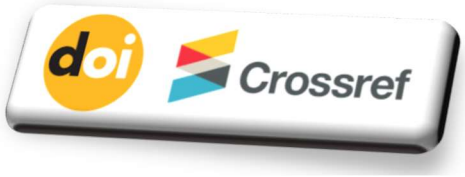
Journal, Advances in Mathematical & Computational Sciences
Vol. 12 No. 4, 2024 Series
www.isteams.net/mathematics-computationaljournal



Journal of Advances in Mathematical & Computational Sciences
An International Pan-African Multidisciplinary Journal of the SMART Research Group
International Centre for IT & Development (ICITD) USA

Available online at https://www.isteams.net/ mathematics-computationaljournal.info
CrossREF Member Listing - https://www.crossref.org/06members/50go-live.html

Ensemble Machine Learning-based Predictive Modelling of Interruption-in-Treatment of People Living with HIV/AIDS

[1]Joseph P. Chikezie, *[1]Patience U. Usip, [2]Uduak A. Umoh, [3]Edward N. Udo & [2]Ifiok J. Udo
Department of Computer Science, University of Uyo, Uyo, Nigeria
Department of Information Systems, University of Uyo, Uyo, Nigeria
Department of Software Engineering, University of Uyo, Uyo, Nigeria
Corresponding Author: patienceusip@uniuyo.edu.ng

ABSTRACT

Interruption in treatment among People Living with HIV/AIDS (PLWHA) poses significant challenges to the efficacy of antiretroviral therapy (ART) programs and the global fight against HIV/AIDS. Predicting these interruptions enables timely interventions, enhancing treatment adherence and health outcomes. This research proposes an ensemble machine learning model to predict treatment interruptions in PLWHA. The model integrates multiple machine learning algorithms, including decision trees, AdaBoost, Random Forest to leverage their combined predictive power. Utilizing a dataset encompassing demographic, clinical factors, the model demonstrates superior accuracy, precision. Among the algorithms tested, AdaBoost demonstrated the highest accuracy, outperforming Decision Tree and Random Forest models. This indicates that AdaBoost's ability to handle complex patterns in patient data makes it a superior tool for predicting HIV treatment adherence issues. The findings underscore the potential of ensemble models in healthcare decision-making, offering a robust tool for clinicians and policymakers to identify at-risk individuals and develop targeted intervention strategies to enhance retention in care. This study contributes to the growing body of research leveraging machine learning for improving public health outcomes for chronic diseases.

Keywords: Antiretroviral therapy, Machine Learning Algorithm, Interruption in Treatment, Ontology, Knowledge representation, Predictive Model

Joseph P. Chikezie, Patience U. Usip, Uduak A. Umoh, Edward N. Udo & Ifiok J. Udo (2024): Ensemble Machine Learning-based Predictive Modelling of Interruption-in-Treatment of People Living with HIV/AIDS. Journal of Advances in Mathematical & Computational Science. Vol. 12, No. 3. Pp 57-72. Available online at www.isteams.net/mathematics-computationaljournal.
dx.doi.org/10.22624/AIMS/MATHS/V12N4P5

1. INTRODUCTION

Human Immunodeficiency Virus (HIV) continues to be a global public health challenge, with millions of people living with HIV/AIDS (PLHIV) requiring lifelong treatment and monitoring.

Antiretroviral therapy (ART) has been instrumental in improving the quality of life and life expectancy of PLHIV. However, treatment adherence remains a critical challenge, as interruptions in ART can lead to drug resistance, disease progression, and increased risk of transmission. Managing HIV/AIDS, therefore, requires not only medical interventions but also psychosocial support systems to ensure patients remain engaged in their treatment. Efforts have been made by different medical personnel to ensure improvement in the treatment and prevention of this virus. Scaling up care for thousands of patients require a care management system with robust counseling support information to ensure that each patient and his or her health status can be identified and monitored effectively.

To improve quality of healthcare of people with chronic conditions, Amith et al. (2019) utilized an ontology-based method to model the dialogue for the counseling of HPV vaccine. Amith et al. (2020) added that ontology method was efficient in handling dialogue for PrEP (Pre-exposure prophylaxis) and PEP (Post-exposure prophylaxis) counseling for PLWH and HIV-negative people with high risks of HIV infection, the system handles questions that are queried from a knowledge base, called the Ontology of PrEp and PEP (OPP) used in the management of reactive and non-reactive patients. More so, Innovative approaches of identifying individuals at risk of LTFU and tailored activities to prevent them from disengagement to care are needed to ensure optimal client health and sustained epidemic control (UNAIDS, 2022).

The utilization of machine learning algorithms has been seen as a sure way to cause machines to learn some of the procedures by experts from existing datasets in several domains such as chatbot for rural farmers (Usip et al., 2022), text classification and machine translation (Oboho-Etuk et al., 2023), Cyber Physical Systems (2022), etc. This research proposes the development of an ensemble model designed to predict potential treatment interruptions and assist care giver to take proactive measures to enhance retention in care. The model utilizes machine learning algorithms to analyze clinical data, generating predictions about the likelihood of treatment interruption.

2. RELATED WORKS

2.1 Interruption in Treatment Prediction

Over the years, researchers have adopted different machine learning technique to identify patient group at risk of disengagement and how to target available interventions to these groups while still in engaged in care. Denney *et al*., (2019) proposed a solution to predict the probability of a patient not showing up for a medical appointment and demonstrate the effectiveness of a machine learning model. The research team demonstrated a process to intake clinical, demographic, financial, and other types of data and successfully used a modern machine learning algorithm to predict the probability of no-show appointments. They used a survival of the fittest approach to nine different algorithms, including a classical statistical model and newer machine learning techniques such as Adaptive Boosting (AdaBoost), Logistic Regression, Naïve Bayes, Support Vector Machine, Stochastic Gradient Descent, Decision Tree Classifier, Extra Trees Classifier (Extremely Randomized Trees), Random Forest Classifier, and eXtreme Gradient Boosting (XGBoost), yielded a clear winner, AdaBoost. AdaBoost outperformed all the other algorithms in a statistically meaningful manner. A direct comparison of AdaBoost and Logistic Regression shows AdaBoost to be more effective at predictions on this data set.

In other three studies, the authors used the same public dataset to build their predictive model. Public Dataset refers to medical appointment scheduled in public hospitals in the city of Victoria, in the state of Expirito Santo, Brazil. Praveena *et al*., (2019) started from data cleaning and preprocessing, explanatory analysis and finally had the highest accuracy on the test set with the Decision Tree algorithm. Similarly, in Salazar *et al*., (2022) the authors explored the dataset following the same machine learning steps and obtained the same accuracy for both Decision Tree and Random Forest algorithm. Lastly, Batool *et al*. (2022) in their study built a solution based on the Decision Tree algorithm. However, a scheduling system was implemented such that the overall model detects whether a patient has a risk of missing appointment with a 95% accuracy, upon which it automatically enables the risky patient's schedule slot for overbooking and notifies medical staff or administration to contact them accordingly.

Ogbechie *et al*. (2023) described their experience in developing a machine learning (ML) model to predict interruption in treatment (IIT) at 30 days among people living with HIV newly enrolled on ART in Nigeria and the integration of the model into the routine information system. Routine program data collected from January 2005 through February 2021 and was used to train and test an ML model (boosting tree and Extreme Gradient Boosting) to predict future IIT. Maskew et al. (2022) in their study evaluated three classification algorithms (Logistical regression, random forest and AdaBoost) for building predictive models. These ML algorithms were used to anonymised, patient level HIV Programmatic data from two districts in South Africa, 2016-2018. In another study in Tanzania, Fahey *et al*., (2022) used an ensemble decision tree approach to predict risk of disengagement from HIV care.

This approach used routing Electronic Medical Records (EMR) from the time of ART initiation through 24 months of follow up for 241 patients. They added that aggregated decision tree learning approach is appropriate for this analysis that involves a smaller sample size where prediction accuracy can be substantially influenced by missing data (common in EMR data). Similarly, Luiz H. et al., (2022) explored the main causes that contribute to a patient's no show and develop a prediction model able to identify whether the patient will attend their scheduled appointment or not. The study was based on data from clinics that serve the Unified Health System (SUS) at the University of Vale do Itajaí in southern Brazil. Thus, they selected the Random Forest classifier as the best classification algorithm able to achieve the objective of the analysis.

Lee *et al*., (2017) presented a feature engineering approach to predict a patient's risk of clinic no-show. The authors applied text mining techniques in order to extract useful information from the records collected. They build a no-show XGBoost model with the 15 top features and achieve an AUC of 0.793. Moreover, the authors describe the insights gained, and discuss modeling considerations on the trade-offs between model accuracy and complexity of deployment. Although the model developed in this work had good ability to identify clinic no-shows, it was based on a premise enterprise data warehouse that could be hardly adapted to other environments. Other studies used statistical predictive models capable of predicting whether a patient will be absent from appointment based on their historical data, Alaeddini *et al*. (2011) they developed a hybrid probabilistic model based on logistic regression and empirical Bayesian inference to predict the probability of no-shows in real time using both general patient social and demographic information and individual clinical appointments attendance records

Knowledge Representation and Ontologies

Knowledge representation is a sub-field of Artificial Intelligence where domain knowledge are properly represented in a bid to making machines make use of them and think the way humans do. Grosan et al., (2011) agreed that representing knowledge is a key issue in artificial intelligence. The question of how human knowledge of all kinds can be represented by a computer language, and in such a way that computers can use this knowledge for purposes of reasoning have led to generalized use of knowledge representations in various contexts, including search, simulations, web semantic ontology description. Knowledge representation is of immense importance in the field of Artificial Intelligence. An intelligent agent should be able to acquire information (or knowledge) from environment, to represent and understand it, and to be able of reasoning, that is to infer the implications of what it knows and of the choices it has.

Taxonomies, rules, semantic network, logic programs, categories, frames have been a standing knowledge representation formalism but recently we have seen an explosion of interest in ontologies as artifacts to represent human knowledge and as a critical component in knowledge management. Ontologies have evolved in computer science as computational artefacts to provide computer systems with a conceptual yet computational model of a particular domain of interest. Knowledge-based systems have a computational model of some domain of interest in which symbols serve as surrogates for real world domain artefacts, such as physical objects, events, relationships, etc. The domain of interest can cover any part of the real world or any hypothetical system about which one desires to represent knowledge for computational purposes. A knowledge-based system maintains a knowledge base which stores the symbols of the computational model in form of statements about the domain, and it performs reasoning by manipulating these symbols.

Ontologies have been experimented in several research domains including e-health. The definition and use of ontology in the medical domain, is an active research field, as it has been recognized that ontology-based systems can be used to improve the management of health systems (Valls et al., 2010). Recently, most works that require intelligence and adopted ontology is seen in Abdollahi (2021) where ontology-based approach was proposed for augmenting medical documents by considering concepts of terms and phrases in the discharge notes. This approach substitutes the vocabulary and expressions with their scientific names if they relate to a concept in the clinical area. Furthermore, an incremental ontology-based and dictionary-based method is proposed at the same time to argument new medical documents and Biomedical Natural Language Processing (NLP) played a role in extracting information in medical discharge notes.

Hong (2021) proposed a reasoning-reuse method to construct Brain Areas-Autism (BAA) Ontology to support the domain knowledge discovery i.e., discovering inherent relationship between autism and brain areas based on BAA ontology. NLP technique was also used to extract and fuse knowledge from scientific literatures; The Reasoning-reuse method was used to construct the brain areas-autism ontology (BAA Ontology) to ensure the accuracy of the ontology and then uses rule-based reasoning to expand the scale of ontology. Paganelliet et al., (2010) in their study used ontologies to improve home care for patients with chronic diseases that require constant care from different caregivers. They argued that there is a need for better long-term care because most chronic diseases develop slowly overtime. The diseases have a great impact on the patient's personal life and relationships. They built a system around a care network.

A care network is all the people involved with patient care such as nurses, doctors, and relatives. The "person" entity takes care of all the people in the care network, and the "Place" entity models the location like home, hospital, and test center. The ontology also stores biomedical properties of the patient, like temperature, and social context of the patient like information regarding members of the care network. The system also had the ability of alerting the patient's care network when some medical parameters get out of range.

Kataria et al., (2010) in their study used an ontological layering in classifying requests imposed by healthcare professional when retrieving e-health information from heterogeneous sources and reducing semantic heterogeneities across repositories. The layered model was based on Generic Ontology for Context-aware, Interoperable and Data sharing (GO-CID) software applications, applicable to e-health environments. The problems of sharing, retrieval of e-health data is solved based on these ontologies and their layering. In the study by Phan et al., (2017), a domain ontology was integrated into a neuro-symbolic architecture with a restricted Boltzmann machine model to predict and explain human behaviors for health care intervention systems in health social networks. Similarly, Che et al., (2015) proposed a health diagnosis prediction system that uses medical ontologies to learn (embedded) representations for medical nodes in the knowledge and their parent codes.

These are then used to learn the input representations of patient data, which are then loaded into a neural network architecture. The system uses an attention mechanism that learns weights to improve the prediction accuracy and allow the interpretation of the importance of various pieces of information Massari et al., (2022) focused on building an ontology-based intelligent system for preventing the diagnosis of diabetes patients based on the result of a decision tree model. The research created an artificially intelligent predictive model which could identify patients with diabetes and was mapped into ontology using SWRL rules which were extracted from the decision tree model with the help of a java program.

More so, Chang et al. (2014) developed an ontology for depression diagnosis. They used the Description Logics variant of the Web Ontology Language (OWL) in developing the ontology, and GeNIe & SMILE were used as reasoning engines for making inferences on the ontology. Patient personal ontology was fetched from the cloud. The user interface was designed to get data that will be used to infer whether the patient is suffering from depression. Bayesian Network was used as a tool for computing the probability of a user having a disease from the users' description. Chen et al., (2019) gathered comprehensive diabetic knowledge from individuals in order to develop an ontology-based model for diabetic patients that will aid in future diagnosis and therapy. When creating an ontology-based model for diabetic patients, the model makes use of semantic web ontology language, which helps medical professionals diagnose diabetes patients accurately. The results generated suggest that the proposed solution is more accurate in managing diabetes compared to other medical applications. Despite this, it cannot be applied on the Internet of Medical Things (IoMT)-based applications.

Athanasios et al., (2012) in their study proposed an Integrated Ontology-Based Model for the Early Diagnosis of Parkinson's Disease (PD). They proposed the general architecture of an integrated multi-scale ontology-based modelling technology for early diagnosis of PD and progress monitoring. The model was used to identify physics and biology-based biomarkers of processes occurring at radically different scales, from cell to whole body. It is however a complex system.

The proposed model would help physicians to decide more easily and faster what treatment to give to patients with PD, providing different treatment options in individual patients. An early diagnosis may enable physicians to provide medical care at an earlier stage, at a time when clinical diagnosis using only signs and symptoms of disease is challenging. Early diagnosis combined with future drugs and prevention strategies will also delay or stop the onset or the progress of PD. This can be shifted directly to reduced costs of PD prevention and treatment.

3. MACHINE LEARNING-BASED MODEL

The proposed model for the prediction of interruption-in-treatment for people living with HIV/AIDS features the adoption of an ensemble model. The adoption of the Adaboost, an ensembled ML model is due to its ability to create multiple models and choosing the best from the various choices. Figure 1 shows the architecture of the proposed model.

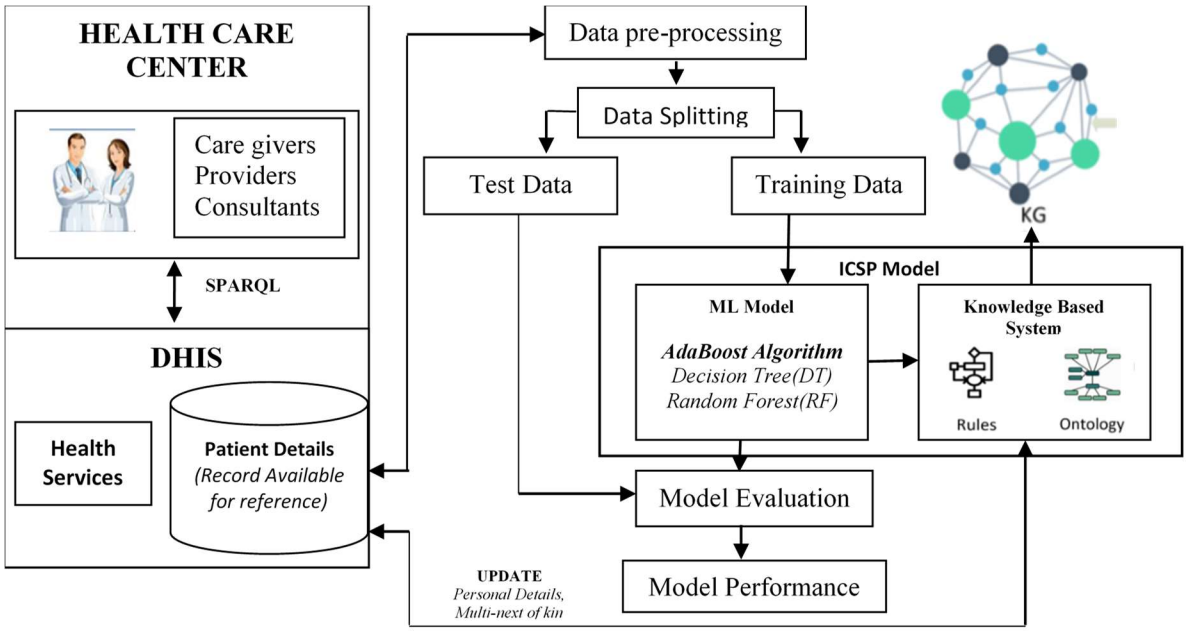


Figure 1: Proposed Architecture

3.1 Data Collection

The predictor variable selected that will be used for model building are preselected as they were strongly associated with treatment interruption by a group DHIS project staff and HIV prevention and treatment experts. The variables selected for this model include age, gender, marital status, local government area, baseline clinic stage, care entry point. A patient is deemed LTFU if a patient did not attend within 28 days of their scheduled appointment i.e. they were more than 28 days late. If they missed their appointment, they were considered LTFU and not retained in care.

3.2 Model Training

1. AdaBoost Algorithm

a. The input to the ICS Model is the training data which represent a patient level data for reactive patient which includes schedule appointments and treatment progress for patient. The algorithm first assigns weight (initial weight) to all the data points. Initially, all the weights will be equal and sum to 1. The initial weight is given in Equation 1

$$Initial\ Weight = w(x,y)i = \frac{1}{N}, i = 1,2, \dots n \qquad \text{Equation 1}$$

Where N is the total number of data points.

b. After assigning weight, it classifies the samples and create a decision stump for each of the features and then calculate the Gini index of each tree. The tree with lowest Gini index will be the first stump. For all stumps created, "Amount of Say" or "Importance", "influence" for the classifier in classifying the data points is calculated using Equation 2

$$Amount\ of\ Say = \frac{1}{2}\log\frac{1-Total\ Error}{Total\ Error} \qquad \text{Equation 2}$$

Where $Total\ Error$ is the summation of all sample weights of misclassified data points.

c. After finding the Amount of say and total error, the weight for the samples is updated because if identical weight is maintained for the subsequent model, the output will mirror what was obtained in the initial model. The wrong predictions will be given more weight, whereas the correct predictions weight will be decreased, and more preference will be given to the points with higher weight. The new sample weight is given in Equation 3

$$New\ Sample\ Weight = old\ weight * e^{\pm Amount\ of\ say\ (\alpha)} \qquad \text{Equation 3}$$

i. The amount of, say (alpha) will be *negative* when the sample is correctly classified.
ii. The amount of, say (alpha) will be *positive* when the sample is miss-classified.

d. After adjusting the weight, the new weights are normalized to sum to 1, the algorithm divides data point in buckets based on the new sample weight and then select random numbers from 0-1. Since incorrectly classified records have higher sample weights, the probability of selecting those records is very high. A new dataset of wrongly classified data points is selected because it has a higher weight.

e. Given the new dataset, it repeats the above steps (a,b,c,d) Iterates through these steps until and unless a low training error is achieved.

Suppose with respect to a dataset, we construct 3 decision stumps (DS1. DS2, DS3) in a sequential manner. If we send test data, it will pass through all the decision stumps and finally arrive at a class that has the majority, and based on that the prediction for the test dataset is made

2. Decision Tree Algorithm

Decision Tree algorithm splits the dataset into subsets based on the most significant feature at each node. The goal is to create branches that best separate the data into homogenous subsets (in terms of the target variable). This process continues recursively, with each node being split until a stopping criterion is met (such as a maximum depth, a minimum number of samples, or perfect classification). For classification, the algorithm typically uses impurity measures like Gini Index, Entropy (Information Gain), or Chi-square to decide the best split at each node. Based on the selected feature, the data is divided into subsets that are as homogenous as possible with respect to the target variable. At each node of the tree, the algorithm selects the best feature to split the data. The goal is to find a feature that best separates the data into different classes (for classification). The most common measures for splitting are

a. **Gini Impurity: Gini impurity is a measure o**f how often a randomly chosen element would be incorrectly classified. The Gini Index is given in Equation 4

$$\text{Gini (t)} = 1 - \sum_{i=1}^{k} p_i^2 \quad \text{Equation 4}$$

Where p_i is the probability of choosing an element from class 'i' in note t and k is the number of classes.

b. **Entropy & Information Gain:** Entropy measures the disorder or impurity in a dataset. Information gain is the reduction in entropy after a split. Entropy is given in Equation 5

$$Entropy\ (t) = -\sum_{i=1}^{k} p_i \log_2 p_i \quad \text{Equation 5}$$

Where p_i is the proportion of samples in class i at node t. the goal is to maximize the information gain.

$$IG(t) = Entropy(parent) - \sum \frac{|S_j|}{|S|}.Entropy\ (S_j) \quad \text{Equation 6}$$

Where S_j are the subsets created by splitting on feature j

Based on the selected feature, the data is divided into subsets that are as homogenous as possible with respect to the target variable. The splitting process continues recursively for each subset until one of the stopping conditions is met, such as: A predefined maximum depth is reached, A node contains fewer than a certain number of samples, All data points in the node belong to the same class. Once the tree reaches its leaf nodes, it assigns an output to each node. the final prediction for a new instance is made by traversing the tree from the root to a leaf. The class label at the leaf node is assigned as the predicted class.

3. Random Forest Algorithm

Random Forest generate multiple bootstrap samples from the training dataset. A bootstrap sample is created by randomly selecting data points with replacement from the original dataset. This means some data points may be repeated, while others may be omitted. When the original dataset have N data points, the bootstrap sample will also have N data points, but selected randomly with replacement.

$$S_{bootstrap} = \{x_1, x_2, \dots, x_N\}(where\ some\ of\ the\ x_1\ values\ may\ be\ repeated) \quad \text{Equation 7}$$

Once the bootstrap samples are created, **multiple decision trees** are trained on these samples. Each tree is trained by recursively splitting the data based on feature values, which maximizes **information gain** or minimizes **impurity** at each split. The decision trees are built by recursively splitting the dataset into subsets based on the chosen feature at each node. The trees are typically grown **without pruning,** meaning they are fully grown until a stopping criterion (such as a maximum depth or minimum number of samples per leaf) is reached.

Once the trees are built, **prediction** is done by aggregating the outputs from all individual trees in the forest. For classification, **majority voting** is used determine the final class label. If there are N_{trees} trees in the forest, and each tree makes a prediction $\hat{y}_i$ for a given input x, the final predicted class $\hat{y}$ is the class that is predicted by the majority of the trees. This is given in Equation 8

$$\hat{y} = \arg max \sum_{i=1}^{N_{trees}} 1\ (\widehat{\hat{y}_i} = c) \qquad \text{Equation 8}$$

Where $\hat{y}_i$ is the prediction from tree i, c is a class, $1(\hat{y}_i = c)$ is an indicator function that is 1 if the tree predicted class c and 0 otherwise.

Since each tree is trained on a bootstrap sample, some data points are not included in the training of each tree. These data points are known as **out-of-bag** (OOB) data. OOB data can be used to estimate the performance of the Random Forest without needing a separate validation set. The OOB error rate is given in Equation 9

$$OOB\ Error\ Rate = \frac{1}{N_{OOB}} \sum_{i=1}^{N_{OOB}} 1\ (y_i \neq \hat{y}_i) \qquad \text{Equation 9}$$

Where y_i is the true label of the i-th data point, $\hat{y}_i$ is the OOB prediction for that point and N_{OOB} is the total number of OOB predictions

4. Model Development with WEKA

The dataset was uploaded to WEKA and pre-processed using WEKA filters after which a predictive Adaboost algorithm. For model selection was comparatively confirmed when the same data is trained on Decision tree (DT) model using J4.8 classifier algorithm and Random Forest (RF) classifier. J4.8 evaluates the data using its gain ratio attribute evaluator was created to make predictions on the dataset given. All settings were kept at a default level and the training and test set was divided into 80% and 20% respectively, and validated using 10-fold cross-validation. Adaboost algorithm had 82.2 % accuracy, Decision tree had 71.1% accuracy and 79.9% accuracy in Random Forest. Figure 2, 3 and 4 shows the classifier output of the adaboost, decision tree classifier and Random Forest applied respectively.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        5747               82.1117 %
Incorrectly Classified Instances      1252               17.8883 %
Kappa statistic                          0.629
Mean absolute error                      0.2489
Root mean squared error                  0.3638
Relative absolute error                 51.8122 %
Root relative squared error             74.2146 %
Total Number of Instances             6999

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.843    0.212    0.856      0.843    0.850      0.629    0.876     0.907     LTFU
                 0.788    0.157    0.771      0.788    0.779      0.629    0.876     0.775     ACTIVE
Weighted Avg.    0.821    0.190    0.822      0.821    0.821      0.629    0.876     0.854

=== Confusion Matrix ===

    a    b   <-- classified as
 3534  657 |    a = LTFU
  595 2213 |    b = ACTIVE
```

Figure 2: AdaBoost Classifier output

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        5002               71.4674 %
Incorrectly Classified Instances      1997               28.5326 %
Kappa statistic                          0.3622
Mean absolute error                      0.391
Root mean squared error                  0.4342
Relative absolute error                 81.3772 %
Root relative squared error             88.5928 %
Total Number of Instances             6999

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.902    0.564    0.705      0.902    0.791      0.391    0.767     0.830     LTFU
                 0.436    0.098    0.748      0.436    0.551      0.391    0.767     0.677     ACTIVE
Weighted Avg.    0.715    0.377    0.722      0.715    0.695      0.391    0.767     0.768

=== Confusion Matrix ===

    a    b   <-- classified as
 3779  412 |    a = LTFU
 1585 1223 |    b = ACTIVE
```

Figure 3: Decision Tree Output

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        5002               71.4674 %
Incorrectly Classified Instances      1997               28.5326 %
Kappa statistic                          0.3622
Mean absolute error                      0.391
Root mean squared error                  0.4342
Relative absolute error                 81.3772 %
Root relative squared error             88.5928 %
Total Number of Instances            6999

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.902    0.564    0.705      0.902    0.791      0.391    0.767     0.830     LTFU
                 0.436    0.098    0.748      0.436    0.551      0.391    0.767     0.677     ACTIVE
Weighted Avg.    0.715    0.377    0.722      0.715    0.695      0.391    0.767     0.768

=== Confusion Matrix ===

    a    b   <-- classified as
 3779  412 |    a = LTFU
 1585 1223 |    b = ACTIVE
```

Figure 4: Random Forest Output

5. Evaluation Results and Discussion

This section presents a comprehensive evaluation of the ICS model (AdaBoost and Ontology) developed for predicting interruption in treatment for People living with HIV/AIDS, assessing its performance using key metrics

AdaBoost Model Evaluation

The metrics namely Recall, Precision, Receiver Operating Characteristics, True Positive Rate(TPR). False Positive Rate (FPR) were used to compare the prediction models' overall performance. Table 2 presents the results of the different evaluation metrics. In comparing the prediction results for predicting Interruption in treatment. AdaBoost, Decision Tree, Random Forest achieved 82.1%, 71.4 % and 79.3% accuracy respectively. The graphical representation of classified instances is shown in figure 5. Figure 6, 7, 8 shows the Confusion matrices for Adaboost, Decision Tree and Random Forest respectively.

Table 2: Performance Comparison of Adaboost, DT, RF Algorithms

Evaluation Criteria	AdaBoost	Decision Tree	Random Forest
TPR	0.821	0.715	0.800
FPR	0.190	0.377	0.212
Precision	0.822	0.722	0.801
Recall	0.821	0.715	0.800
F-Measure	0.821	0.695	0.800
ROC Area	0.876	0.767	0.878
MCC	0.629	0.391	0.585
Accuracy	82.1%	71.4%	79.9%

Table 3: Statistic of Classified Instances

	Correctly Classified	Incorrectly Classified
AdaBoost	5747	1252
Decision Tree	5002	1997
Random Forest	5597	1402

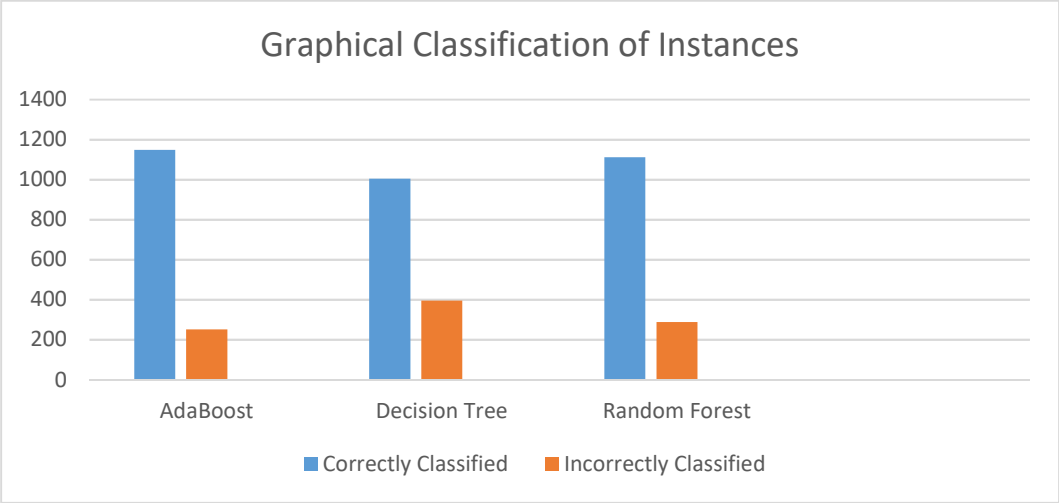


Figure 5: Graphical representation of classified instances

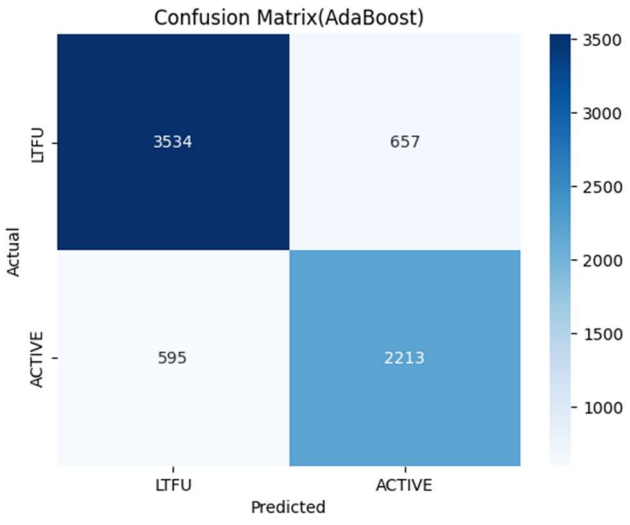


Figure 6: Confusion Matrix for AdaBoost

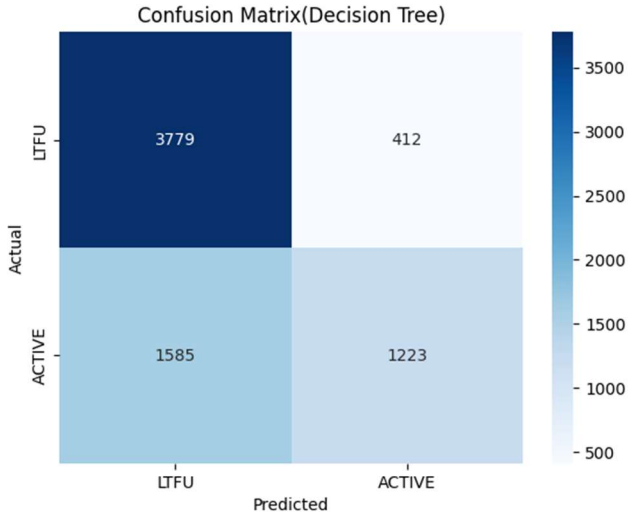


Figure 7: Confusion Matrix for DT

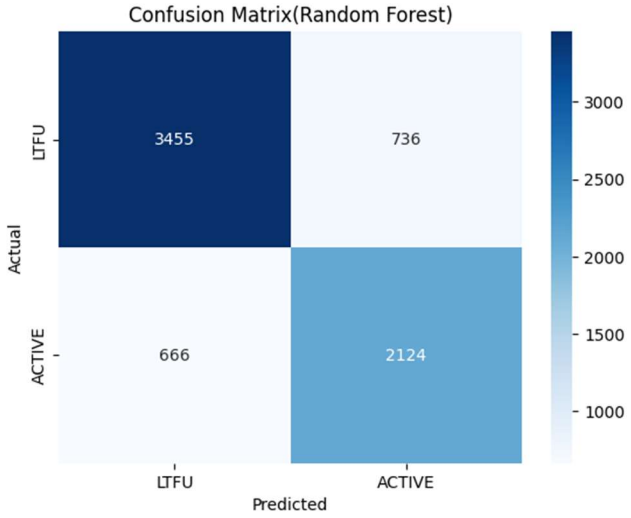


Figure 8: Confusion Matrix for RF

6. CONCLUSION

This study demonstrates the effectiveness of machine learning techniques in predicting interruptions among People Living with HIV/AIDS. By leveraging machine learning algorithms, the model can proactively identify patients at risk of treatment interruption or decline in health, enabling early intervention. Among the models tested, AdaBoost outperformed both Decision Tree and Random Forest models in terms of prediction accuracy, recall, precision; underscoring its suitability for this complex classification task. By leverging an ensemble approach, AdaBoost effectively mitigated the limitations of weak learners, delivering more reliable predictions. The findings emphasize the importance of data-driven methodologies in healthcare, particularly in identifying at-risk patients and enabling timely interventions to improve adherence to antiretroviral therapy.

REFERENCES

Abdollahi, M. (2021). Improving Medical Document Classification via Feature Engineering.

Alaeddini, A.; Yang, K.; Reddy,C.; Yu,S. (2011). A probabilistic model for predicting the probability of no-show in hospital appointments. Health Care Manag. Sci., 14, 146–157

Amith, M., Cui, L., Roberts, K., & Tao, C. (2020, July). Towards an ontology-based medication conversational agent for PrEP and PEP. In Proceedings of the conference. Association for Computational Linguistics. Meeting (Vol. 2020, p. 31). NIH Public Access.

Batool, T.; Abuelnoor, M.; El Boutari, O.; Aloul, F.; Sagahyroon (2021). Predicting Hospital No-Shows Using Machine Learning. In Proceedings of the 2020 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS), BALI,Indonesia; pp. 142–148

Che, Z., Kale, D., Li,W., Bahadori, MT., Liu, Y..(2015, August) Deep computational phenotyping. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2015 Presented at: KDD '15;; Sydney, Australia [doi: 10.1145/2783258.2783365]

Chen, L., Lu, D., Zhu, M., Muzammal, M., Samuel, O. W., Huang, G., ... & Wu, H. (2019). OMDP: An ontology-based model for diagnosis and treatment of diabetes patients in remote healthcare systems. *International Journal of Distributed Sensor Networks*, *15*(5), 1550147719847112.

Denney, J., Coyne, S., & Rafiqi, S. (2019) "Machine Learning Predictions of No-Show Appointments in a Primary Care Setting," SMU Data Science Review: Vol. 2: No. 1, Article 2. Available at: https://scholar.smu.edu/datasciencereview/vol2/iss1/2

Drancé, M., Boudin, M., Mougin, F. and Diallo, G. (2021) Neuro-symbolic XAI for Computational Drug Repurposing. DOI: 10.5220/0010714100003064 In Proceedings of the 13th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K) - Volume 2: KEOD, pages 220-225 ISBN: 978-989-758-533-3; ISSN: 2184-3228 Copyright c2021 by SCITEPRESS – Science and Technology Publications, Lda

Fahey CA, Wei L., Njau, PF., Shabani S., Kwilasa S., Maokola W. (2022) Machine learning with routine electronic medical record data to identify people at high risk of disengagement from HIV care in Tanzania. PLOS Glob Public Health 2(9): e0000720.https://doi.org/10.1371/ journal.pgph.0000720

Grosan, C., & Abraham, A. (2011). Knowledge Representation and Reasoning. Intelligent Systems, 131–147. doi:10.1007/978-3-642-21004-4_6

Hatzilygeroudis,I., & Prentzas, J. (2011) “Fuzzy and neuro-symbolic approaches to assessment of bank loan applicants,” in Artificial Intelligence Applications and Innovations, pp. 82–91, Springer,.

Hendawi R., Li, J., Roy S.(2023) A Mobile App That Addresses Interpretability Challenges in Machine Learning–Based Diabetes Predictions: Survey-Based User Study JMIR Form Res;7:e50328 URL: https://formative.jmir.org/2023/1/e50328 doi: 10.2196/50328 PMID: 37955948

Hong, L., Xu, H., & Shi, X. (2021). Constructing Ontology of Brain Areas and Autism to Support Domain Knowledge Exploration and Discovery. International Journal of Computational Intelligence Systems, 14(1), 834-846.

Joint United Nations Programme on HIV/AIDS (UNAIDS). (2021). End inequalities. End AIDS. Global AIDS Strategy 2021-2026. Joint United Nations Programme on HIV/AIDS (UNAIDS), Geneva.

Lee, G. Wang, S., Dipuro, F.; Hou, J. Grover, P., Low, L., Liu, N., Loke, C.(2017) Leveraging on Predictive Analytics to Manage Clinic No Show and Improve Accessibility of Care. In Proceedings of the IEEE International Conference on Data Science and Advanced Analytics (DSAA), Tokyo, Japan, 19–21; pp. 429–438

Liu, Y. X., Liu, X., Cen, C., Li, X., Liu, J. M., Ming, Z. Y., ... & Zheng, S. S. (2021). Comparison and development of advanced machine learning tools to predict nonalcoholic fatty liver disease: An extended study. *Hepatobiliary & Pancreatic Diseases International*, *20*(5), 409-415.

Madani, H. M., Djezzar, M., Mounir, H., Zianou, A. S., & Maimour, M. (2022). Semantic Models and Machine Learning Approach in CPS: A Survey. In *SIoT-2022: International Workshop on Semantic IoT, Co-located with the KGSWC-2022* (pp. 21-23).

Maskew, M., Sharpey-Schafer, K., De Voux, L., Crompton, T., Bor, J., Rennick, M., ... & Pisa, P. (2022). Applying machine learning and predictive modeling to retention and viral suppression in South African HIV treatment cohorts. Scientific reports, 12(1), 12715.

Massari, H. E., Mhammedi, S., Sabouri, Z., & Gherabi, N. (2022). Ontology-based machine learning to predict diabetes patients. In *Advances in Information, Communication and Cybersecurity: Proceedings of ICI2C'21* (pp. 437-445). Springer International Publishing.

Oboho-Etuk, A., Usip, P. U., Adeoye, O. S., & Ollawa, I. (2023). A Machine-Learning Based Text Classification And Machine Translator For Selected Under-Resourced Languages. In *SWIoT+ MSW@ KGSWC* (pp. 98-105).

Ogbechie, MD., Fischer, C., Lee, MT., Abba, G A, Oduola A, Idemudia, A., Edor, M., Harris, EL., Stephens, J., Gao, X., Chen, PL., (2023) Persaud NE Predicting Treatment Interruption Among People Living With HIV in Nigeria: Machine Learning Approach JMIR AI;2:e44432 URL: https://ai.jmir.org/2023/1/e44432 doi: 10.2196/44432 PMID:

Paganelli, F., & Giuli, D. (2010). An ontology-based system for context-aware and configurable services to support home-based continuous care. *IEEE Transactions on Information Technology in Biomedicine*, *15*(2), 324-333.

Praveena, M.; Krupa, J.; SaiPreethi, S. (2019) Statistical Analysis of Medical Appointments Using Decision Tree. In Proceedings of the Fifth International Conference on Science Technology Engineering and Mathematics (ICONSTEM), Chennai, India, 14–15 March 2019; Volume 1, pp. 59–64

Silva, C., Marreiros, G., & Silva, N. (2014). Development of an ontology for supporting diagnosis in psychiatry. In *Distributed Computing and Artificial Intelligence, 11th International Conference* (pp. 343-350). Springer International Publishing.

Usip, P.U., Udo, E.N., Asuquo, D.E., James, O.R. (2022). A Machine Learning-Based Mobile Chatbot for Crop Farmers. In: Ortiz-Rodríguez, F., Tiwari, S., Sicilia, MA., Nikiforova, A. (eds) Electronic Governance with Emerging Technologies. EGETC 2022. Communications in Computer and Information Science, vol 1666. Springer, Cham. https://doi.org/10.1007/978-3-031-22950-3_15

Valls, A., Gibert, K., Sánchez, D., & Batet, M. (2010). Using ontologies for structuring organizational knowledge in Home Care assistance. *international journal of medical informatics*, *79*(5), 370-387.