Multi-Class prediction of Viral Load levels among Children and Adolescent HIV Patients Using Supervised Machine Learning

Stephen Kalyesubula[1], Yusuf Kyambadde[2]

skalyesubula@gmail.com[1], Department of Computer Science[1], kyambaddeyusuf@gmail.com[1] Department of Computer and Electrical Engineering[2], Makerere University, Kampala Uganda

Introduction

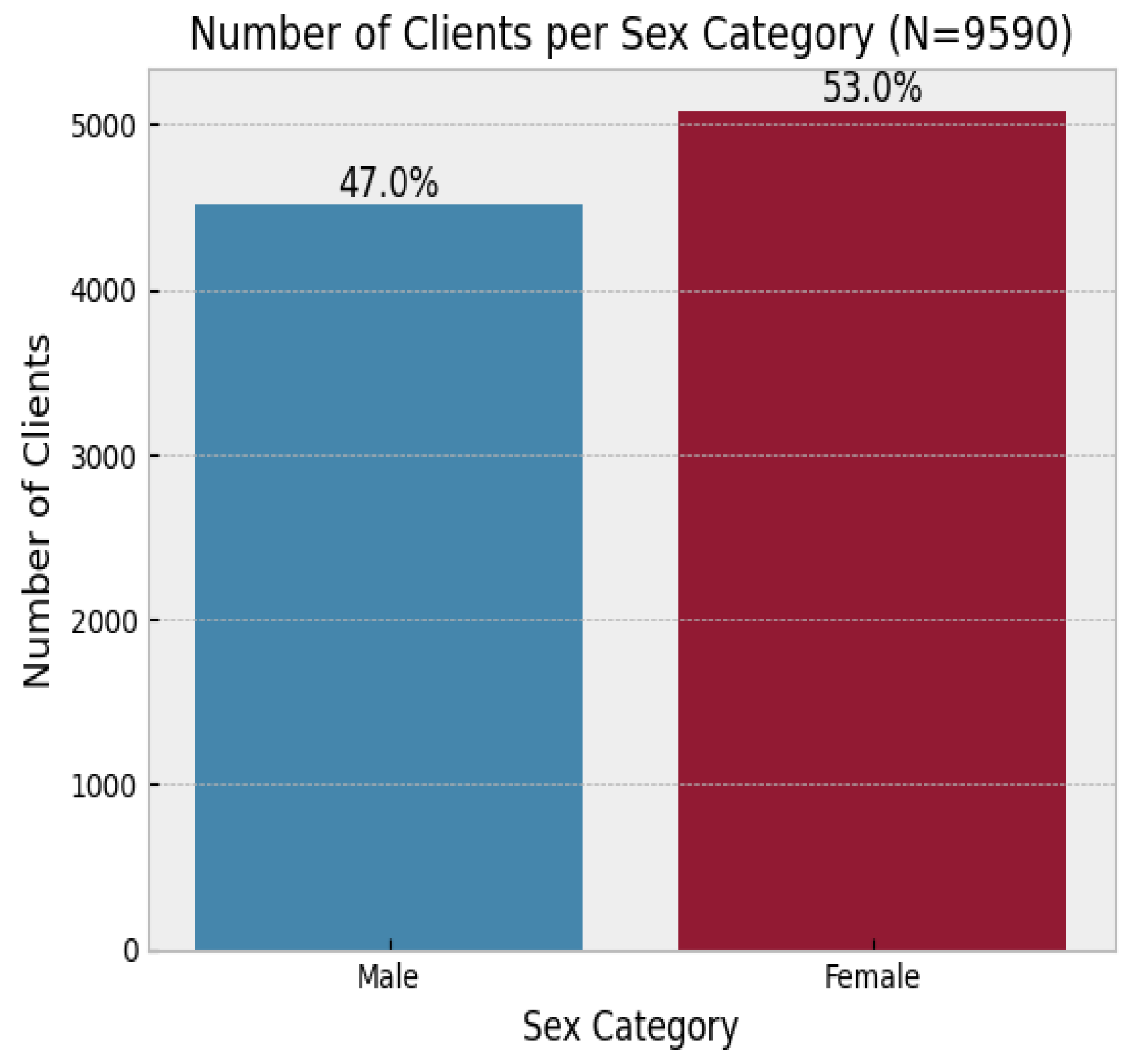
Human immunodeficiency virus (HIV) leads to a haunting chronic infection that is progressive. Currently, there is no cure, once one gets it, they have it for the rest of their life. But with proper and effective ART medical care, HIV can be controlled and can live long, have healthy lives, and protect their partners. According to the 2020 Uganda Population-based HIV Impact Assessment (UPHIA) 2020, 15-19 years of age had the lowest HIV prevalence. The report further highlights a higher prevalence in young women than in young men. The HIV prevalence among young people aged 15 to 24 years was 1.8% and it was three times higher among women at 2.9% compared to men at 0.8%. [1]

To effectively and efficiently endure ART treatment, it is essential to track HIV viral load (VL) levels routinely as a treatment outcome. According to the Ministry of Health National Antiretroviral Treatment Guidelines for Adults, Adolescents, and Children (2019), Viral load (VL) results are classified into three levels namely < 50copies/mL, ≥50 999copies/mL, and ≥ 1,000copies/mL. All patients with VL < 1000 copies/mL are regarded as suppression and the above is taken as non-suppression. [2], [3]

This paper delves into the application of supervised machine-learning techniques to predict the viral load classes for HIV+ patients based on demographic and clinical variables. The predictive performance of these models is evaluated, and exploratory data analysis is performed to visualize the trends and partners of different key variables.

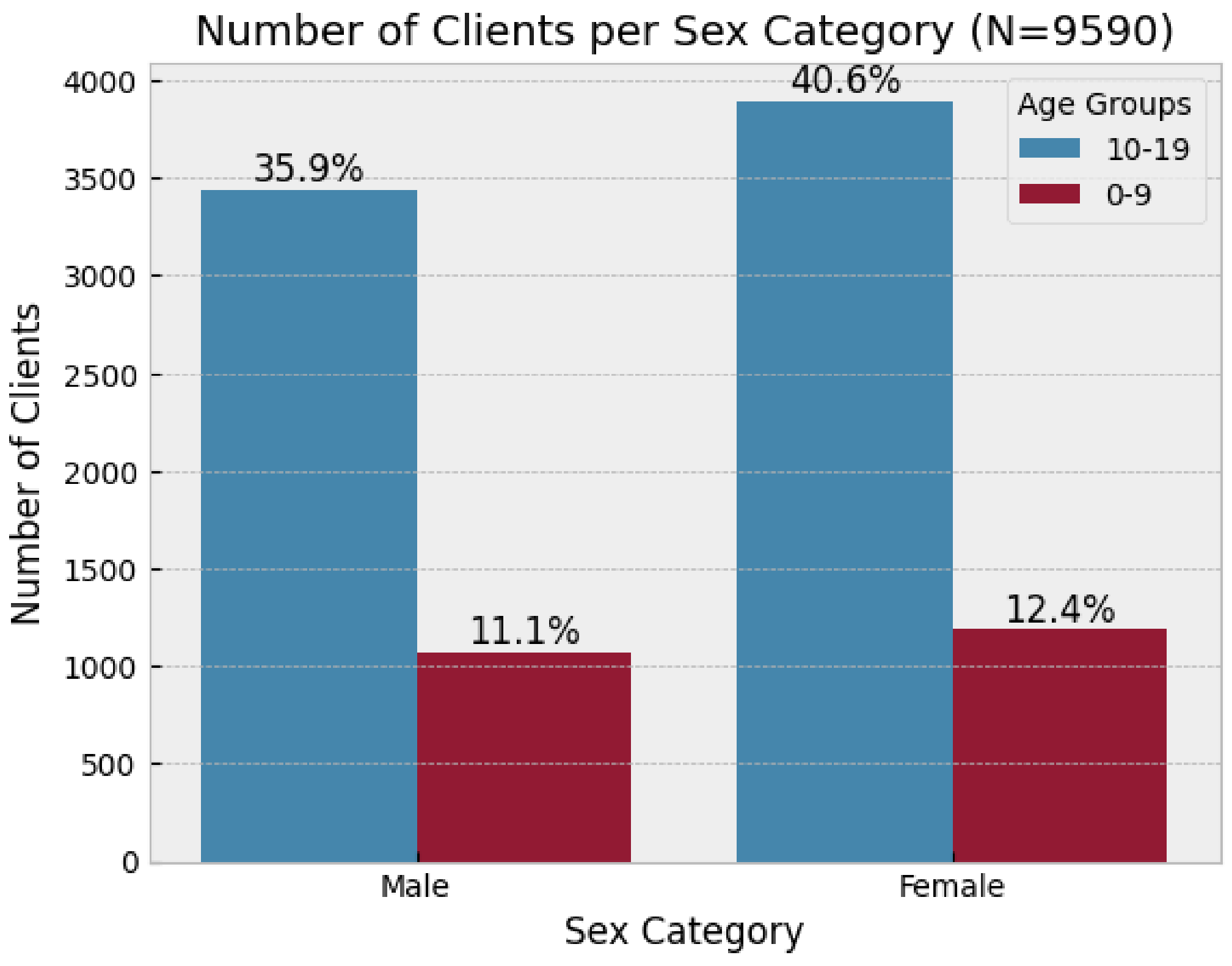
Data Representation

Our research utilized demographic and clinical point of care service data from 188 ART Health facilities in Uganda. The total data set constitutes of 9590 HIV+ patients.



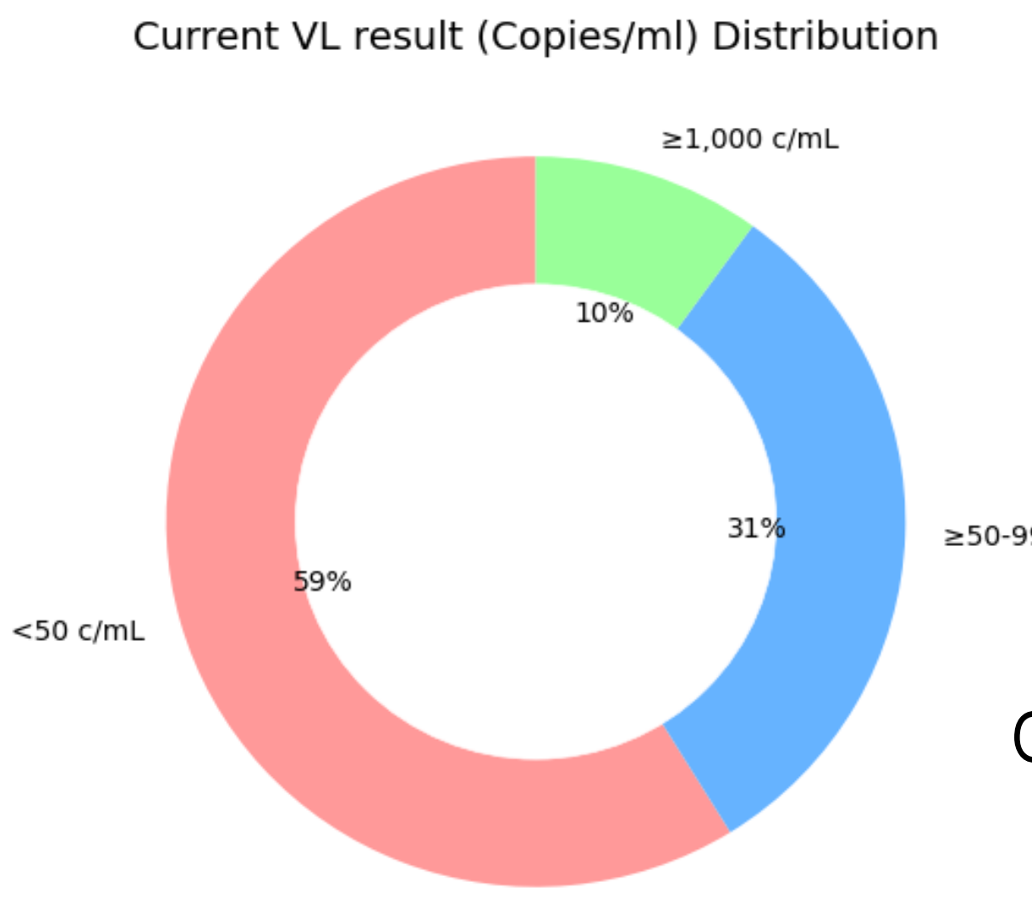
(Figure 1) Gender Distribution

53% (9590) are female, 47% (9590) are male. Looking a disaggregation by age group; approximately 41% female clients are between 10-19 years and 36% male clients are between 10-19 years. The data set has more HIV+ clients in 10-19 age group compared to 0-9 years age group.



(Figure 2) Gender Distribution by age-groups

A. Target Variable Distribution [Viral Load Levels]



(Figure 3 VL Distribution

59%, 31%, 10% represent HV+ clients with viral load levels < 50copies/mL, < 50copies/mL and ≥ 1,000copies/mL respectively.

Overall, 70% HV+ clients have suppressed VL results.

B. Impact of Drug Distribution towards VL

One of the key interventions offered to HIV+ clients is to create multiple differentiated service delivery approaches which vary the "when, where, who, and what" of service delivery to create models that respond to different needs, contexts, and groups of people. The aim of differentiated service delivery is to improve patient outcomes, enhance patient satisfaction, and optimize resource utilization. By providing patient-centered care that considers these different dimensions, healthcare systems can ensure that HIV patients receive the right care, at the right time, in the right place, and from the right providers.

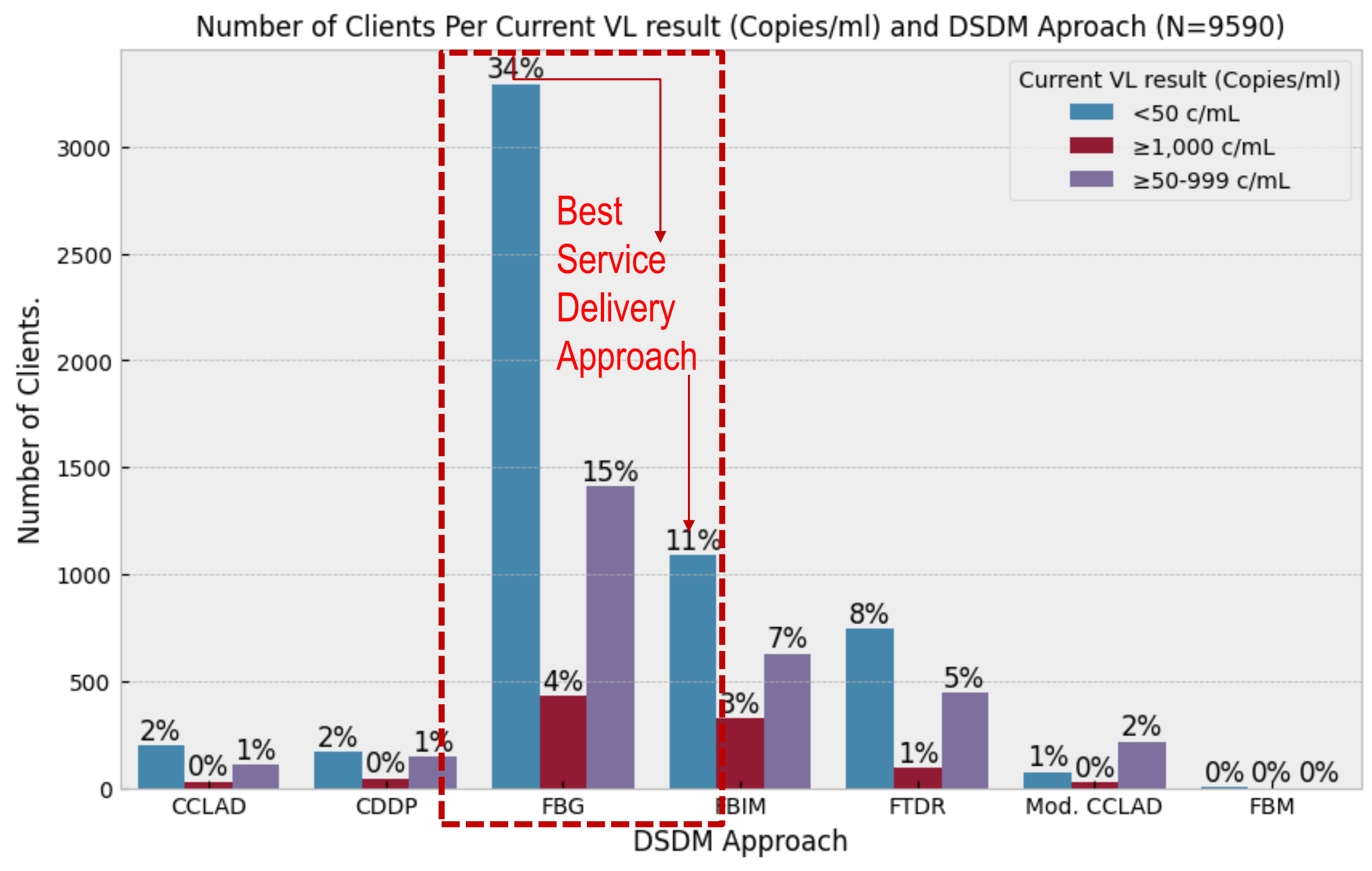


Figure 3 Best Differentiated Service Delivery Model

34% of HIV+ clients on Facility Based Group [FBG] are having a very viral load level of < 50 copies/mL. 11% are on the Facility Based Individual Management [FBIM] differentiated service model. Therefore the above two models are regarded as the best service distribution approach to achieve maximum suppression.

Methodology

A methodological approach to data preparation and data anonymization was taken.

Data Anonymization: The data set was further abstracted to exclude any patient-identifiable information including the names, Art numbers, date of birth, phone numbers, location, and health facility. Random unique IDs were generated for the different clients to support the model training and testing. Features (columns) with more than 10% of missing data (null values) were dropped to reduce the underfitting or overfitting within the models.

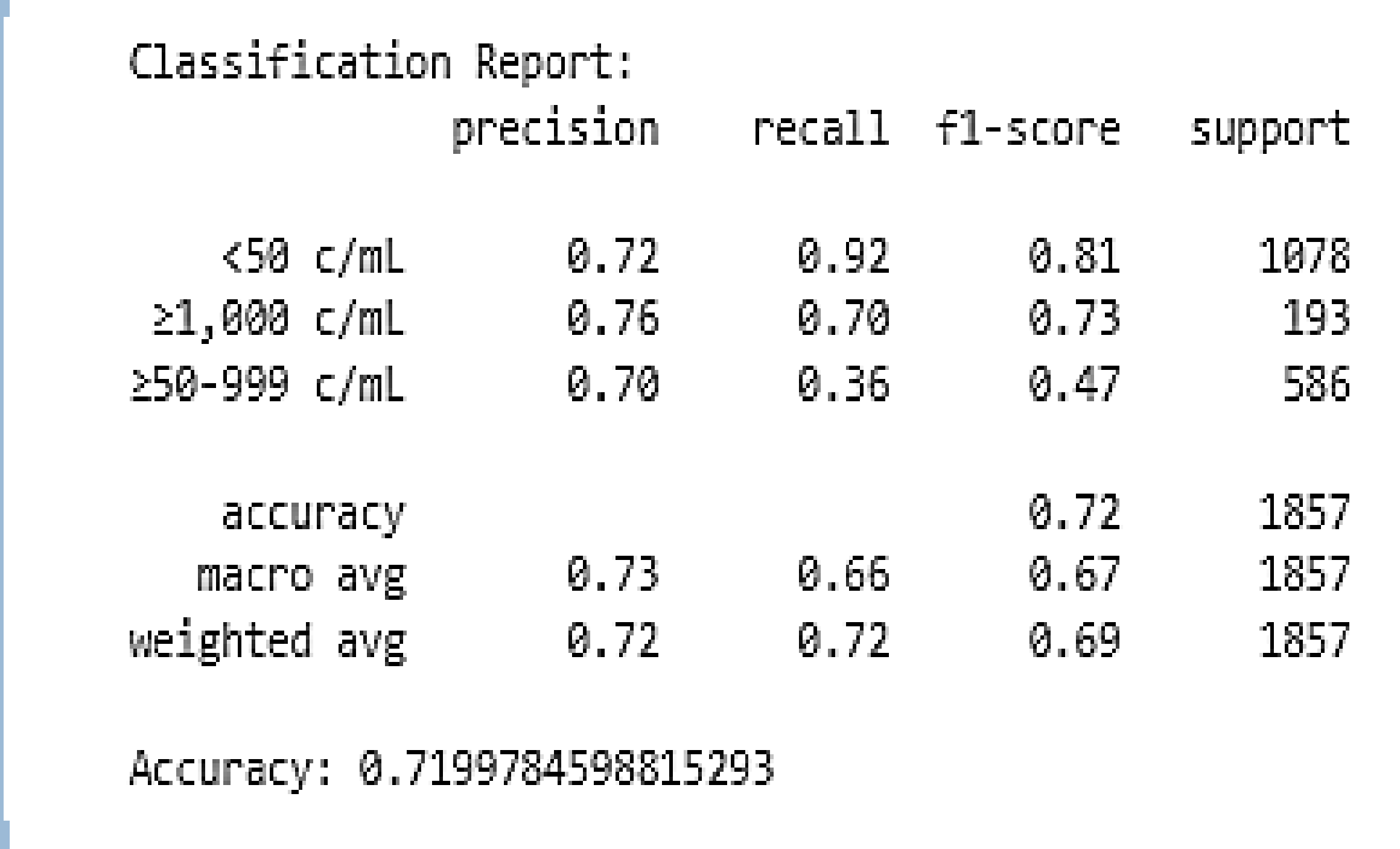
Data Imputation: Since this is health data related to service delivery, proper imputation approaches need to be followed, for categorical variables, a *"Missing Data"* categorical variable was created to cater to the missing values and also allow the model to learn from data sets with null values. For the null numerical values, KNNImputer with *neighbors* of 10 is used to fill in missing values in a dataset by estimating them based on the values of their k-nearest neighbors. Data-time variables were also converted to numerical variables.

Data Encoding: For categorical variables, the label encoding technique was used to convert categorical values into numerical values by assigning a unique integer to each category. Each category is mapped to a different integer, allowing categorical data to be represented in a format that machine learning algorithms can work with.

Model Preparation: The Current VL result (Copies/ml) feature was selected as a target variable. The data set with a shape (9285 clients, 30 labels) was split into the test and train data sets using a test size of 0.2

Results: Random Forest Classifier [RFC]

Using the 100 estimators and a random state of 42, the model was trained on the train set and tested on the test set. Below is the classification report. Using model evaluation metrics, an **accuracy of 71.9% was achieved.**



```
Classification Report:
              precision    recall  f1-score   support

    <50 c/mL       0.72      0.92      0.81      1078
  ≥1,000 c/mL       0.76      0.70      0.73       193
  ≥50-999 c/mL       0.70      0.36      0.47       586

     accuracy                           0.72      1857
    macro avg       0.73      0.66      0.67      1857
 weighted avg       0.72      0.72      0.69      1857

Accuracy: 0.7199784598815293
```

RFC Hyperparameter Tuning & Confusion Matrix

Increasing the number of estimators to 300 while running the model in a for loop to obtain the maximum accuracy. **The accuracy increased to 73.4%.**

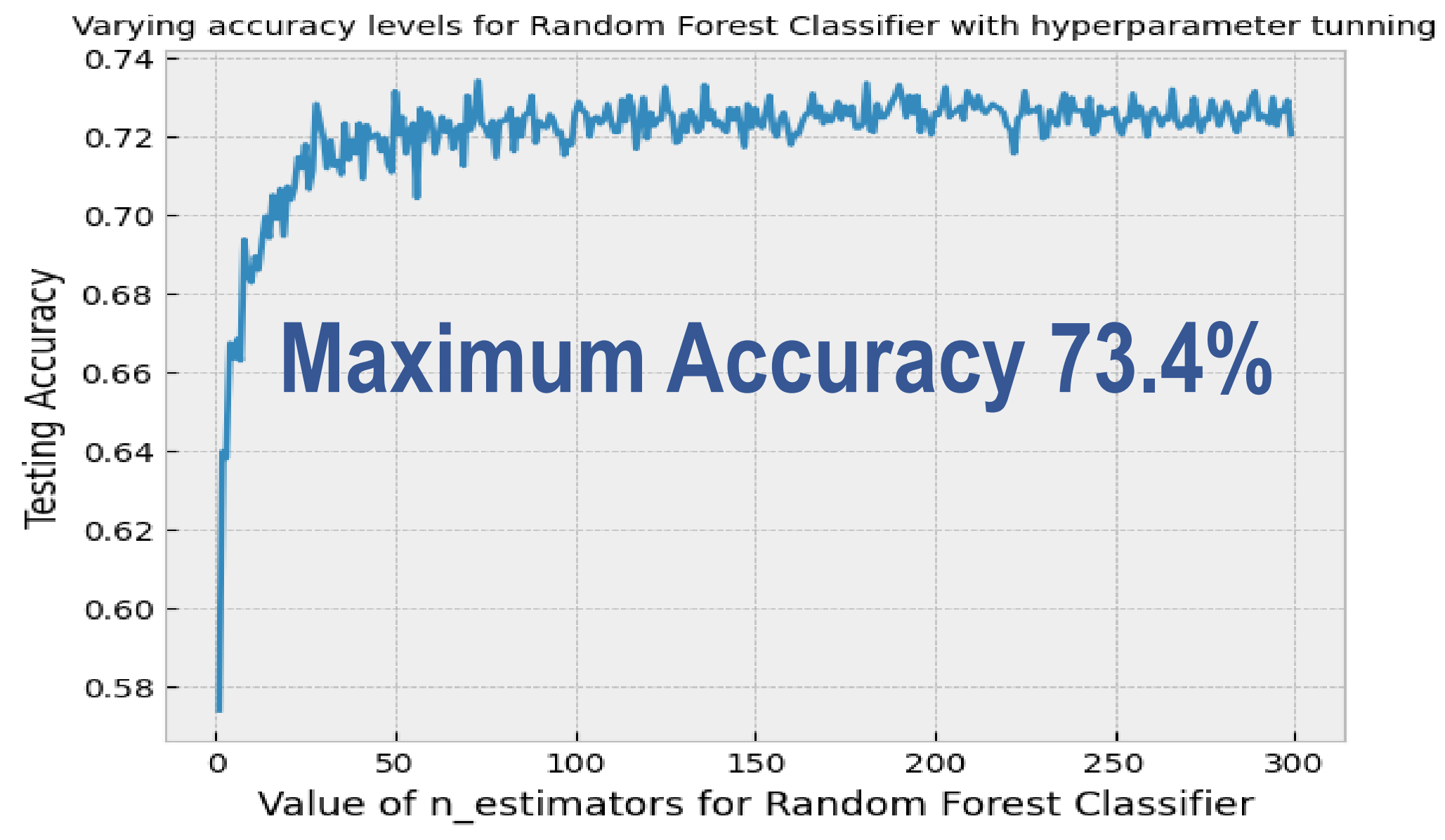


Figure 4 Varying Accuracy

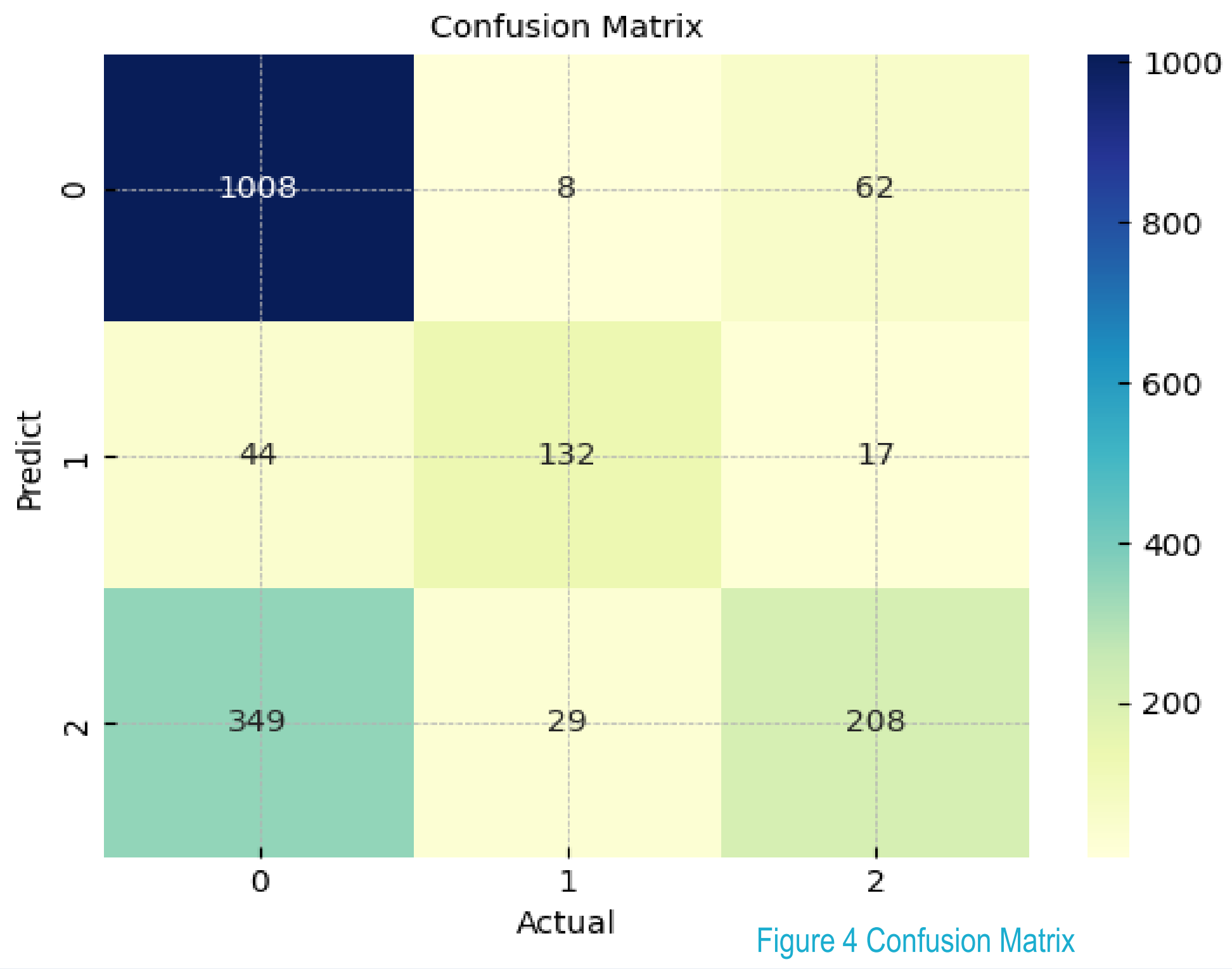
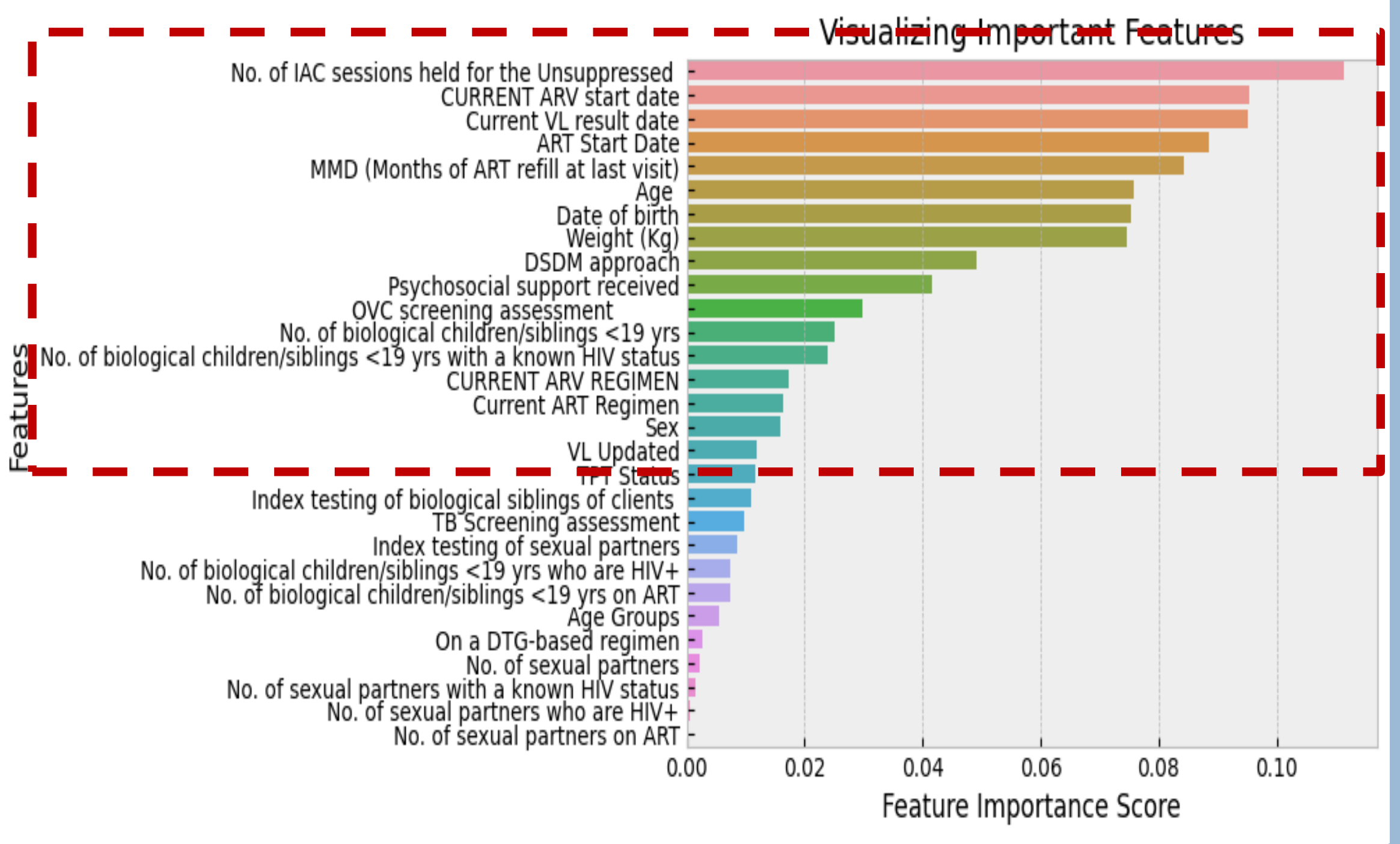


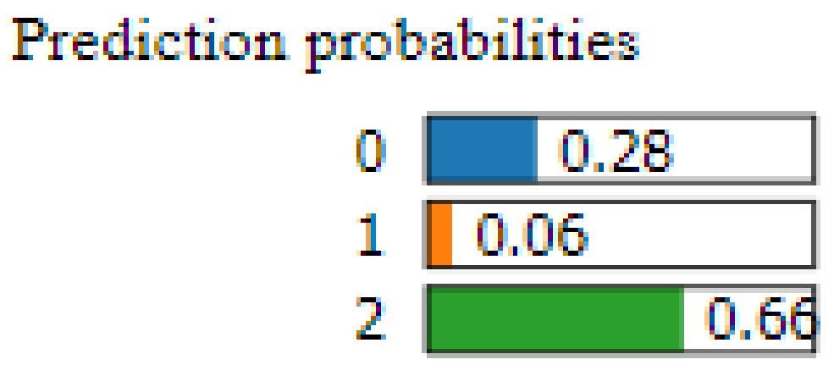
Figure 4 Confusion Matrix

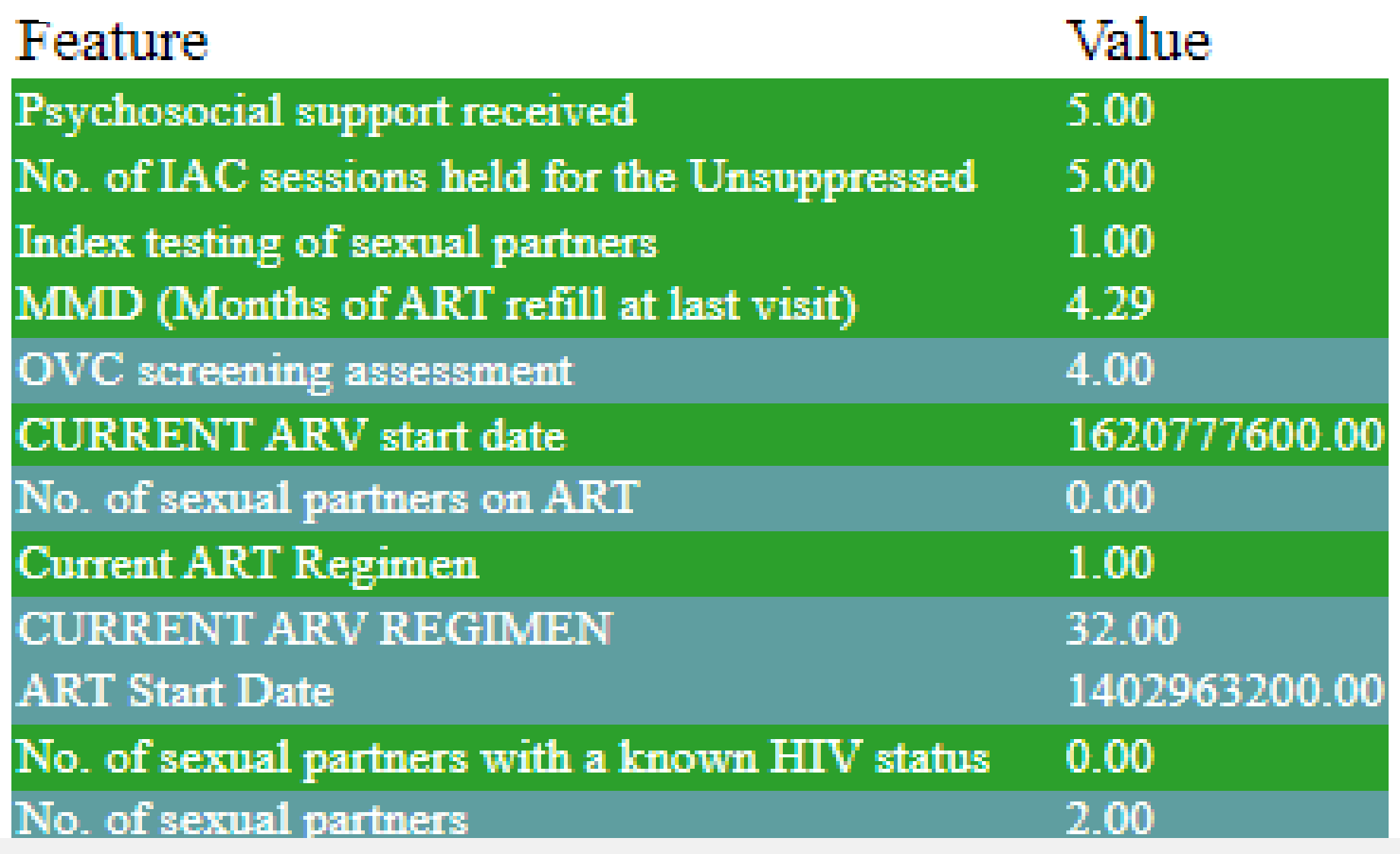
Key Predictor Services and Variables



Explainable AI with LIME

Considering patient with an ID: 6030. We are able to know the key services that determine the VL levels for the patient. [0, 1, 2] = [< 50copies/mL, ≥50 999copies/mL, and ≥ 1,000copies/Ml]





Feature	Value
Psychosocial support received	5.00
No. of IAC sessions held for the Unsuppressed	5.00
Index testing of sexual partners	1.00
MMD (Months of ART refill at last visit)	4.29
OVC screening assessment	4.00
CURRENT ARV start date	1620777600.00
No. of sexual partners on ART	0.00
Current ART Regimen	1.00
CURRENT ARV REGIMEN	32.00
ART Start Date	1402963200.00
No. of sexual partners with a known HIV status	0.00
No. of sexual partners	2.00

Conclusion

The study shows how to use supervised machine learning and explainable AI techniques to understand the key services that determine the VL levels of HIV+ clients. Random Forest classifier with adequate calibration properties is able to achieve a desirable accuracy of 73.4% classification and general predictive performance. Therefore, ML has the capability of tracking cohorts of patients who are likely to be unstable over a long period of time and also informs on the quality improvement strategies to ensure that unstable patients become stable.

References

[1] Uganda Population-based HIV Impact Assessment (UPHIA), 2020

[2] Uganda Ministry of Health, https://www.health.go.ug/2022/05/31/uganda-records-significant-drop-in-mother-to-child-hiv-infections/

[3] National Antiretroviral Treatment Guidelines for Adults, Adolescents, and Children, 3rd Edition June 2009, STD/AIDS Control Programme, Ministry of Health