Tanzania Journal of Engineering and Technology 2022, **41(3)**:64-71
OPEN ACCESS articles distributed under Creative Commons Attribution Licence [CC BY-ND]
Websites: https://ajol.org/tjet; https://tjet.udsm.ac.tz

ISSN 1821-536X (**print**); ISSN 2619-8789 (**electronic**)
https://doi.org/10.52339/tjet.v41i3.845

*Special Issue – Sixth Tanzania Annual ICT (TAIC) Conference, 26 – 28, October 2022, Zanzibar, TANZANIA*

# A Robust Random Forest Prediction Model for Mother-to-Child HIV Transmission Based on Individual Medical History

**Rebecca B. Chaula[1†] and Godfrey N. Justo[2]**

[1]Directorate of Information and Communication Technology, Muhimbili University of Health and Allied Sciences, P.O. BOX 65001, Dar es Salaam, Tanzania.
[2]Department of Computer Science and Engineering, University of Dar es Salaam, P.O. BOX 33335, Dar es Salaam, Tanzania.
[†]Corresponding author: chaularebecca@gmail.com; https://orcid.org/0000-0001-9812-6602

## ABSTRACT

**Human Immunodeficiency Virus (HIV) continues to be a leading cause of mortality and reduces manpower throughout the world. HIV transmission from mother to child is still a global challenge in health research. According to UNAIDS, in every 7 girls, 6 are found to be newly infected among adolescents whereby 15-24 years are likely to be living with HIV which is the maternal age and likely to transfer to the child. Machine learning methods have been used to predict HIV/AIDS transmission from mother to child but left behind some important considerations including the use of patient-level information and techniques in balancing the dataset which may impact models' performance. A robust prediction model for mother-to-child HIV/AIDS transmission is vital to alleviate HIV/AIDS detrimental effects. The Random Forest Machine Learning method was employed based on features from the individual medical history of HIV-positive mothers. A total of 680 balanced data tuples were used for model development using the ratio of 75:25 for training and testing the dataset. The Random Forest model outperformed the most commonly used learning algorithms achieving the performance of 99% accuracy, recall and F1-score of 0.99 and an error of 0.01, thus improving the prediction rate.**

ARTICLE INFO

**First Submitted**: June 30, 2022

**Revised:** Sept. 30, 2022

**Presented:** Oct. 26 – 28, 2022

**Accepted**: Nov. 18, 2022

**Published:** Nov. 30, 2022

**Keywords:** *Machine learning/AI, prediction model, mother-to-child HIV/AIDS transmission, data imbalance.*

## INTRODUCTION

The human immunodeficiency virus (HIV) is a major global health emergency, which affects all regions of the world, causing millions of deaths and suffering to millions more (WHO, 2003). HIV is a virus that weakens the human immune system of an individual exposing the body to several opportunistic infections. HIV continues to be a leading cause of mortality and reduces manpower throughout the world, and 68% of its effect is in sub-Saharan Africa (UNAIDS, 2019).

Mother-to-child transmission (MTCT) is the most important mode of HIV-1 acquisition among infants and children and it can occur in utero, intrapartum, and postnatal through breastfeeding. Great progress has been made in preventing MTCT through the use of antiretroviral regimens during gestation, labor/delivery and breastfeeding (Ellington et al., 2018). HIV transmission from mother to child is the vertical transmission from mother to child and this transmission can occur during pregnancy, delivery and breastfeeding if care is not taken and if the mother did not receive HIV treatment when she was pregnant (WHO&CDC, 2008). As we have the Prevention of mother-child transmission (PMTCT) program, the program helps to reduce transmission whereby the mother will have early ART and therefore reduces the number of infected infants. About 93% of pregnant women living with HIV were receiving effective ART in 2018, compared to 75% in 2010.

Machine learning methods have received growing attention in the health domain for medical diagnosis, medical case prediction and others forms of decision-making support. The work by (James et al., 2018) used the Resilient Backpropagation Neural Network (RBNN) algorithm to predict HIV transmission from mother to child. This method works well only when all the features are kept constant, otherwise, the accuracy of the prediction weakens significantly. RBNN also requires a large amount of data to lend a high prediction rate, thus, higher computational expense. Further, the work by (James et al., 2018), considered the during pregnancy, delivery and breastfeeding factors, along with the features: of CD4 count, delivery mode and ART drug used; but singled out other features. However, the study used accuracy metric only and the results achieved 95%.

### Related work

Machine Learning (ML) has been widely used in past studies for the prediction of HIV transmission from different perspectives. In (Nan & Gao, 2018; Girum, et al., 2018) ML models were designed to monitor HIV/AIDS transmission trends. Studies by (Shen et al., 2016; Ekpenyong et al., 2019) used ML models to predict HIV/AIDS patient drug response. In (Campos Coelho et al., 2019) a forecast model is designed for HIV-1 mother-to-child transmission prevalence using machine learning and in (James et al., 2018) the RBNN algorithm is used to design a prediction model for MTCT of HIV. The study by (Negussie Deyessa, 2015) assessed the determinants of mother-to-child HIV transmission.

This study builds on (James et al., 2018), where the prediction for HIV transmission from mother to child is proposed by considering the during pregnancy, delivery and breastfeeding factors. However, the study left behind individual features of the mother and child status which can mimic the real environment and the factors that cause the transmission of HIV from mother to child.

### Problem statement

Previously proposed models for mother-to-child transmission made use of aggregate data, paid little attention to consideration of balancing the dataset and did not consider combined metrics for model validation, a factor that may impact the model's performance. This study advocate balancing of the dataset, use of disaggregated dataset (individual features) and comprehensive model validation as a strategy to improve model robustness and performance.

### Objectives

The objective of the study is to:

(i) Determine the relevant features for HIV transmission from mother to child.

(ii) Develop an improved prediction model for MTCT of HIV using relevant features.

(iii) Evaluate the performance of the model.

## METHODS AND MATERIALS

The research revolved around main prediction model development tasks which include secondary data collection, data pre-processing, data exploration/analysis, model development and model validation. Machine Learning methods reviewed from the literature, feature engineering tools used to identify important features and tools for the model development used to build and validate the prediction models.

Data were collected from National AIDS Control Program (NACP) CTC2 historical database. The prediction model development is modeled as a supervised learning problem that involves the use of secondary labeled data (Al-Zaiti et al., 2020). Through data pre-processing data inconsistency and unclean data were resolved as model performance also depends on quality data (Grafberger et al., 2021).

Different performance metric measurements were employed for model evaluation, which included precision, accuracy, recall and F1-score rates. Further, the confusion matrix (Xu et al., 2020), cross-validation and area under the curve were used to validate the model to warrant performance robustness.

## RESULTS AND DISCUSSIONS

This section presents the results of this study. It summarizes the results obtained from the feature analysis and selection, and model development includes the splitting of the dataset and evaluation of the model in the prediction of HIV Transmission from Mother-To-Child. Tables and figures were used to show the development and performance of the model.

The *Scikitlearn* library from the python software was used to develop and evaluate the performance of the model using the metric measurement. The individual medical history of a mother and the status of the child using historical data from 2010 to 2020 were used for the model training and testing since the number of patients with HIV increased to visit the hospital from 2010 according to UNAIDS. Data were pre-processed and balanced to have clean data, the cleaned data were used in the model development to avoid overfitting and underfitting the model. From the individual medical history of a mother, additional features were also used in building the model, which had a high chance of improving the performance of the model, and metric measurements were used to evaluate the performance of the model.

### Data Pre-Processing

A total of 5295 data tuples were collected of which 34 tuples were from infected children and 5261 tuples were from uninfected children. The dataset was imbalanced at a rate of 0.60%, whereby the minority class (Positive 1) are the children with infection and the majority are the ones with no infection (Negative 0), as summarized in Table 1.

**Table 1: Imbalanced dataset**

| Class | Count |
|---|---|
| Positive | 34 |
| Negative | 5261 |

Dataset imbalanced problem commonly occurs in real-world data, whilst most machine learning algorithms are only tuned to perform well on a balanced dataset (Thabtah et al., 2020). Hence, a novel technique needs to be employed to balance such datasets before model development to attain a robust model.

The data were sampled to reduce the minority and majority classes by multiplying 10 times the minority class.

The application of the data balancing technique resulted in a total of 680 data tuples. The pre-processed dataset consisted of the individual medical history of a mother with the health status of their child.

### Feature analysis and selection

Feature analysis is a process that identifies feature relevance (Aggarwal et al., 2014). Identification of feature relevance led to the optimum selection of features used for model development. Python software provided tools for feature analysis. Figure 1 shows the feature's importance score. The procedure provided by the Random Forest helps medical practitioners to understand which feature has high impact in predicting the status of the child than the other which leads to the transmission of HIV from mother to child whereby this prediction can reduce HIV transmission and death rate.

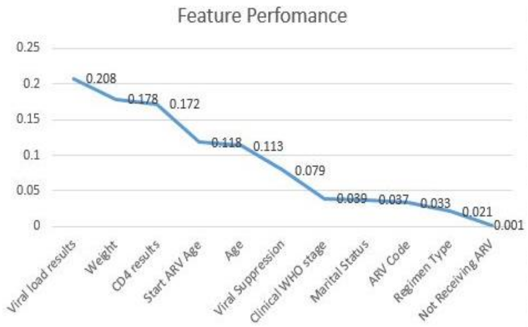

**Figure 1: Feature importance score**

### Model development

Figure 2 presents the core model development process. The dataset was split into training and test datasets by a 75:25 ratio and in turn used to build models under commonly used algorithms in the health domain. According to (Rácz et al., 2021) when the partition ratio is between 70% and 80%, the model prediction performance increases with the increase of the dataset size. Thus, a default split ratio of 75:25 which is between 70% and 80% was applied. A total of 510 tuples were used for model training and 170 tuples for testing the model.

### Random forest model

The Random Forest classifier with ensemble technique was used with the parameters *n_estimators (2,8,10)* to set the number of trees before making the decision, *max_depth (50,150,250)* to set the longest path between the leaf node and the root node, *min_samples_split (2,3,4) is* the minimum number set in a node before the split and *min_samples_leaf* (*1*,*2*,*3*) is the minimum number set in a leaf before the split. Table 2 shows the Random Forest model confusion Matrix and Table 3 shows the corresponding performance.

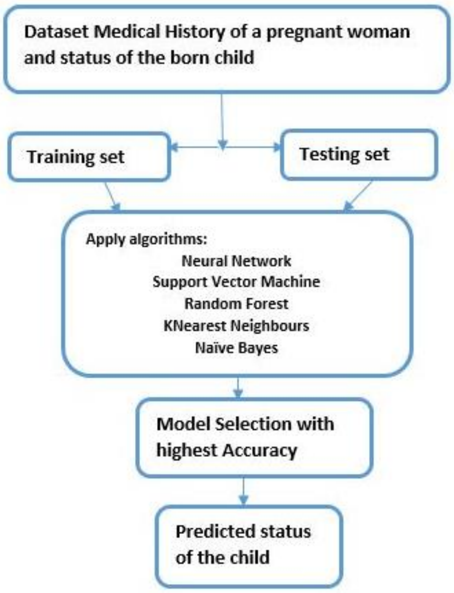
Dataset Medical History of a pregnant woman and status of the born child
Training set
Testing set
Apply algorithms:
Neural Network
Support Vector Machine
Random Forest
KNearest Neighbours
Naïve Bayes
Model Selection with highest Accuracy
Predicted status of the child

**Figure 2: Flow Chart for the Model Development**

**Table 2: Random-forest model confusion matrix**

| | | True Label | |
|---|---|---|---|
| | | True Positive | False Positive |
| Predicted label | True Positive | 90 | 0 |
| | False Negative | 1 | 79 |

**Table 3: Metric performances for random forest model**

| Accuracy | Error (1-Accuracy) | Recall | Precision | F1-Score |
|---|---|---|---|---|
| 0.99 | 0.01 | 0.99 | 1 | 0.99 |

### Models for other common ML methods in health domain

Through the extensive literature review conducted, commonly used Machine Learning methods were selected for model development on the same dataset to compare the performance of the machine learning methods, as depicted in Table 4.

The Neural Network classifier model had 0.95% performance whereby performed well with the hidden layer sizes of (*16*,*16*), solver *lbfgs* and the activation *relu* and the maximum iteration of 400.

The Support Vector Machine model had 82% accuracy, using *gamma parameter (0.01, 0.03)* to influence a single training example reaches whereby a low number means far and a high number means close. It used kernel *rbf* and sigmoid function to set the mathematical functions used to manipulate data and the C parameter tells the SVM optimization of how much to avoid misclassifying each training example, the larger the C the smaller margin hyperplane will be chosen.

The Naïve Bayes had 65% accuracy and 0.35 errors. The GaussianNB classifier was used in the prediction model development.

The KNearest had 97% accuracy with 0.03 errors. The KNearest classifier was used for model development.

Table 4 presents a summary of different model's performances other than accuracy, with respective attained accuracy; Random Forest (0.99), Neural network (0.95), Support vector machine (0.82), Naïve Bayes (0.65) and KNearest (0.97) on the same dataset.

**Table 4: Different models performances**

| SN | Model Name | Precision | Error | Recall | F1-Score |
|---|---|---|---|---|---|
| 1 | Random Forest | **1** | **0.01** | **0.99** | **0.99** |
| 2 | Neural Network | 1 | 0.05 | 0.92 | 0.96 |
| 3 | KNearest | 1 | 0.03 | 0.95 | 0.97 |
| 4 | Support Vector Machine | 0.93 | 0.18 | 0.77 | 0.65 |
| 5 | Naïve Bayes | 0.36 | 0.35 | 0.97 | 0.53 |

## Model validation

The key to a fair comparison of machine learning algorithms is ensuring that each algorithm is evaluated in the same way on the same data. This can be achieved by forcing each algorithm to be evaluated on a consistent test harness. A 10-fold cross-validation procedure was used to evaluate each algorithm, importantly configured with the same random seed to ensure that the same splits to the training data are performed and that each algorithm is evaluated in precisely the same way. To achieve the comparison, common seed 7 and split 10 were used on the same datasets. The cross-validation showed that Random Forest has a better chance to distinguish between classes. Furthermore, its higher accuracy implied a better performance for the model's ability to distinguish the status of a child whether positive or negative, compared to other algorithms as depicted in Figure 3.

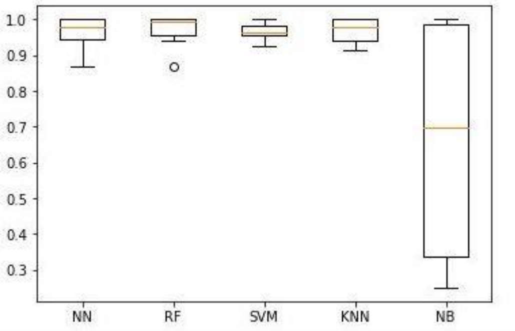

**Figure 3: Models cross-validation**

In the supervised learning problem, the confusion matrix was used to summarize the performance, since the accuracy alone may not satisfy the performance of the model. The confusion matrix shows exactly how the model is getting right, and wrong and what types of errors are made. The test dataset used to validate the model shows that the total number of the child is 80 tuples with actual negative status and 90 tuples with the actual positive status of the child making a total of 170 test tuples. Table 2 shows results for the Random Forest model where the correct predicted values are shown in the diagonal line, which is 79 and 90. This implies that 1 error was made in predicting the negative status of the child as positive while 0 errors in predicting the positive status as negative. All models were validated using the Area under the curve (AUC) (in which the higher the AUC the better performance of a model) as depicted in Figure 4. The results show that the Random Forest model achieved a higher AUC compared to other models. This implies that the Random Forest model has a higher ability to distinguish whether the child is positive or negative compared to the other models.

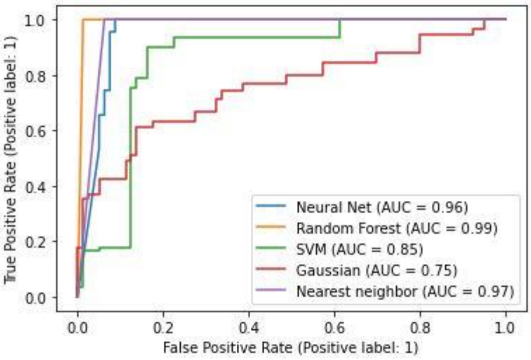

**Figure 4: Models receiver operating characteristics (AUC) curve**

## DISCUSSION

The goal of this work was to improve the prediction of HIV transmission from mother to child using machine learning models. The Random Forest model showed higher performance compared to existing best-performing models that are commonly used in the health domain. The random forest has an in-built mechanism for feature selection ahead of model development, which helps to eliminate the least important features and focus on the most relevant ones, which influence the model's prediction rate (Aggarwal et al., 2014; Zhang, 2005). Random forest models also have high predictive accuracy and the ability to be used in the complex dataset (Hjerpe, 2016). The random forest method performs best in solving classification problems when the dataset is balanced (Singh & Purohit, 2015), whereby the dependent variable must be linear to the independent variable for good performance (McAlexander & Mentch, 2020). Moreover, the results correlate with existing literature which confirms that the random forest method has high accuracy with good tolerance for inconsistency and noise datasets (Gao et al., 2019).

Correcting the data imbalance problem helped to boost the model's performance, indeed, this is prevalent in the real-world dataset (Gao, 2020). As (Sánchez & Valdovinos, 2020; Singh & Purohit, 2015), the sampling procedure was a novel approach used for handling the data imbalance.

## CONCLUSION AND RECOMMENDATIONS

This research aimed at improving the prediction of HIV transmission from mother to child using the individual medical history of a mother and the status of the child. To achieve that, the study determined relevant features that influence the transmission of HIV from mother to child, the use of balancing of the dataset and the use of metrics measurement. The random forest method along with other commonly applied supervised learning methods in the health domain was used to determine the best prediction models.

The study contributes to knowledge on the importance of individual datasets and data balancing for model performance, in addition to the potential to improve decision-making and planning for HIV transmission from mothers and appropriate medication.

## ACKNOWLEDGMENT

The authors are grateful to UDSM for providing a supportive research environment, the Tanzania National AIDS Control Program for granting permission to use the CTC2 data and the Muhimbili University of Health and Allied Sciences under the project of the East Africa Centre of Excellence and Cardiovascular Sciences for sponsoring the study.

## DECLARATION

The authors declare that there is no conflict of interest regarding this research.

## REFERENCES

Aggarwal, C. C., Kong, X., Gu, Q., Han, J., and Yu, P. S. (2014). Active learning: A survey. *Data Classification: Algorithms and Applications*, 571–605. https://doi.org/10.1201/b17320

Al-Zaiti, S., Besomi, L., Bouzid, Z., Faramand, Z., Frisch, S., Martin-Gill, C., Gregg, R., Saba, S., Callaway, C., and Sejdić, E. (2020). Machine learning-based prediction of acute coronary syndrome using only the pre-hospital 12-lead electrocardiogram. *Nature Communications*, **11**(1): 1–10. https://doi.org/10.1038/s41467-020-17804-2

Campos Coelho, A. V., Campos Coelho, H. F., Arraes, L. C., and Crovella, S. (2019). HIV-1 mother-to-child transmission in Brazil (1994–2016): a time series modeling. Brazilian Journal of Infectious Diseases, **23**(4): 218–223. https://doi.org/10.1016/j.bjid.2019.06.012

Ekpenyong, M. E., Etebong, P. I., and Jackson, T. C. (2019). Fuzzy-multidimensional deep learning for efficient prediction of patient response to antiretroviral therapy. Heliyon, **5**(7): e02080. https://doi.org/10.1016/j.heliyon.2019.e02080

Ellington, S. R., King, C. C., and Kourtis, A. P. (2018). Predictors of HIV-1 serostatus disclosure: a prospective study among HIV-infected pregnant women in Dar es Salaam, Tanzania. *HHS Public Access*, **6**(2): 1451–1469. https://doi.org/10.2217/fvl.11.119.Host

Gao, J. I. E. (2020). Data Augmentation in Solving Data Imbalance Problems. *Degree Project Computer Science and Engineering*.

Gao, X., Wen, J., and Zhang, C. (2019). An Improved Random Forest Algorithm for Predicting Employee Turnover. *Mathematical Problems in Engineering, 2019*. https://doi.org/10.1155/2019/4140707

Girum, T., Wasie, A., and Worku, A. (2018). Trend of HIV/AIDS for the last 26 years and predicting achievement of the 90-90-90 HIV prevention targets by 2020 in Ethiopia: A time series analysis. BMC Infectious Diseases, **18**(1): 1–10. https://doi.org/10.1186/s12879-018-3214-6

Grafberger, S., Munich, T. U., Stoyanovich, J., and Schelter, S. (2021). *Lightweight Inspection of Data Preprocessing in Native Machine Learning Pipelines*. https://github.com/tensorflow/transform

Hjerpe, A. (2016). Degree Project in the Field of Technology *Computing Random Forests Variable Importance Measures ( VIM ) on Mixed Continuous and Categorical Data Computing Random Forests Variable Importance Measures (VIM) on Mixed Numerical and Categorical Data Beräknin. Vim*.

James, T. O., Gulumbe, S. U., and Danbaba, A. (2018). Resilient Back-Propagation Algorithm in the Prediction of Mother to Child Transmission of HIV. *OALib*, **5**(5): 1–7. https://doi.org/10.4236/oalib.1104538

Nan, Y., and Gao, Y. (2018). A machine learning method to monitor China's AIDS epidemics with data from Baidu trends. PLoS ONE, **13**(7): 1–12. https://doi.org/10.1371/journal.pone.0199697

McAlexander, R. J., and Mentch, L. (2020). Predictive inference with random forests: A new perspective on classical analyses. *Research and Politics*, **7**(1): https://doi.org/10.1177/2053168020905487

Negussie Deyessa, A. B. (2015). Determinants of Mother to Child HIV Transmission (HIV MTCT); A Case Control Study in Assela, Adama and Bishoftu Hospitals, Oromia Regional State,Ethiopia. *Cell & Developmental Biology*, **4**(2): 1–12. https://doi.org/10.4172/2168-9296.1000152

Rácz, A., Bajusz, D., and Héberger, K. (2021). Effect of dataset size and train/test split ratios in qsar/qspr multiclass classification. *Molecules*, **26**(4): 1–16. https://doi.org/10.3390/molecules26041111

Sánchez, J. S., and Valdovinos, R. M. (2020). Jou. *Expert Systems With Applications*, 114301. https://doi.org/10.1016/j.eswa.2020.114301

Shen, C., Yu, X., Harrison, R. W., and Weber, I. T. (2016). Automated prediction of HIV drug resistance from genotype data. BMC Bioinformatics, **17**(8): https://doi.org/10.1186/s12859-016-1114-6

Singh, A., and Purohit, A. (2015). A Survey on Methods for Solving Data Imbalance Problem for Classification. *International*

*Journal of Computer Applications*, **127**(15): 37–41. https://doi.org/10.5120/ijca2015906677

Thabtah, F., Hammoud, S., Kamalov, F., and Gonsalves, A. (2020). Data imbalance in classification: Experimental evaluation. *Information Sciences*, **513**: 429–441. https://doi.org/10.1016/j.ins.2019.11.004

UNAIDS. (2019). Global HIV and AIDS statistics 2019 Fact sheet. *Global HIV and AIDs Ststistics, World AIDS Day 2019 Fact Sheet*, **1**: 1–6.

WHO&CDC. (2008). *PREVENTION OF MOTHER-TO-CHILD TRANSMISSION OF HIV Generic Training Package D R A F T Participant Manual*. *January*, 369.

WHO. (2003). HIV / AIDS: Confronting a Killer. *World Health Organization*, 41–56. https://doi.org/10.1146/annurev.ecolsys.35.021103.105711

Xu, J., Zhang, Y., and Miao, D. (2020). Three-way confusion matrix for classification: A measure driven view. *Information Sciences*, **507**: 772–794. https://doi.org/10.1016/j.ins.2019.06.064

Zhang, H. (2005). Exploring conditions for the optimality of naïve bayes. *International Journal of Pattern Recognition and Artificial Intelligence*, **19**(2): 183–198. https://doi.org/10.1142/S0218001405003983