

# Multi-class Discriminative Classification Model on Antiretroviral Therapy Reaction and Failure Developed on the Unique Records of the Akwa Ibom HIV Database

Nicelle Sernadilla Macaspac

August 2023

## Introduction

- HIV has no known cure, but the infected patient is treated with highly active antiretroviral therapy (HAART) [3], mainly for the purpose of suppressing the viral load (the amount of HIV in the blood stream) and prolonging the life expectancy of the patient (Ekpenyong, Etebong, and Jackson 2019, 1).

-Current antiretroviral therapy does not cure HIV infection but allows a person's immune system to get stronger. This helps them to fight other infections (World Health Organization, n.d.)

- machine learning

This undertaking is part of the capstone in the Professional Certificate Program in Data Science of Harvard Online. The corresponding R Markdown and R files are in the GitHub of nsmacaspac.

## Unique Records of the Akwa Ibom HIV Database

In a previous study on patient response to antiretroviral therapy, Ekpenyong, Etebong, and Jackson (2019, 3) used a database of patients who received treatment for HIV from thirteen health centers in Akwa Ibom, Nigeria, between 2015 and 2018. Two years later, they published the processed dataset (Ekpenyong et al. 2021b, Appendix) with minor oversight in the accompanying article, which were easily reconciled through the 2019 study and were appropriately referenced throughout this project. The processed dataset is composed of an Individual Treatment Change Episodes table with a column for each antiretroviral drug administered and a concatenated Unique Records table with the drugs combined into a column for each antiretroviral therapy of three drugs administered. For the purpose of this project, we utilize only the Unique Records table.

The Unique Records table was imported with the corresponding `read_xlsx` function in the language R. The dataset is composed of 1,056 patient records, each with 15 columns: patient identification, sex, baseline CD4 count, follow-up CD4 count, baseline RNA load, follow-up RNA load, baseline weight, follow-up weight, drug combination, and patient response and drug reaction classifications 1 to 5 (fig. 1).

```
dataset <- as.data.frame(read_xlsx("mmc1.xlsx", range = "N27:AB1083")) # range reads  
# the Unique Records table, which is the combined version of the Individual Treatment Change Episodes
```

The immunological marker CD4 count is given in cells per cubic millimeter (Ekpenyong et al. 2021a, 8). The viral RNA load is expressed in times  $10^2$  copies (Ekpenyong, Etebong, and Jackson 2019, 10). The

Unique Records										[Target Classes]				
PID	SEX	BCD4	FCD4	BRNA	FRNA	BWt(kg)	FWt(kg)	DRUGCOMB	PR	C1	C2	C3	C4	C5
1	F	148	106	3	1.3	42	43	TDF+3TC+EFV	53.56	0	0	1	0	0
2	F	145	378	2.5	1.3	57	60	AZT+3TC+NVP	55.33	0	0	0	1	0
3	M	78	131	4.1	1.7	70	75	AZT+3TC+NVP	50.00	0	1	0	0	0
4	M	295	574	4.4	1.9	64	66	AZT+3TC+NVP	50.00	0	0	1	0	0
5	F	397	792	1.9	1.3	52	55	AZT+3TC+NVP	76.00	0	0	0	0	1

Figure 1: First rows of the Unique Records table of the Akwa Ibom HIV Database.

weight ranges from 4.7 to 125 kg on account of the presence of pediatric patients (Ekpenyong, Etebong, and Jackson 2019, 2). The three-drug combinations of antiretroviral therapy are a complementary mix of nucleoside reverse transcriptase inhibitors tenofovir (TDF), lamivudine (3TC) and zidovudine (AZT), and non-nucleoside reverse transcriptase inhibitors efavirenz (EFV) and nivarapine (NVP) given in the first 6 months of treatment (Ekpenyong et al. 2021a, 8).

Patient response and drug reaction were quantified and classified in the 2019 study using expert domain knowledge and the advanced method of interval type-2 fuzzy logic system (Ekpenyong, Etebong, and Jackson 2019, 4, 7). The drug reaction classification uses a binary system to indicate very high interaction (C1), high interaction (C2), low interaction (C3), very low interaction (C4), and no interaction (C5; Ekpenyong, Etebong, and Jackson 2019, 11). A low response rate signifies high to very high drug interactions and treatment failure, whereas a high response rate denotes low to no drug interactions (Ekpenyong, Etebong, and Jackson 2019, 10, 13).

### Tidy Dataset

The dataset was rendered into tidy format to prepare it for preprocessing. The fifteen columns were re-named consistently with their aforementioned descriptions, with `vhi_tf` corresponding to very high interaction\_treatment failure and `ni` corresponding to no interaction. Missing values were not detected.

```
colnames(dataset) <- c("id", "sex", "bcd4", "fcd4", "brna", "frna", "bweight", "fweight",
"therapy", "response", "vhi_tf", "hi_tf", "li", "vli", "ni")
head(dataset, n = 5)
##   id sex bcd4 fcd4 brna frna bweight fweight   therapy response vhi_tf hi_tf
## 1  1  F  148  106  3.0  1.3    42    43  TDF+3TC+EFV 53.56199      0      0
## 2  2  F  145  378  2.5  1.3    57    60  AZT+3TC+NVP 55.33422      0      0
## 3  3  M   78  131  4.1  1.7    70    75  AZT+3TC+NVP 50.00000      0      1
## 4  4  M  295  574  4.4  1.9    64    66  AZT+3TC+NVP 50.00000      0      0
## 5  5  F  397  792  1.9  1.3    52    55  AZT+3TC+NVP 76.00000      0      0
##   li vli ni
## 1  1  0  0
## 2  0  1  0
## 3  0  0  0
## 4  1  0  0
## 5  0  0  1
```

The `brna` and `frna` columns were multiplied by  $10^2$  to simplify the unit from times  $10^2$  copies to just copies. This aligns them with the unit used for viral RNA load in the WHO definition of HIV (World Health Organization, n.d.).

The `vhi_tf`, `hi_tf`, `li`, `vli`, and `ni` columns were verified to have only one value per row. Hence, the binary system was relabeled as `vhi_tf` to `ni` using the `case_when` function and merged under a newly defined `dreaction` column. This brought down the number of columns to eleven.

```

dataset1 <- dataset |>
  mutate(brna = brna*10^2) |>
  mutate(frna = frna*10^2) |> # simplifies the unit from times 10^2 copies to just copies
  mutate(dreaction = case_when(vhi_tf == 1 ~ "vhi_tf",
                                hi_tf == 1 ~ "hi_tf",
                                li == 1 ~ "li",
                                vli == 1 ~ "vli",
                                ni == 1 ~ "ni")) |> # relabels drug reactions as
  # vhi_tf to ni and merges them under a newly defined dreaction column
  select(-vhi_tf, -hi_tf, -li, -vli, -ni)
head(dataset1, n = 5)
##   id sex bcd4 fcd4 brna frna bweight fweight therapy response dreaction
## 1  1  F  148  106  300  130    42    43 TDF+3TC+EFV 53.56199      li
## 2  2  F  145  378  250  130    57    60 AZT+3TC+NVP 55.33422      vli
## 3  3  M   78  131  410  170    70    75 AZT+3TC+NVP 50.00000     hi_tf
## 4  4  M  295  574  440  190    64    66 AZT+3TC+NVP 50.00000      li
## 5  5  F  397  792  190  130    52    55 AZT+3TC+NVP 76.00000      ni

```

## Preprocessed Dataset

The sex, therapy, and dreaction variables of the dataset were changed from character to numeric class using the factor and as.numeric functions to allow for numerical data examination. All eleven variables showed good data variability.

## ??? DISTRIBUTION

The correlation coefficients across variables were calculated and visualized using the cor and corrplot functions (fig. 2; Wei and Simko 2021).

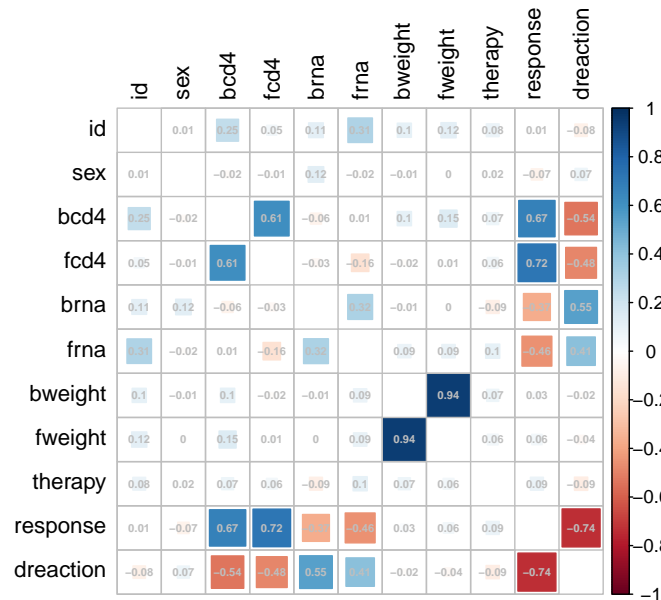


Figure 2: Matrix of the correlation coefficients between variables.

Patient response and drug reaction highly correlated with each other, having an absolute coefficient of 74% (fig. 2). This was expected as they refer to similar information. Hence, only the dreaction variable was retained as an outcome for the purpose of this project.

On the other hand, CD4 count was negatively correlated with drug reaction (-48% and -54%) whereas RNA load was positively correlated with it (41% and 55%), while having minimal correlation with each other (fig. 2). This meant that the bcd4, fcd4, brna and frna variables contained distinct information relevant to drug reaction and were kept as predictors.

The preprocessed dataset was streamlined to contain only these pertinent predictors and outcome for faster classification modeling. The dreaction variable was changed from character to factor class as required for outcomes in classification models.

```
dataset2 <- dataset1 |>
  select(bcd4, fcd4, brna, frna, dreaction) |> # keeps CD4 count and RNA load as
  # predictors and drug reaction as outcome
  mutate(dreaction = factor(dreaction, c("ni", "vli", "li", "hi_tf", "vhi_tf")))
head(dataset2, n = 5)
##   bcd4 fcd4 brna frna dreaction
## 1  148  106  300  130         li
## 2  145  378  250  130        vli
## 3   78  131  410  170      hi_tf
## 4  295  574  440  190         li
## 5  397  792  190  130         ni
```

## Predictors and Outcome

The predictors CD4 count and RNA load are measures of immune system health and HIV status, respectively (Ekpenyong, Etebong, and Jackson 2019, 10; World Health Organization, n.d.). They are crucial markers of drug reaction and treatment failure that can be obtained from simple laboratory tests in settings with limited resources for the costly HIV genotyping to determine drug resistance (Isaakidis et al. 2010, 7; Revell et al. 2010, 605).

The interactions of these predictors with drug reaction were further visualized using ggplot2 functions (fig. 3). The plots showed a general increase in CD4 count and decrease in RNA load at follow-up. This indicated the robustness of the immune system of most patients in suppressing the HIV and positive patient response to the treatment (Ekpenyong, Etebong, and Jackson 2019, 10).

However, the plots also exhibited patients who retained a low CD4 count and high RNA load at follow-up (fig. 3). They were the same cases classified as with high drug interaction and treatment failure.

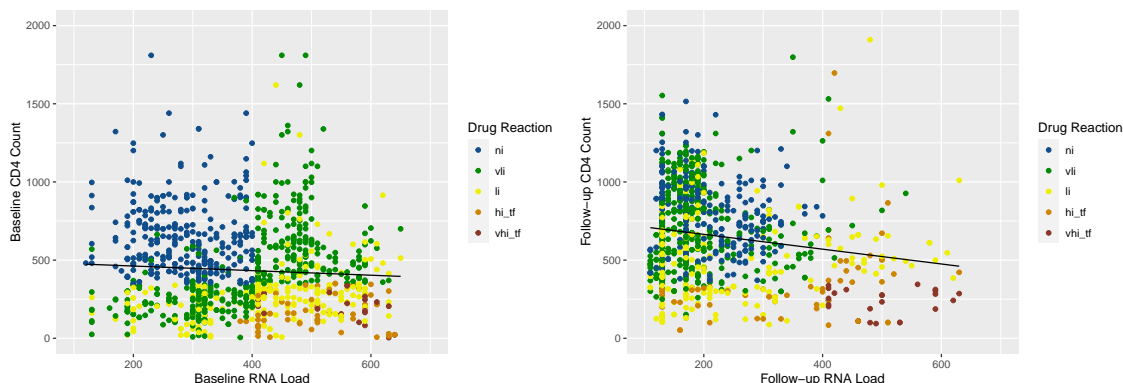


Figure 3: Scatterplots of CD4 count, RNA load, and the corresponding drug reaction at baseline and follow-up.

## Classification Models

In the final dataset with the bcd4, fcd4, brna, and frna variables as predictors and the dreaction variable as outcome, there were only 1,056 observations. The dataset was split into 80% train set and 20% test set using the createDataPartition function to maximize the data available for both training and testing. The function additionally preserves the distribution of the five outcome classes of the dreaction variable: ni, vli, li, hi\_tf, and vhi\_tf (Kuhn 2019).

```
set.seed(20, sample.kind = "Rounding") # if using R 3.6 or later
# set.seed(20) # if using R 3.5 or earlier
# for reproducibility during assessment
train_index <- createDataPartition(dataset2$dreaction, p = 0.8, list = FALSE)
train_set <- dataset2[train_index,]
test_set <- dataset2[-train_index,]
```

Classification models introduced in the program that can possibly accommodate the nonlinear nature of the decision boundaries of the outcome classes (similar to what was observed in fig. 3) were successively optimized on the train set with the train function (Haugh 2017, Irizarry 2002). We discuss them in the following subsections. The streamlined dataset allowed for the use of the default resampling method of the function that is bootstrapping for twenty-five times, which would have been time-consuming otherwise. Thereafter, the models were evaluated for the accuracy of their outcome prediction in the test set. The sensitivity or true positive rate and specificity or true negative rate of their prediction were also checked to rule out the effect of prevalence (Irizarry 2002).

### *k*-Nearest Neighbor Classification

The k-Nearest Neighbor Classification determines the class of a data point based on the majority class among its neighboring data points (Haugh 2017). The number of neighbors k can be optimized to maximize the accuracy (Irizarry 2002). In the train set, the optimal model was tuned at 3 neighbors, with an accuracy of 80.4%.

Its predicted drug reaction was obtained using the predict function. This was compared with the observed drug reaction in the test set utilizing the confusionMatrix function, which resulted in an accuracy of 88%. The accuracy came with good sensitivity and specificity for most outcome classes, but the model had a low true positive detection rate for the class hi\_tf.

```
set.seed(30, sample.kind = "Rounding") # if using R 3.6 or later
# set.seed(30) # if using R 3.5 or earlier
knn_model <- train(dreaction ~ ., train_set, method = "knn",
  tuneGrid = data.frame(k = seq(3, 9, 1))) # tunes neighbor number
knn_model
## k-Nearest Neighbors
##
## 847 samples
## 4 predictor
## 5 classes: 'ni', 'vli', 'li', 'hi_tf', 'vhi_tf'
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 847, 847, 847, 847, 847, 847, ...
## Resampling results across tuning parameters:
##
## k Accuracy Kappa
```

```
## 3 0.804 0.719
## 4 0.797 0.710
## 5 0.799 0.711
## 6 0.797 0.709
## 7 0.796 0.707
## 8 0.798 0.710
## 9 0.798 0.710
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was k = 3.

knn_dreaction <- predict(knn_model, test_set)
head(knn_dreaction)
## [1] hi_tf li vhi_tf li vli vli
## Levels: ni vli li hi_tf vhi_tf
confusionMatrix(knn_dreaction, test_set$dreaction)$overall["Accuracy"]
## Accuracy
## 0.88
confusionMatrix(knn_dreaction, test_set$dreaction)$byClass[, 1:2]
##
## Sensitivity Specificity
## Class: ni 0.932 0.967
## Class: vli 0.893 0.920
## Class: li 0.854 0.957
## Class: hi_tf 0.615 0.995
## Class: vhi_tf 1.000 0.990
```

### *Recursive Partitioning and Regression Trees Model*

The Recursive Partitioning and Regression Trees model forms successive dichotomous data splitting to derive the class of a data point, such that it looks like an inverted tree (ScienceDirect 2007). The complexity parameter at which the tree will be pruned in the case of an unimportant data splitting can be tuned (Therneau, Atkinson, and Ripley 2022). This model was optimized on the train set at a complexity parameter of 0 and an accuracy of 96.1%.

The test accuracy of its outcome prediction was at 97.6%, higher than that from the preceding model and with high sensitivity and specificity for all outcome classes.

```
set.seed(40, sample.kind = "Rounding") # if using R 3.6 or later
# set.seed(40) # if using R 3.5 or earlier
rpart_model <- train(dreaction ~ ., train_set, method = "rpart",
  tuneGrid = data.frame(cp = seq(0, 0.1, 0.01))) # tunes tree complexity
rpart_model
## CART
##
## 847 samples
## 4 predictor
## 5 classes: 'ni', 'vli', 'li', 'hi_tf', 'vhi_tf'
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 847, 847, 847, 847, 847, 847, ...
## Resampling results across tuning parameters:
##
```

```
##      cp    Accuracy    Kappa
##    0.00    0.961      0.944
##    0.01    0.959      0.942
##    0.02    0.950      0.929
##    0.03    0.923      0.890
##    0.04    0.907      0.866
##    0.05    0.898      0.853
##    0.06    0.888      0.838
##    0.07    0.875      0.819
##    0.08    0.866      0.805
##    0.09    0.856      0.790
##    0.10    0.839      0.763
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was cp = 0.

rpart_dreaction <- predict(rpart_model, test_set)
confusionMatrix(rpart_dreaction, test_set$dreaction)$overall["Accuracy"]
## Accuracy
##      0.976
confusionMatrix(rpart_dreaction, test_set$dreaction)$byClass[, 1:2]
##              Sensitivity Specificity
## Class: ni           0.966         1.000
## Class: vli          1.000         0.984
## Class: li           0.958         0.994
## Class: hi_tf        0.923         0.990
## Class: vhi_tf       1.000         1.000
```

### *Rborist Model*

The Rborist model improves on the prediction of the Recursive Partitioning and Regression Trees model using the average of multiple trees to classify a data point (Irizarry 2002). Its accuracy can be optimized through the number of predictors for data splitting and the minimum size of its terminal node. In the train set, the optimal model was tuned at 4 predictors and node size of 1, with an accuracy of 97.7%.

Its predicted drug reaction had a test accuracy of 99%, higher than those from the first two models. The accuracy was supported by excellent sensitivity and specificity for all outcome classes.

```
set.seed(50, sample.kind = "Rounding") # if using R 3.6 or later
# set.seed(50) # if using R 3.5 or earlier
rborist_model <- train(dreaction ~ ., train_set, method = "Rborist",
  tuneGrid = expand.grid(predFixed = seq(1, 4), minNode = seq(1, 4))) # tunes combinations
  # of predictor number and terminal node size
rborist_model
## Random Forest
##
## 847 samples
## 4 predictor
## 5 classes: 'ni', 'vli', 'li', 'hi_tf', 'vhi_tf'
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 847, 847, 847, 847, 847, 847, ...
```

```
## Resampling results across tuning parameters:
##
##   predFixed minNode Accuracy Kappa
##   1         1      0.964  0.949
##   1         2      0.965  0.950
##   1         3      0.964  0.948
##   1         4      0.963  0.947
##   2         1      0.975  0.964
##   2         2      0.975  0.964
##   2         3      0.975  0.964
##   2         4      0.975  0.964
##   3         1      0.977  0.967
##   3         2      0.977  0.967
##   3         3      0.977  0.967
##   3         4      0.977  0.966
##   4         1      0.977  0.968
##   4         2      0.977  0.968
##   4         3      0.977  0.967
##   4         4      0.977  0.967
##
## Accuracy was used to select the optimal model using the largest value.
## The final values used for the model were predFixed = 4 and minNode = 1.

rborist_dreaction <- predict(rborist_model, test_set)
confusionMatrix(rborist_dreaction, test_set$dreaction)$overall["Accuracy"]
## Accuracy
##      0.99
confusionMatrix(rborist_dreaction, test_set$dreaction)$byClass[, 1:2]
##              Sensitivity Specificity
## Class: ni           1.000         1.00
## Class: vli           1.000         1.00
## Class: li            0.958         1.00
## Class: hi_tf         1.000         0.99
## Class: vhi_tf        1.000         1.00
```

### *Quadratic Discriminant Analysis*

The Quadratic Discriminant Analysis classifies data points with flexible quadratic decision boundaries (Haugh 2017). Unlike the previous models, this does not have a parameter for tuning to maximize its accuracy. Accordingly, this model had the lowest training accuracy of 72.7%.

The test accuracy of its outcome prediction was at 76.1%, lower than those from the preceding models. The accuracy was accompanied by good specificity, but the model had mediocre true positive detection rates for the classes li and hi\_tf.

```
set.seed(60, sample.kind = "Rounding") # if using R 3.6 or later
# set.seed(60) # if using R 3.5 or earlier
qda_model <- train(dreaction ~ ., train_set, method = "qda")
qda_model
## Quadratic Discriminant Analysis
##
## 847 samples
## 4 predictor
```



```
## 5 classes: 'ni', 'vli', 'li', 'hi_tf', 'vhi_tf'
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 847, 847, 847, 847, 847, ...
## Resampling results:
##
## Accuracy Kappa
## 0.727 0.601

qda_dreaction <- predict(qda_model, test_set)
confusionMatrix(qda_dreaction, test_set$dreaction)$overall["Accuracy"]
## Accuracy
## 0.761
confusionMatrix(qda_dreaction, test_set$dreaction)$byClass[, 1:2]
##
## Sensitivity Specificity
## Class: ni 0.814 0.980
## Class: vli 0.798 0.800
## Class: li 0.646 0.870
## Class: hi_tf 0.615 1.000
## Class: vhi_tf 1.000 0.995
```

## Multi-class Discriminative Classification Model

- discriminative classification model (Haugh 2017)
- Random Forest
- results of model performances
- the model with the best predictions on new data
- therapy is not a predictor.

-it must be noted that, due to the limited resources in the Nigerian environment, only 5 drugs combined into 3 therapies were possible (Ekpenyong, Etebong, and Jackson 2019, 3)

- not overtrained
- Two RF models were trained using .8000 TCEs without the use of genotypes, one with comprehensive and one with simplified treatment history information. Both models predicted virological response for 400 independent test cases with an accuracy of 82%. The models were able to identify alternative regimens (involving the same restricted range of drugs) that they predicted would have reduced viral load to below 50 copies/mL in almost half of the cases of actual treatment failure.<sup>11</sup> They identified regimens that were predicted to be more effective than those that failed in almost all cases. A secondary analysis of the input variables used for these models revealed that the baseline viral load was by far the most important variable (considerably more so than CD4 counts, for example) for the models in making these predictions. These studies suggest that with viral load monitoring in place, computational models could play an important future role in optimizing antiretroviral therapy in resource-limited settings. (Revell et al. 2010, 606)
- A disadvantage of random forests is that we lose interpretability. An approach that helps with interpretability is to examine variable importance. varImp function to know which predictors are used the most (Irizarry 2002)

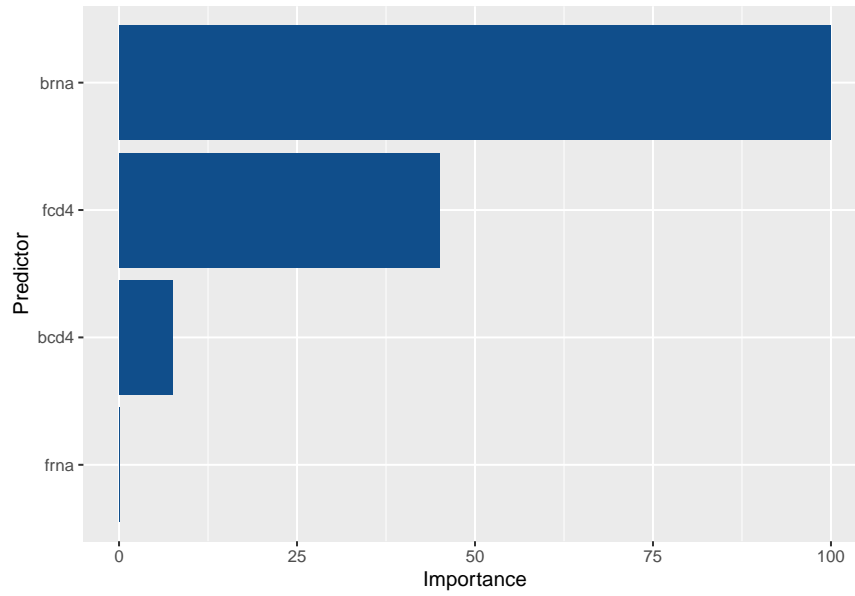


Figure 4: Bar chart of the variable importance of the Rborist model.

- varImp and ggplot2 functions, The baseline RNA load is the most important predictor (fig. 4).

-ensemble, which and tibble functions

```
which_index <- c(which(rborist_dreaction != test_set$dreaction))
# [1] 15 61
tibble(rborist = rborist_dreaction[which_index], rpart = rpart_dreaction[which_index],
       qda = qda_dreaction[which_index], knn = knn_dreaction[which_index],
       test_set = test_set$dreaction[which_index])
## # A tibble: 2 x 5
##   rborist rpart qda   knn   test_set
##   <fct>   <fct> <fct> <fct> <fct>
## 1 hi_tf hi_tf li    li    li
## 2 hi_tf hi_tf li    li    li
```

## Conclusion

- meaningful decisions on antiretroviral therapy administration
- it shall aid physicians on more proactive detection of acute interaction as well as early referrals of patients with failed treatments, for immediate change in treatment episode (Ekpenyong, Etebong, and Jackson 2019, 2)

## References

Ekpenyong, Moses E., Mercy E. Edoho, Ifio J. Udo, Philip I. Etebong, Nseobong P. Uto, Tenderwealth C. Jackson, and Nkem M. Obiakor. 2021a. "A Transfer Learning Approach to Drug Resistance Classification in Mixed HIV Dataset." *Informatics in Medicine Unlocked* 24: 100568. <https://doi.org/10.1016/j.imu.2021.100568>.

- Ekpenyong, Moses E., Philip I. Etebong, and Tenderwealth C. Jackson. 2019. "Fuzzy-Multidimensional Deep Learning for Efficient Prediction of Patient Response to Antiretroviral Therapy." *Heliyon* 5: e02080. <https://doi.org/10.1016/j.heliyon.2019.e02080>.
- Ekpenyong, Moses E., Philip I. Etebong, Tenderwealth C. Jackson, and Edidiong J. Udofa. 2021b. "Processed HIV Prognostic Dataset for Control Experiments." *Data in Brief* 36: 107147. <https://doi.org/10.1016/j.dib.2021.107147>.
- Haugh, Martin. 2017. "IEOR E4525: Machine Learning for OR & FE (Columbia University)." Accessed August 15, 2023. [http://www.columbia.edu/~mh2078/ML\\_ORFE.html](http://www.columbia.edu/~mh2078/ML_ORFE.html).
- Irizarry, Rafael A. 2002. *Introduction to Data Science: Data Analysis and Prediction Algorithms with R*. <http://rafalab.dfci.harvard.edu/dsbook/>.
- Isaakidis, Petros, Marie-Eve Raguenaud, Vantha Te, Chhraing S. Tray, Kazumi Akao, Varun Kumar, Sopheak Ngin, Eric Nerrienet, and Rony Zachariah. 2010. "High Survival and Treatment Success Sustained After Two and Three Years of First-line ART for Children in Cambodia." *Journal of the International AIDS Society* 13: 11. <https://doi.org/10.1186/1758-2652-13-11>.
- Kuhn, Max. 2019. "The caret Package." Last modified March 27, 2019. <https://topepo.github.io/caret/>.
- Revell, A. D., D. Wang, R. Harrigan, R. L. Hamers, A. M. J. Wensing, F. DeWolf, M. Nelson, A.-M. Geretti, and B. A. Larder. 2010. "Modelling Response to HIV Therapy Without a Genotype: An Argument for Viral Load Monitoring in Resource-Limited Settings." *Journal of Antimicrobial Chemotherapy* 65: 605-607. <https://doi.org/10.1093/jac/dkq032>.
- ScienceDirect. 2007. "Recursive Partitioning." Accessed August 18, 2023. <https://www.sciencedirect.com/topics/medicine-and-dentistry/recursive-partitioning>.
- Therneau, Terry, Beth Atkinson, and Brian Ripley. 2022. "Package 'rpart'." Last modified October 21, 2022. <https://cran.r-project.org/web/packages/rpart/rpart.pdf>.
- Wei, Taiyun, and Viliam Simko. 2021. "An Introduction to corrplot Package." Last modified November 18, 2021. <https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html>.
- World Health Organization, n.d. "HIV." Accessed July 21, 2023. [https://www.who.int/health-topics/hiv-aids#tab=tab\\_1](https://www.who.int/health-topics/hiv-aids#tab=tab_1).