoost model. Additionally, it was observed that there was no big difference among the folds; hence, this indicates that training the algorithm on the whole dataset and deploying it in production may produce comparable performance.

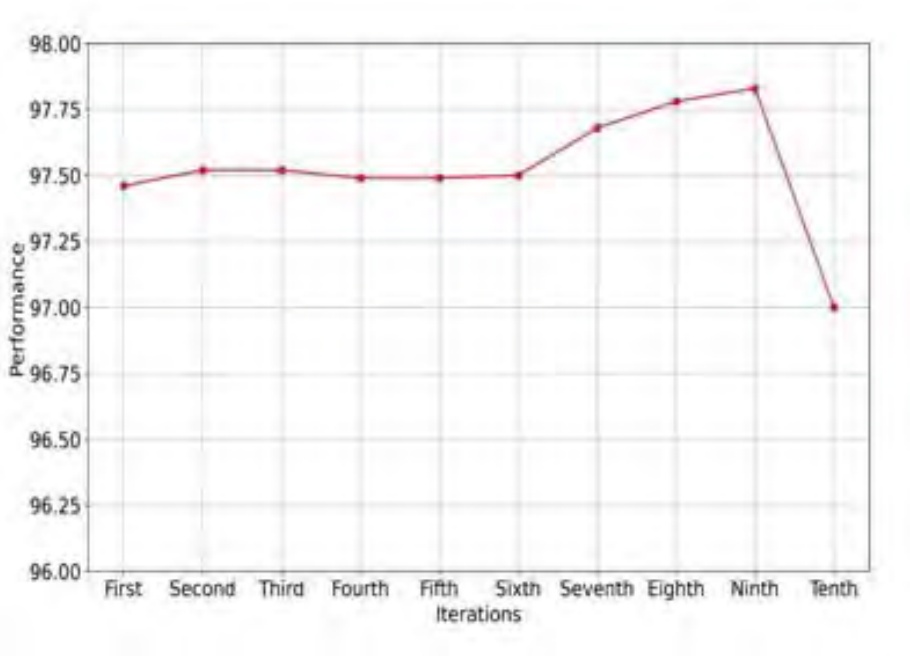

Fig. 7 Results of 10-fold cross validation performance for AdaBoost Model.

Table 4 Results of 10-fold cross validation performance for AdaBoost Model.

Iterations	**1st**	**2nd**	**3rd**	**4th**	**5th**	**6th**	**7th**	**8th**	**9th**	**10th**
Performance	97.46	97.52	97.52	97.49	97.49	97.50	97.68	97.78	97.83	97.00

- Receiver Operating Characteristic Curve (ROC)

The receiver operating characteristic curve (ROC) is a graphical representation of a binary classifier's performance across different classification thresholds. The curve compares the hypothetical true positive rates (TPR) with the false positive rates (FPR). The ROC AUC score is a single value that summarizes the classifier's performance at all conceivable classification thresholds (Kumar & Indrayan, 2011). To calculate the score, measure the area under the ROC curve. The left side of the curve represents the more "confident" thresholds: a higher threshold results in lesser recall and fewer false positive mistakes. The extreme point occurs when both recall and FPR are zero. In this situation, there are no true detections and no erroneous ones. The right side of the curve depicts the "less strict" scenarios when the threshold is low. Both recall and false positive rates increase, eventually reaching 100%. If the model consistently makes accurate predictions, the TPR is always 1.0 and the FPR is zero. It detects all situations and never generates false alarms. The ROC curve shown in Fig. 8 was from the best-performed algorithm; the AUC is approximately 97.45, which was the accuracy of the AdaBoost algorithm.

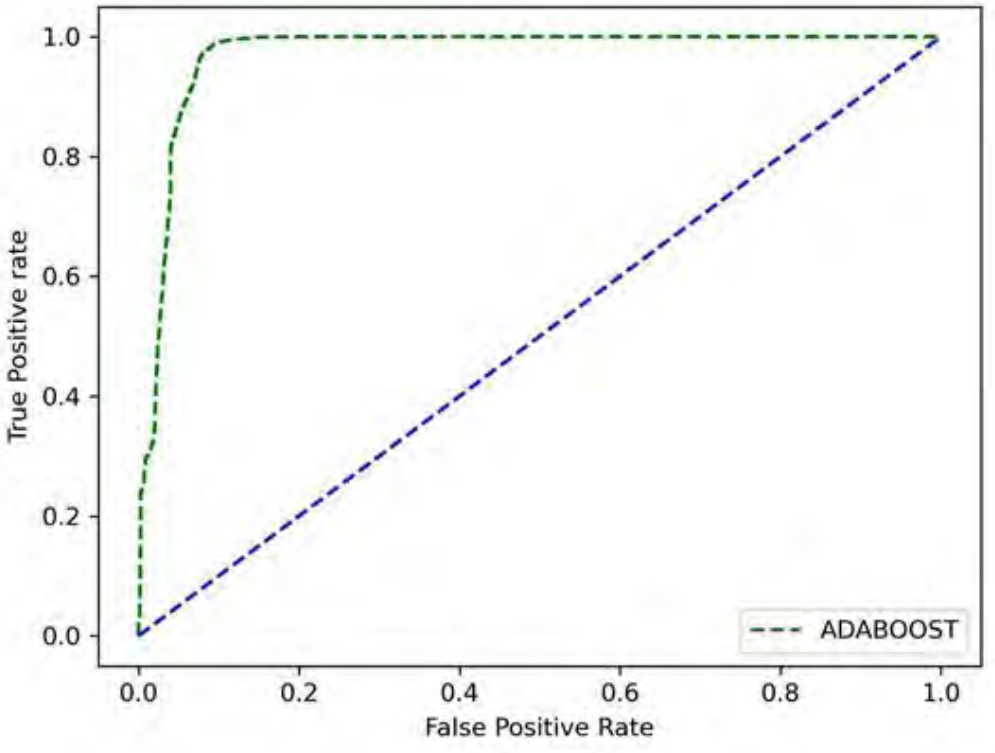

Fig. 8 ROC curve.

4.2 Discussion

This study investigated the design and implementation of a machine-learning model for the early detectionof STIs. The general objective was to use the predictive capacity of machine learning techniques to assist in faster diagnosis and intervention, ultimately improving public health outcomes. The findings showed that several machine learning algorithms can reliably predict the presence of STIs based on clinical data, showing their potential to revolutionise existing diagnostic techniques. The findings underscore the importance of comprehensive data collection and the use of advanced analytics in improving the precision and efficiency of STI detection.All five algorithms; AdaBoost, Support Vector Machine, Random Forest, Decision Tree, and Stochastic Gradient Descent showed significant improvement in the dataset, and the AdaBoost model was able to predict the probability of STI presence in the new dataset.

5 Conclusion and Recommendation

In many African countries, STIs are considered embarrassing diseases, and many individuals are afraid of being stigmatised if they visit a hospital. Sexually transmitted infections are linked to a high rate of malignancy and infertility as a result of late treatment. Therefore, early diagnosis of STIs is critical for effective therapy and prevention of the related risk associated with late treatment. This study, through the use of machine learning techniques, has shown clearly that it is possible to develop a system that can assist patients with the early diagnosis and recommendation for effective early intervention of STIs among patients. The study therefore recommends the review of the healthcare system and policies to integrate the developed model to enable STI patients to receive the initial recommendation without feeling stigmatised during the process.

6 Future Works

Based on the outcomes of this study on constructing a machine learning model for the early detection of sexually transmitted infections (STIs), many areas for future research are suggested to improve the model's performance, generalisability, and real-world applicability.

- Data Source Expansion: Future research should include more diverse and broad datasets from other geographic regions, healthcare settings, and demographic groupings. This will increase the model's generalisability and robustness across populations.
- Incorporation of Additional Predictors: Including new predictive elements, such as genetic markers, lifestyle factors, and more detailed sexual behaviour data, may improve the model's accuracy. Working with epidemiologists and healthcare specialists can help find new relevant variables.
- Model Optimisation and Complexity: Experimenting with more advanced machine learning algorithms, such as ensemble methods, deep learning approaches, and neural networks, may help to increase predictive performance. Hyperparameter tweaking and model optimisation strategies should also be investigated.
- Validation in Clinical Settings: Pilot studies and clinical trials are essential for validating the model in real-world healthcare settings. The model will be tested with actual patient data to determine its practical utility, accuracy, and impact on early diagnosis and treatment outcomes.
- Deploying the system on various platforms, including using Unstructured Supplementary Service Data (USSD), so that the system can be accessible using feature phones.

References

Aral SO. 2001. Sexually transmitted diseases: magnitude, determinants and consequences. International Journal of STD & AIDS, 12(4): 211-215

Beja-Battais P. 2023. AdaBoost: A theoritical review.

Deng N, Tian Y, Zhang C. 2012. Support Vector Machines: Optimization Based Theory, Algorithms, and Extensions. CRC Press, USA

Gerbase AC, Zemouri C. 2020. Global epidemiology of sexually transmitted infections in the twenty-first century: Beyond the Numbers. In: Sexually Transmitted Infections. 3-12, Springer

Książek W, Abdar M, Acharya UR, Pławiak, P. 2019. A novel machine learning approach for early detection of hepatocellular carcinoma patients. Cognitive Systems Research, 54: 116-127

Kumar A, Mukherjee S, Luhach AK. 2019. Deep learning with perspective modeling for early detection of malignancy in mammograms. Journal of Discrete Mathematical Sciences and Cryptography, 22(4): 627-643

Kumar R, Indrayan A. 2011. Receiver operating characteristic (ROC) curve for medical researchers. Indian Pediatrics, 48: 277-287

Lichtenstein B. 2003. Stigma as a barrier to treatment of sexually transmitted infection in the American deep south: issues of race, gender and poverty. Social Science and Medicine, 57(12): 2435-2445

Louppe G. 2014. Understanding random forests: From theory to practice. arXiv preprint, arXiv: 1407.7502.

Mihaylov I, Nisheva M, Vassilev D. 2019. Application of machine learning models for survival prognosis in breast cancer studies. Information, 10(3): 93

Myles AJ, Feudale RN, LiuY, Woody NA, Brown SD. 2004. An introduction to decision tree modeling. Journal of Chemometrics: A Journal of the Chemometrics Society, 18(6): 275-285

Naomi CA. Juliana, Saikat Deb, Sander Ouburg, Aishwarya Chauhan, Jolein Pleijster, Said M Ali, Servaas A. Morré, Sunil Sazawal EA. 2020. The Prevalence of Chlamydia trachomatis and Three Other Non-Viral Sexually Transmitted Infections among Pregnant Women in Pemba Island Tanzania

Naser SSA, Al-Hanjori MM. 2016. An expert system for men genital problems diagnosis and treatment. International Journal of Medicine Research, 1

Patoli AQ. 2007. Role of Syndromic Management using Dynamic Machine Learning in Future of e-Health in Pakistan. Studies in Health Technology and Informatics, 129(1): 601

Radhika S, Shree SR, Divyadharsini VR, Ranjitha A. 2020. Symptoms based disease prediction using decision tree and electronic health record analysis. European Journal of Molecular and Clinical Medicine, 7(4): 2060-2066

Susmaga R. (2004. Confusion matrix visualization. In: Intelligent Information Processing and Web Mining. 107-116, Springer

Thompson T, Sowunmi O, Misra S, Fernandez-Sanz L, Crawford B, Soto R. 2017. An expert system for the diagnosis of sexually transmitted diseases–ESSTD. Journal of Intelligent and Fuzzy Systems, 33(4): 2007-2017

Valero-Carreras D, Alcaraz J, Landete M. 2023. Comparing two SVM models through different metrics based on the confusion matrix. Computers and Operations Research, 152: 106131

WHO. 2019. Sexually transmitted infections (STIs). https://www.who.int/news-room/fact-sheets/detail/sexually-transmitted-infections-(stis)