2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

Insights on prediction of patients' response to anti-HIV therapies through machine learning

Rogério S. Rosa*, Rafael H. S. Santos*, Ádamo Y. Brito † and Katia S. Guimarães*
*Informatics Center
†Center of Biological Sciences
Federal University of Pernambuco, Recife, Brazil
E-mail: katiag@cin.ufpe.br

***Abstract*—We collect data from the HIV Resistance Drug Database and, based on CD4+ and viral load measures, together with RNA sequences of the reverse transcriptase and of the protease of the virus, we design models using machine learning techniques MultiLayer Perception (MLP), Radial Basis Function (RBF), and Support Vector Machine (SVM), to predict the patient's response to anti-HIV treatment. In this work we applied the SMOTE Algorithm to deal with the enormous difference between the number of case and control samples, which was crucial for the accuracy of the models. Our results show that the SVM model proved more accurate than the other two, with a ROC curve area of 0.9398. We observe that, from 1000 patients, there are 646 samples for which the three methods delivered correct predictions. On the other hand, for 69 patients all three models fail. We analyzed the data for those patients more carefully, and we identified codons and properties that are important for a response/non-response result. Among the codons that our models identified, there are several with strong support from the literature and also a few new ones. Our analysis offers numerous insights that can be very useful to the prediction of patients' response to anti-HIV therapies in the future.**

I. Introduction

The HIV (Human Immunodeficiency Virus) is a retrovirus that attacks the humans CD4+ T cells, causing the decline in their natural defenses against pathogenic microorganisms. Given the high rate of mutation that retroviruses present [1], fighting them is a very difficult task. There are many variants of a same type of HIV in a single individual. This variation is even higher among viral strains from different patients. Such mutations can cause a particular patient not to have a good response to antiretroviral treatment, since the virus installed on your system may have developed resistance to some drugs used to combat the infection. Recognizing HIV mutations that lead to resistance of the virus to certain medications and predicting whether the patient will have a satisfactory response to therapy are challenges that require the use of computational and statistical techniques to be faced in a timely fashion.

Many machine learning methods have been applied in attempts to predict whether a patient will or not have a satisfactory response to HIV cocktail drugs. To this end, properties of the virus RNA sequences of each patient were considered. Neural Network models have been developed considering such characteristics [2], Support Vector Machines models have also been introduced [3] [4]. A comparative study was conducted considering the performance of these tools and human experts [5], which showed evidences that the computational methods are more efficient than human experts. Methods that do not consider the genotype of the HIV have also been presented [6] [7] [8]. Recently, a method based on multilabel classification exploring cross-resistence was presented [9].

In this work, we fitted prediction models and developed an extensive analysis involving the machine learning techniques MultiLayer Perception (MLP), Radial Basis Function (RBF), and Suport Vector Machine (SVM). These approaches were selected because they are well known and they can deliver accurate results when precisely tunned. We applied SMOTE (Synthetic Minority Over-sampling Technique) [10], a synthetic oversampling algorithm, to resolve the unbalance between the case and control cases without causing overfitting. In this paper we answer the following questions: (1) Which of those approaches has the best performance considering accuracy rate?, (2) Are there significant differences among the wrong results of each method?, (3) Are there patterns or properties of the data that induce the approaches to fail?, and (4) Mutations in RNA and clinical measures have the same predictive power on all models?

The remainder of this paper is organized as follows. The next section presents aspects of HIV virus and AIDS (Acquired Immunodeficiency Syndrome) treatment that will be considered in the study. In Section III, the experiments and the data used are described. The results obtained are presented in Section IV. Section V contains a discussion of the results and our concluding remarks.

II. HIV and AIDS treatment

The structure of the HIV is composed of an outer lipoprotein membrane containing specific receptors (gp120 on the surface and gp14 that crosses the membrane), an inner cover which insulates its genetic material, the proteins necessary for virus replication in cell (p11 protease, integrase p31, and reverse transcriptase p51), and two structural proteins (the former of capsid p24 and the former of nucleocapsid p17).

The cycle of HIV is extremely fast. Every day, about 100 copies of the virus are produced. The viral RNA has a high rate of mutations observed. The mutations occur mainly on the reverse transcriptase viral RNA replication step, since the enzyme responsible does not have the control mechanisms required to repair errors in base pairing occurring at this stage. These mutations produce changes in the structure of proteins, causing the virus not to be recognized by the immune system of its host, and making the new subtype resistant to the drugs used.

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

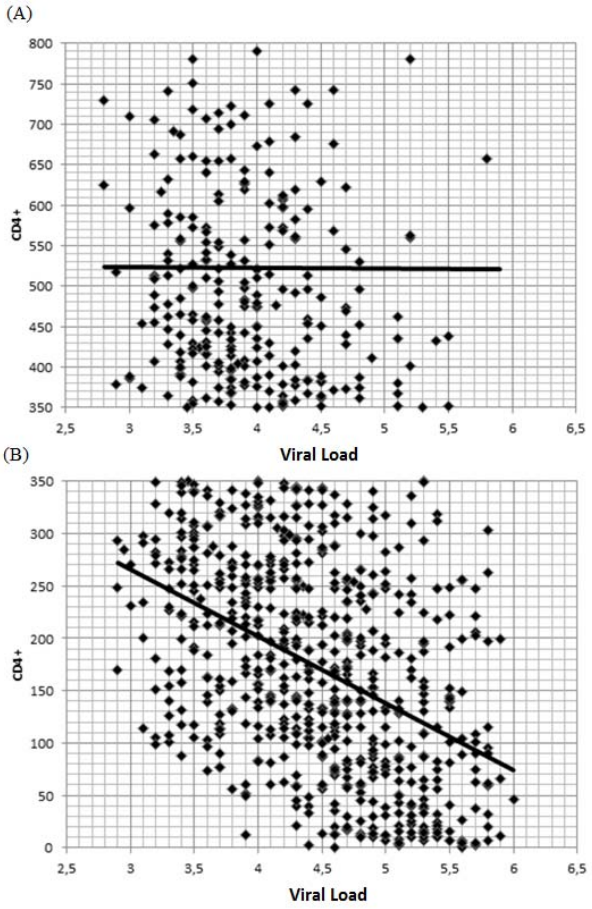

Fig. 1. Correlation between viral load and CD4+ cells count in patients (A) with CD4+ $>$ 350 cells/mL, and (B) with CD4+ $\leq$ 350 cells/mL.

The parameters for evaluation of the patient state are the laboratory tests of rates of CD4+ and Viral Load (VL). According to the values obtained, anti-HIV therapy is indicated or not. In this work we applied the guideline treatment for HIV from the National Institute of Health (NIH) until February 2013, which recommended the monitoring of patients for initiation of therapy based on rate of CD4+ cells, and the patterns were as follows: (Group 1) CD4+ $<$ 350 $cells/mL$; (Group 2) 500 $cells/mL$ $>$ CD4+ $>$ 350 $cells/mL$; and (Group 3): CD4+ $>$ 500 $cells/mL$.

III. Experiments Design

A. Data preparation

The data used in this work were collected from the HIV Resistance Drug Database [11]. It includes 1000 patients with the following attributes: CD4+ at the beginning of the treatment (CD4), viral load at the beginning of treatment (in log base 10) (VL), RNA sequence of the reverse transcriptase of virus (RT), sequence of the protease (PR). We compute the response of the patient to treatment after 16 weeks. It is considered a positive response to the treatment when after 16 weeks there is a reduction of the viral load by a factor of at least 10 (1 in log basis). Patients with positive response are called responders (class 1) and patients with negative response are called non-responders (class 0).

One interesting observation about this data is illustrated in Figure 1, the correlation between the viral load and the CD4+ cells count. In Figure 1 (A) are patients with CD4+ count $>$ 350 at the beginning of the treatment, while in Figure 1 (B) are counts for patients with CD4+ $\leq$ 350. It is expected that the larger the viral load, the least is the CD4+ cell count, since with a larger number of viruses circulating in the body fluids, represents a stronger aggression to the immunological system. Nonetheless, analyzing Figure 1 (A), it is easy to notice the null correlation between those two variables (line parallel to the x axis), as opposed to the strong inverse correlation in Figure 1 (B).

The data presented three problems that should be dealt with before being used to train a classifier: (1) 80 of the 1000 patients do not have the PR sequence available, (2) The RT sequences have different numbers of bases, and (3) The Responders and Non-responders classes are unbalanced, there are 794 samples (79.4%) of class 0 and 206 samples (20.6%) of class 1.

For efficiency purposes, each amino acid was represented by a single natural number between 1 and 22, resulting in approximately 565 entries to the Neural Net. Since RT have different number of nitrogenous bases sequenced in each patient, the codons were translated into peptide sequences, using EMBOSS Transec [12] [13] software, and then they were aligned using the tool called Clustal Omega [14] [13]. The output of Clustal Omega is composed by 26 possible symbols (22 amino acids and 4 special situations, for example, missing data and gaps). The symbols ”underline” and ”gap” used to represent missing data resolved the problems of (1) missing codons on RT and (2) RTs of different lengths. With the special characters inserted by the Clustal Omega, the number of inputs for the neural model was 593. A binary representation of those symbols was also considered in our experiments, but the results did not present any improvement.

The two types of approaches used to solve the problem of unbalanced classes are oversampling (replication of data from smaller class) and undersampling (data reduction of larger class). As the amount of examples for training is not large, discarding some of them would negatively affect the training; for that reason, undersampling was discarded. Simply replicating the data from minority class increases the bias of the classifier for this class, however it makes the model very specific to these replicate cases, damaging the generalization. We used the oversampling synthetic technique called SMOTE (Synthetic Minority Over-Sampling Technique) [10], which creates artificial data based on spatial features between individuals of the minority class, in order to extend the decision region, thus increasing the generalization power of the classifiers generated for these data. SMOTE was applied in order to equal the number of minority class examples to the majority class, resulting in a base with 794 instances (50%) of class 0 and 794 instances (50%) of class 1. The solution of the above-mentioned three problems resulting thus in a database with 1588 (1000 real + 588 synthetic) patients and 596 columns (593 of (PR+RT) + CD4 + VL + response to treatment). The use of the SMOTE technique improved our results considerably. Due to lack of space, the charts indicating such improvements cannot be presented in this paper.

B. Classifiers considered

Three different computational intelligence techniques were used for the design of the classifiers: MultiLayer Perceptron (MLP), Radial Basis Function (RBF), and Support Vector Machine (SVM).

For defining the best MLP configuration, we used multilayer perceptron networks with two intermediate layers, which are able to approximate any continuous function [15]. Different numbers of intermediate layers were considered, but none of them showed better results than two.

Because a very large number of connections can lead the network to memorize the training patterns instead of extracting their general characteristics and provide generalization, several values were tested for the number of nodes in the first intermediate layer(100, 150, 200, 250, and 300), and for the number of nodes in the second intermediate layer (50, 75, 100, 125, and 150). We also varied widely the rate of learning to observe the effect on the error, since a high learning rate makes the training very unstable and prevents the convergence of the learning process. The activation function chosen was hyperbolic tangent.

The learning algorithms tested were: Gradient descent with momentum and adaptive learning rate backpropagation (traingdx), since the momentum accelerates the training process and reduces the possibility of local minimum, whereas an adaptive learning rate attempts to keep the learning step size as large as possible, while keeping learning stable [16]; resilient backpropagation (trainrp), which is fast and uses little memory [16], and one-step secant backpropagation (trainoss), which usually converges faster than conjugate gradient methods [17].

As stopping criterion, we used the maximum number of iterations equal to 1500 or max fail equal to 10, which indicates that if the error in the validation set grows for ten consecutive iterations, the training is stopped to avoid overfitting. The best configuration found for MLP had 250 nodes for the first intermediate layer, 125 nodes for the second intermediate layer, 0.001 for the learning rate and one-step secant backpropagation as learning algorithm.

The RBF has a fixed number of intermediate layers equal to one and number of neurons in the hidden layer is defined at run time, which gradually grows to reach the established mean square error taken as a parameter. Therefore, only one parameter was varied in search for the best possible value: the spread of radial basis functions that are used by each neuron. The values tested in this search were: from 1 (default) to 21 by a factor of 2. These values were chosen because this parameter is directly associated with the smoothness of the approximation function, so that the higher the value, the smoother the function. Thus, a high spread means many neurons will be required to fit a fast-changing function while a low spread means a lot of neurons will be necessary to fit a smooth function and the network may not generalize well [16]. As stopping criterion we used a mean squared error goal equal to 0.05. The RBF starts with no neuron in the hidden layer and, as the inputs are being provided, a new neuron is inserted with center equal to the input vector with the smallest error until it reaches the mean squared error specified as parameter [18].

For SVM, according to Hsu, Chang and Lin [19] there are three main parameters to be varied in search of the best classifier: kernel function, box constraint and rbf sigma. Kernel function is used by SVM to map the training data in the kernel space. Box constraint represents a restriction value for the soft margin and rbf sigma represents a scaling factor in the radial basis function kernel. In search for the best parameter settings, Hsu, Chang, and Lin [19] advise starting with the kernel function "rbf" and default parameters. Then, in order to obtain a better accuracy, different box constraint values were tested (11 values, from $1e^{-5}$ to $1e^{5}$ by a factor of 10.) and rbf sigma (11 values, from $1e^{-5}$ to $1e^{5}$ by a factor of 10).

For training the MLP, the data were divided into three sets: training, testing and validation. Worrying maintaining the proportions of the two classes in the three sets, the division was made as follows: we separated the the instances in accordance with the classes (0 and 1); randomized both groups; each one was divided in ten equal parts, in order to realize a 10-fold cross validation; for each one of the ten runs of the training, one fold class 0 and one class 1 were put together at random to form the test set, while the other nine of each class were united randomly and divided into training set (2/3 of the remaining instances) and validation (1/3 of the remaining instances). This process was repeated 5 times to obtain a number of accuracy rates relevant to the statistical tests. For the training of RBF and SVM was performed similar process of dividing data. The difference is that they were only divided into two sets: training (80%) and test (20%).

For the experiments we used MATLAB R2011b [16], which has a family of toolboxes with functions used to solve various types of problems and represent several processing structures. For the experiments with SVM, we used the Statistics Toolbox, and for MLP and RBF we used the Neural Network Toolbox.

IV. Results

A. Statistical Performance of the Methods

The accuracy of the models created is presented in Table I. According to the results, the SVM Model had the best performance, with 87.11% of correct predictions, while the RBF Model presented 84.34%, and the MLP Model, 71.02%. The distance between MLP and SVM is about 16 percentage points, and between SVM and RBF is 2.77 percentage points. Based on this observation, we can conclude that the MLP Model had the worst performance, because the difference of the MLP results to the other methods was highly significant. SVM and RBF had close percentages of correct predictions, for checking if there is statistical difference between them, an appropriate test was performed, as described next.

To verify the normality of the data and thus identify the most appropriate statistical test to find out if there is statistical evidence that SVM was better than RBF, two Kolmogorov-Smirnov tests [20] were applied, one for each model. For RBF we have $H_0 : X \sim N(\mu = 84.33613; \sigma^2 = 6.227)$ and $H_1 : X \sim other\ distribution$. For SVM we have $H_0 : X \sim N(\mu = 87.11225; \sigma^2 = 6.7943)$ and $H_1 : X \sim other\ distribution$. The values of the test statistic obtained for RBF and SVM were D = 0.1367 and D = 0.1251, respectively. As the critical value for a sample of size 50 and a significance

TABLE I. PERFORMANCE OF MODELS FOR EACH CLASS: (μ) MEAN PERCENTAGE OF CORRECT PREDICTIONS; AND (σ) STANDARD DEVIATION.

Model	Class	μ	σ
MLP	0	71.25 %	6.72
	1	70.80 %	10.66
	Total	71.02 %	5.38
RBF	0	78.56 %	3.42
	1	90.11 %	3.70
	Total	84.34 %	2.49
SVM	0	85.71 %	3.33
	1	88.51 %	3.72
	Total	87.11 %	2.61

level of 0.05 is equal to 0.1923, we do not reject the hypothesis that the data come from a normal distribution.

Given that the data follow a normal distribution, a t-test was performed to check for statistical evidence that the SVM showed better results than the RBF. For $H_0 : \mu_{SVM} = \mu_{RBF}$ and $H_1 : \mu_{SVM} > \mu_{RBF}$, with confidence level of 95%, the test statistic was 5.3853 and p-value 2.489e-07. Thus, as the rejection region for the null hypothesis, considering 98 degrees of freedom, is t $>$ 1.658, we reject the null hypothesis with α = 0.05 and we conclude that the SVM has accuracy rates higher than those obtained by the RBF.

Another evidence of the superiority of SVM can be seen in the ROC curves. Figure 2 illustrates the ROC curves for the three classifiers, considering all patients of 50 folds. The MLP had an area under the curve of 0.7824, while the RBF had a larger area of 0.925, very close, but still smaller than the area obtained by SVM, 0.9398.

The accuracy of models for each class is presented in Table I. According to the results, the SVM and the MLP models showed no significant differences in performance, indicating that the synthetic patients created by the SMOTE were able to extend the decision region of the minority class without increasing the bias of the classifier for that class, and thus increased the generalization power of classifiers. The RBF model was slightly more sensitive to synthetic data. In this model, the minority class showed a mean accuracy 11.55 percentage points higher than the other class, indicating a greater bias towards the minority class than other models, but not considerably affecting the generalization of the network, since it presented high accuracy.

B. Qualitative analysis of results

For each patient, a consensus response was computed for each model, using a simple metric: The majority response of 5 predictions. Figure 3 presents a Venn Diagram of consensus responses predicted incorrectly. From a universe of 1000 patients, 646 patients (called the EASY group) had responses correctly predicted by all models, while for 69 patients (called the HARD group) all models failed in finding the correct response. With 19 exclusive (not shared with any other model) incorrect predictions, SVM was the most accurate model. The model with the largest number of predictions, both exclusive and in general, was MLP (99).

For the remainder of this Section, we focus our attention on identifying which properties in the data would separate the 69 (incorrectly predicted by all three methods) patients in the HARD group from the 646 (correctly predicted by all three methods) individuals in the EASY group.

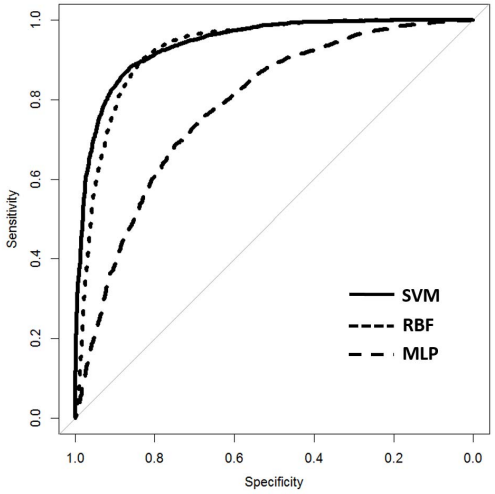

Fig. 2. ROC curves considering results of 50 runs of three Models.

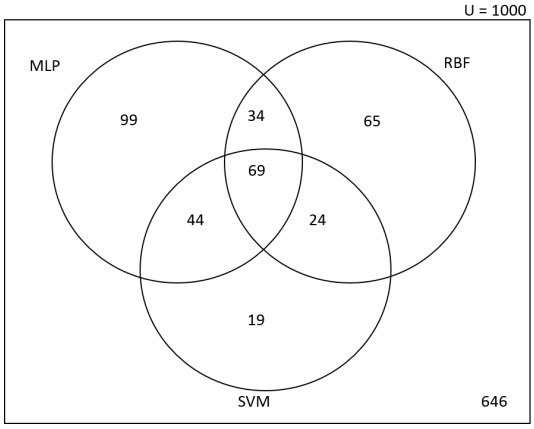

Fig. 3. Venn Diagram of patient's response predicted incorrectly. The response for each patient was considered the consensus class in 5 predictions in each model.

Initially, the VL and CD4+ count were analyzed. Table II presents the average of VL and CD4+ in these two groups. The average VL was very close in the two groups (4.19 in the EASY and 4.86 in the HARD group). We did not find evidences that there is correlation between incorrect inferences and VL. On the other hand, CD4+ count presented significant difference between the two groups. While in the EASY group the CD4+ count average was 293.39, in the HARD group it was 184.88.

TABLE II. THE AVERAGE(STANDARD DEVIATION) OF VIRAL LOAD AND CD4+ FOR EASY AND HARD GROUPS.

Property	EASY group	HARD group
Viral Load	4.19(0.7)	4.86(0.5)
CD4+	293.39(197.08)	184.88(156.12)

We also investigated the distribution of classes in each group. Table III shows the percentage of responders and

non-responders in each group. There is significant difference between the frequency of each class in each group (9.49 percentage points). However, we did not find a linear correlation between CD4+ and occurrence of errors in predictions. We fitted decision trees for identifying the frequency of responders/non-responders according to VL and CD4+. Figure 4 shows the decision tree to CD4+ count, the frequency of patients responders with CD4+ count $\leq$ 151 is 70.4% while in patients with count $>$ 151 is 83.3%. A similar behavior of distribution of responders and non-responders is observed in VL. Figure 5 shows that patients with VL $\leq$ 3.85 had more difficult for responding to treatment (96.9% non-responders and 3.1% responders). The concentration of responders grows with VL, when VL $>$ 4.65 the frequency of responders is 38.9%.

TABLE III. THE DISTRIBUTION OF THE CLASSES IN EASY AND HARD GROUP.

Class	non-responders	responders
EASY group	80.5%	19.5%
HARD group	71.01%	28.99%

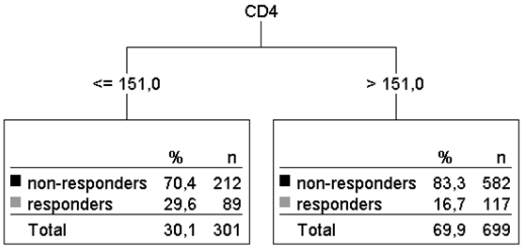

Fig. 4. Decision Tree fitted considering CD4+ count at start of treatment.

Additional information is required for investigating which properties would be able to characterize each patient's group individually. We used decision trees for identifying differences among the RT and PR amino acid sequences in each group.

Figure 6 presents two models: (RESP_NONRESP) the raw data, 1000 patients separated by actual class, where 1 denotes responders and 0 denotes non-responders, and (EASY_HARD) Patients from groups EASY (646 samples) and HARD (69 samples). Figure 7 presents six models: MLP_PRED, RBF_PRED and SVM_PRED were estimated considering the predicted classes by each model, and MLP_CORRECT_PRED, RBF_CORRECT_PRED and SVM_CORRECT_PRED the outputs of each model are classified in correct or incorrect prediction.

Figure 6 (RESP_NONRESP) presents the model where we are looking for the codons that are more relevant for the classification task. According to the tree built, pr10 is the most significant codon for separating responsers from non-responders in general. We did not observe any difference in the frequencies of amino acids on this codon.

The tree Figure 6 (EASY_HARD) models the codon which best separates patients correctly and incorrectly predicted. rt184 was the most important codon for it. The group of patients predicted incorrectly presented two possible amino acids translated in codon rt184: Methionine (M), and Valine (V). The relative frequencies of these amino acids in this group were: (M) 76.82% and (V) 23.18%. The following amino acids occur in this codon in the EASY group: (I) Isoleucine 1.4%, Methionine 39.16% (M), Valine 55.88%(V), and a non standard amino acid 3.56% (X). Codons rt211 and rt296 were indicated as two of the most important codons. The literature contains literature references to codon rt211 [21], but we could not find references to codon rt296.

The trees in Figure 7 (MLP_PRED-MLP_CORRECT_PRED) were fitted for comparing if there are differences on the prediction power of codons, considering predicted classes and correct and incorrect predictions in the MLP model(MLP_PRED-MLP_CORRECT_PRED, respectively). When the tree to consider predicted classes the most important codon is pr63. The most important codon for separating correct and incorrect predictions in the MLP model is pr82. Considering codon pr82 we observed two significant differences between frequencies of amino acids in correct and incorrect predictions. In the EASY group predictions: Alanine 16.16% (A) and missing codons 3.84%. In the HARD group predictions: Alanine 3.27% (A) and missing codons 20.82%. So, there is, indeed, considerable differences regarding this codon in the two scenarios.

The trees in Figure 7 (RBF_PRED-RBF_CORRECT_PRED) were fitted for the RBF model. In the tree that considers predicted classes, the most important codon is rt184. The most important codon for separating correct and incorrect predictions in the RBF model is pr211. Considering codon rt184 we observed two significant differences between frequencies of amino acids in responders and non-responders. In responders: Methionine 59.14% (M) and Valine 36.88% (V). In non-responders: Methionine 41.49% (M) and Valine 53.36% (V).

SVM is the method that presented the best performance, its results are analyzed in Figure 7 (SVM_PRED-SVM_CORRECT_PRED). The most important codon for separating the patients considering the class predicted by SVM is pr10, and for separating its correct and incorrect predictions is rt184. Considering codon rt184 we observed two significant differences between frequencies of amino acids in correct and incorrect predictions. In correct: Methionine 42.30% (M) and Valine 52.37% (V). In incorrect: Methionine 71.15% (M) and Valine 26.92% (V).

The differences detected in the frequencies of the identified amino acids are a strong evidence to validate the binary tree models developed.

V. DISCUSSION AND CONCLUSION

We did an extensive study of different types of genomic data taken from the HIV Resistance Drug Database and its relation to the prediction of patient's response to anti-HIV treatment.

Due to the lack of balance between the case and control samples available, the use of SMOTE technique was central to the success of our analysis. The SMOTE algorithm was applied for resolving the unbalance problem of the data, creating synthetic patients rather than simply replicating a portion of the smaller class. The analysis was based on three well known computational intelligence methods: MLP, RBF, and SVM.

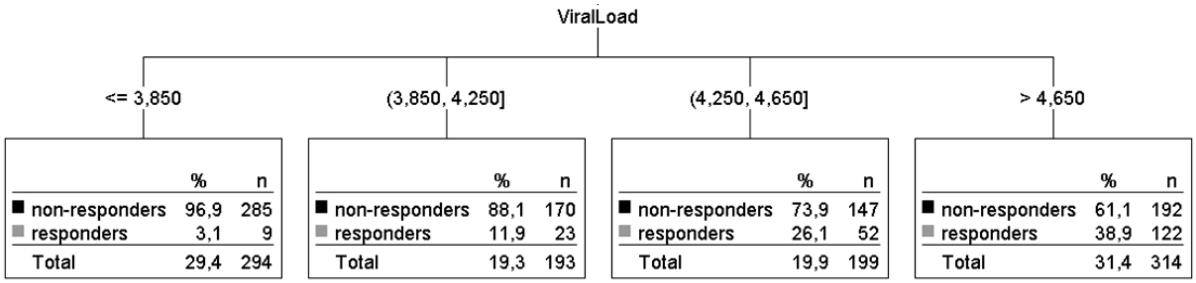

Fig. 5. Decision Tree fitted considering Viral Load at start of treatment.

Fig. 6. Decision Trees estimated considering all codons of RT and PR for: (RESP_NONRESP) 1000 patients and their actual classes: responders and non-responders; and (EASY_HARD) EASY and HARD group.

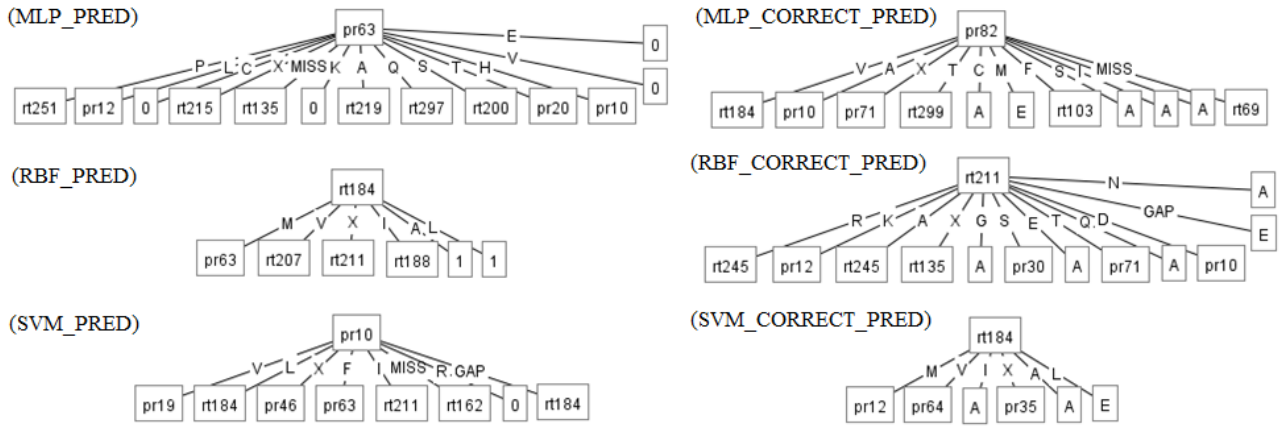

Fig. 7. Decision Trees estimated considering all codons of RT and PR for: predicted classes of all models (MLP_PRED, RBF_PRED and SVM_PRED); and correct and incorrect predictions in each model (MLP_CORRECT_PRED, RBF_CORRECT_PRED and SVM_CORRECT_PRED).

We used statistical tests and ROC analysis for comparison of the accuracy of the methods. Decision trees helped the identification or confirm relationships between: (1) RT/PR codons and predicted classes, (2) RT/PR codons and correct and incorrect predictions, and (3) CD4+/VL and responder and non-responder patients. The results of our analysis offer valuable insights on the behavior of those methods.

We found that the RBF and the SVM models have similar accuracy, although statistically SVM was the best, both being more accurate than MLP by a large margin. Interestingly, RBF is more sensitive than SVM to synthetic samples created by the SMOTE algorithm. While predicted classes by SVM were proportional to frequency of actual classes, RBF was not.

A Venn Diagram was built for analyzing the number of patients' responses predicted incorrectly that are shared by the methods. We identified 646 samples predicted correctly by all methods in all tests, and 69 samples predicted incorrectly in the same situation. SVM had 19 incorrect exclusive predictions, while RBF had 65, and MLP had 99. Based on this observation we can conclude that SVM tends to fail only when the others methods also fail.

We fitted a decision tree for recognizing patterns that separate 69 patients predicted incorrectly from 646 patients correctly predicted. Codon rt184 was identified as the best property for characterizing these two patient groups. Several works have related mutations on this codon to resistance to therapy anti-HIV [22] [23] [24].

Indeed, three decision trees had as root codon rt184 (Figures 6, (RESP_NONRESP), and 7 (RBF_PRED-SVM_CORRECT_PRED)). We computed the relative frequency of amino acids on this codon for each class considered by the trees. Significant differences of frequency of amino acids between the classes of the trees Figure 6 (EASY_HARD) and Figure 7 (SVM_CORRECT_PRED) were found, considering as classes correct and incorrect predictions. Tree Figure 6 (EASY_HARD) was fitted with 69 patients predicted incorrectly from 646 correct for all methods, and tree Figure 7 (SVM_CORRECT_PRED) considers the

TABLE IV. RELATIVE FREQUENCY OF PATIENTS BY CLASS, CD4+ COUNT AND VIRAL LOAD.

CD4+ data	# Patients	Class(%)	CV $\mu(\sigma)$	CD4+ $\mu(\sigma)$
All	1000(100%)	0(79.4)	4.18(0.68)	291.60(193.65)
		1(20.6)	4.80(0.56)	233.50(207.96)
≤ 350	701(70.1%)	0(76.6)	4.32(0.69)	183.11(095.33)
		1(23.4)	4.85(0.54)	155.11(105.51)
> 350	299(29.9%))	0(86.3)	3.85(0.52)	516.36(141.40)
		1(14.7)	4.62(0.60)	548.71(243.69)

correct and incorrect prediction of the SVM model only. An inverse correlation was observed in frequency of Methionine (M) and Valine (V) between classes in both trees. In the tree Figure 6 (EASY_HARD): EASY group (M=39.16% and 55.88%); and HARD group (M=76.82% and V=23.18%). In tree Figure 7 (SVM_CORRECT_PRED): EASY group (M=42.30% and V=52.37%); and HARD group (M=42.30% and 52.37%). Methionine was more abundant in the patients inferred incorrectly. Methionine at rt184 is found in wild-type HIV, which is later replaced by Valine [25].

Tree Figure 6 (EASY_HARD) has rt184 as principal property to separate correct from incorrect predictions. Tree Figure 6 (RESP_NONRESP) has codon pr10, largely mentioned in the literature [26] [27], as the main observed mutation which leads to failure of the treatment of the patients studied. Model SVM had the best performance in the test done, and, curiously, codon rt184 was also identified as the one which may induce SVM to go wrong (Figure 7 (SVM_CORRECT_PRED)). For model SVM, codon pr10 had the most predictive power (Figure 7 (SVM_PRED)), as well as in the tree constructed considering the actual classes (Figure 6 (RESP_NONRESP)). For all that was presented here, we can conclude that codon rt184 needs special treatment in the models, since, although it is associated to the anti-HIV drug resistance, it may also induce the preditors to errors.

Table IV presents the distribution of patients by class. From the patients that started treatment for having (CD4+ $>$ 350), 86.3% did not present a good response to therapy, while 13.7% responded satisfactorily (with considerable reduction in their viral load). That could be explained by a number of reasons. Such patients have average viral load of 3.85, versus 4.62, for the others. Since the immunologic system of those patients were normal and with low viral load, the reason to start the anti-HIV treatment was not to strike out the virus directly, but to prevent an injury of the immune system, or an infection from mother to son, for instance. Since they are healthy, the CD+ and VL values remain unchanged, and they are considered as non-responders.

The goal of the anti-HIV therapy is to reduce the viral load, thus decreasing the aggression to the immune system, enabling it to recover and increase the amount of CD4+ cells. When the viral load is very high it causes damages that are difficult to repair. A very low CD4+ value is evidence that the immune system is highly beaten. Thus, it would be logical to think that the higher the VL value, the greater the chances of the patient not responding to the treatment, as well as patients with low CD4+ measures. But our analysis of the data points in another direction (Figures 4 and 5).

The response to the virus therapy is immediate, since the virus lasts about 48 hours, and the decrease in the VL is used to determine whether the patient had or not a good response to the treatment. On the other hand, the immune response (increased CD4+) is slower. A patient with low CD4+ may have had a 10-fold reduction of VL at 16 weeks, but his CD4+ may never recover due to the high aggression that the immune system may have suffered. This scenario is not analyzed in our study, since the available data are old, from a time that survival was limited, so there is no information regarding for a long period of patient follow up.

Figure 4 shows that the lower the CD4+ (values less than or equal to 151), the greater the chance of viral load reduction (responders); the average VL for responders with CD4+ $\leq$ 151 was 5.029, while for patients with CD4+ $>$ 151 the average reduction was 4.617.

Figure 5 tells us that the higher the viral load, the higher chances of it being reduced. Responders with CD4+ $\leq$ 151 had higher viral loads than those with higher CD4+.

We hypothesize that the greater the amount of virus, the greater the chances of them being reduced at 16 weeks, regardless of the state of the patient's immune system, however, this reduction means just a chance for the immune system to recover, which may or may not happen. We note that in our data there is a considerable proportion of deaths among patients who entered late in therapy (very low CD4+). For example, from 2003 to 2006, Brazil recorded a total of 50.358 patients late in their health system. Of those, 14,457 (28.7%) died within 20 days, and 3,299 (6.55%) had symptoms of developing AIDS [28].

We believe that the insights provided by our analysis can be used for a more effective choice of models, and can also be explored in the design of better approaches for prediction of patients' response to anti-HIV therapies.

REFERENCES

[1] L. M. Mansky, "Retrovirus mutation rates and their role in genetic variation," *Journal of General Virology*, no. 79, p. 13371345, 1998.

[2] B. Larder, D. Wang, A. Revell, J. Montaner, R. Harrigan, F. De Wolf, J. Lange, S. Wegner, L. Ruiz, M. J. Pérez-Elías, S. Emery, J. Gatell, A. D. Monforte, C. Torti, M. Zazzi, and C. Lane, "The development of artificial neural networks to predict virological response to combination HIV therapy." *Antivir Ther*, vol. 12, no. 1, pp. 15–24, 2007.

[3] M. Zazzi, F. Incardona, M. Rosen-Zvi, M. Prosperi, T. Lengauer, A. Altmann, A. Sonnerborg, T. Lavee, E. Schlter, and R. Kaiser, "Predicting response to antiretroviral treatment by machine learning: the EuResist project," *Intervirology*, vol. 55, no. 2, pp. 123–7, 2012.

[4] A. Altmann, M. Däumer, N. Beerenwinkel, Y. Peres, E. Schülter, J. Büch, S.-Y. Rhee, A. Sönnerborg, W. J. Fessel, R. W. Shafer, M. Zazzi, R. Kaiser, and T. Lengauer, "Predicting the response to combination antiretroviral therapy: retrospective validation of geno2pheno-THEO on a large clinical database," *J Infect Dis*, vol. 199, no. 7, pp. 999–1006, 2009.

[5] M. Zazzi, R. Kaiser, A. Sönnerborg, D. Struck, A. Altmann, M. Prosperi, M. Rosen-Zvi, A. Petroczi, Y. Peres, E. Schülter, C. Boucher, F. Brun-Vezinet, P. Harrigan, L. Morris, M. Obermeier, C. Perno, P. Phanuphak, D. Pillay, R. Shafer, A. Vandamme, K. van Laethem, A. Wensing, T. Lengauer, and F. Incardona, "Prediction of response to antiretroviral therapy by human experts and by the euresist data-driven expert system (the eve study)." *HIV Med*, vol. 12, no. 4, pp. 211–8, 2011.

[6] A. D. Revell, D. Wang, R. Wood, C. Morrow, H. Tempelman, R. L. Hamers, G. Alvarez-Uria, A. Streinu-Cercel, L. Ene, A. M. J. Wensing, F. Dewolf, M. Nelson, J. S. Montaner, H. C. Lane, and B. A. a. Larder, "Computational models can predict response to HIV therapy without a

genotype and may reduce treatment failure in different resource-limited settings," *J Antimicrob Chemother*, 2013.

[7] A. D. Revell, D. Wang, R. Harrigan, R. L. Hamers, A. M. J. Wensing, F. Dewolf, M. Nelson, A.-M. Geretti, and B. A. Larder, "Modelling response to hiv therapy without a genotype: an argument for viral load monitoring in resource-limited settings," *J Antimicrob Chemother*, vol. 65, no. 4, pp. 605–7, 2010.

[8] M. C. F. Prosperi, A. A. Michal Rosen-Zvi, S. D. G. Maurizio Zazzi, R. Kaiser, E. Schülter, D. Struck, D. A. v. d. V. Peter Sloot, A.-M. Vandamme, and A. Sönnerborg, "Antiretroviral therapy optimisation without genotype resistance testing: a perspective on treatment history based models." *PLoS One*, vol. 5, no. 10, p. e13753, 2010.

[9] D. Heider, R. Senge, W. Cheng, and E. Hüllermeier, "Multilabel classification for exploiting cross-resistance information in HIV-1 drug resistance prediction," *Bioinformatics*, 2013.

[10] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *J. Artif. Int. Res.*, vol. 16, no. 1, pp. 321–357, 2002.

[11] S.-Y. Rhee, M. J. Gonzales, R. Kantor, B. J. Betts, J. Ravela, and R. W. Shafer, "Human immunodeficiency virus reverse transcriptase and protease sequence database." *Nucleic Acids Research*, vol. 31, no. 1, pp. 298–303, 2003.

[12] P. Rice, I. Longden, and A. Bleasby, "EMBOSS: the european molecular biology open software suite," *Trends Genet*, vol. 16, no. 6, pp. 276–7, 2000.

[13] M. Goujon, H. McWilliam, W. Li, F. Valentin, S. Squizzato, J. Paern, and R. Lopez, "A new bioinformatics analysis tools framework at EMBLEBI," *Nucleic Acids Research*, vol. 38, no. suppl 2, pp. W695–W699, Jul. 2010.

[14] F. Sievers, A. Wilm, D. Dineen, T. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Sding, J. Thompson, and D. Higgins, "Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega." *Mol Syst Biol*, vol. 7, 2011.

[15] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics of Control, Signals, and Systems (MCSS)*, vol. 2, no. 4, pp. 303–314, 1989.

[16] MATLAB, *version 7.13.0.564 (R2011b)*. Natick, Massachusetts: The MathWorks Inc., 2011.

[17] Z. Zakaria, N. A. M. Isa, and S. A. Suandi, "A study on neural network training algorithm for multiface detection in static images," in *International Conference on Computer, Electrical Systems Science, and Engineering*, 2010, p. 170173.

[18] V. S. Abbiramy and A. Tamilarasi, "A comparative study on human spermatozoa images classification with artificial neural network based on fos, glcm and morphological features," in *Advances in Digital Image Processing and Information Technology*, 2011, pp. 220–228.

[19] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A Practical Guide to Support Vector Classification," 2000.

[20] D. Montgomery and G. Runger, *Applied statistics and probability for engineers*, 4th ed. LTC, 2003.

[21] A.-G. Marcelin, P. Flandre, A. Furco, M. Wirden, J.-M. Molina, and V. a. Calvez, "Impact of HIV-1 reverse transcriptase polymorphism at codons 211 and 228 on virological response to didanosine." *Antivir Ther*, vol. 11, no. 6, pp. 693–9, 2006.

[22] R. Kulkarni, K. Babaoglu, E. B. Lansdon, L. Rimsky, V. Van Eygen, G. Picchio, E. Svarovskaia, M. D. Miller, and K. L. White, "The HIV-1 reverse transcriptase m184i mutation enhances the E138K-associated resistance to rilpivirine and decreases viral fitness." *J Acquir Immune Defic Syndr*, 2011.

[23] L. Anta, J. M. Llibre, E. Poveda, J. L. Blanco, M. Alvarez, M. J. Pérez-Elías, A. Aguilera, E. Caballero, V. Soriano, and C. de Mendoza, "Rilpivirine resistance mutations in hiv patients failing non-nucleoside reverse transcriptase inhibitor-based therapies." *AIDS*, 2012.

[24] H.-T. Xu, S. P. Colby-Germinario, E. L. Asahchop, M. Oliveira, M. McCallum, S. M. Schader, Y. Han, Y. Quan, S. G. Sarafianos, and M. A. Wainberg, "The effect of mutations at position e138 in hiv-1 reverse transcriptase and their interactions with the m184i mutation in defining patterns of resistance to the non-nucleoside reverse transcriptase inhibitors rilpivirine and etravirine." *Antimicrob Agents Chemother*, 2013.

[25] S. D. Frost, M. Nijhuis, R. Schuurman, C. A. Boucher, and A. J. Brown, "Evolution of lamivudine resistance in human immunodeficiency virus type 1-infected individuals: the relative roles of drift and selection." *J Virol*, vol. 74, no. 14, pp. 6262–8, 2000.

[26] C. Alteri, A. Artese, G. Beheydt, M. M. Santoro, G. Costa, L. Parrotta, A. Bertoli, C. Gori, N. Orchi, E. Girardi, A. Antinori, S. Alcaro, A. d'Arminio Monforte, K. Theys, A.-M. Vandamme, F. Ceccherini-Silberstein, V. Svicher, and C. F. Perno, "Structural modifications induced by specific HIV-1 protease-compensatory mutations have an impact on the virological response to a first-line lopinavir/ritonavir-containing regimen." *J Antimicrob Chemother*, 2013.

[27] A. Haidara, A. Chamberland, M. Sylla, S. A. Aboubacrine, M. Cissé, H. A. Traore, M. Y. Maiga, A. Tounkara, V. K. Nguyen, and C. L. Tremblay, "Drug Resistance Pathways and Impact of Protease Mutation L10I/V in HIV-1 Non-B Subtypes," *Journal of Antivirals and Antiretrovirals*, vol. 4, no. 2, p. 40, 2012.

[28] A. Grangeiro, M. M. Escuder, P. R. Menezes, R. Alencar, E. A. de Castilho, and L. Myer, "Late Entry into HIV Care: Estimated Impact on AIDS Mortality Rates in Brazil, 2003-2006," *PLOS One*, vol. 6, 2011.