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

Rebecca B. Chaula¹ and Godfrey N. Justo²

¹Directorate of Information and Communication Technology, Muhimbili University of Health and Allied Sciences, P.O. BOX 65001, Dar es Salaam, Tanzania.
²Department 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.

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

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 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.

### 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>
<thead>
<tr>
<th>Table 1: Imbalanced dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class</td>
</tr>
<tr>
<td>Positive</td>
</tr>
<tr>
<td>Negative</td>
</tr>
</tbody>
</table>

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.
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.

Table 2: Random-forest model confusion matrix

<table>
<thead>
<tr>
<th>Predicted label</th>
<th>True Positive</th>
<th>False Positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Positive</td>
<td>90</td>
<td>0</td>
</tr>
<tr>
<td>False Negative</td>
<td>1</td>
<td>79</td>
</tr>
</tbody>
</table>

Table 3: Metric performances for random forest model

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>Error (1-Accuracy)</th>
<th>Recall</th>
<th>Precision</th>
<th>F1-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.99</td>
<td>0.01</td>
<td>0.99</td>
<td>1</td>
<td>0.99</td>
</tr>
</tbody>
</table>

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>
<thead>
<tr>
<th>SN</th>
<th>Model Name</th>
<th>Precision</th>
<th>Error</th>
<th>Recall</th>
<th>F1-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Random Forest</td>
<td>1</td>
<td>0.01</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>2</td>
<td>Neural Network</td>
<td>1</td>
<td>0.05</td>
<td>0.92</td>
<td>0.96</td>
</tr>
<tr>
<td>3</td>
<td>KNearest</td>
<td>1</td>
<td>0.03</td>
<td>0.95</td>
<td>0.97</td>
</tr>
<tr>
<td>4</td>
<td>Support Vector Machine</td>
<td>0.93</td>
<td>0.18</td>
<td>0.77</td>
<td>0.65</td>
</tr>
<tr>
<td>5</td>
<td>Naïve Bayes</td>
<td>0.36</td>
<td>0.35</td>
<td>0.97</td>
<td>0.53</td>
</tr>
</tbody>
</table>

**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.

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.

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


