IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 22, 2020, accepted April 24, 2020, date of publication May 4, 2020, date of current version May 20, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2992240

Identification of HIV-1 Vif Protein Attributes Associated With CD4 T Cell Numbers and Viral Loads Using Artificial Intelligence Algorithms

JOSE S. ALTAMIRANO-FLORES[1], SANDRA E. GUERRA-PALOMARES[2], PEDRO G. HERNANDEZ-SANCHEZ[2], JOSE L. RAMIREZ-GARCIALUNA[3], J. RAFAEL ARGÜELLO-ASTORGA[4], DANIEL E. NOYOLA[5], JUAN C. CUEVAS-TELLO[1], (Member, IEEE), AND CHRISTIAN A. GARCÍA-SEPÚLVEDA[2]

[1]Facultad de Ingeniería, Universidad Autónoma de San Luis Potosí, San Luis Potosí 78290, Mexico
[2]Unidad de Genómica Médica del Centro de Investigacion en Ciencias de la Salud y Biomedicina (CICSaB), Facultad de Medicina, Universidad Autónoma de San Luis Potosí, San Luis Potosí 78210, Mexico
[3]Division of Experimental Surgery, Faculty of Medicine, Montréal General Hospital, McGill University, Montreal, QC H3G 1A4, Canada
[4]Centro de Investigación Biomédica, Departamento de Inmunobiología Molecular, Facultad de Medicina, Universidad Autónoma de Coahuila, Torreón 27000, Mexico
[5]Departamento de Microbiología, Facultad de Medicina, Universidad Autónoma de San Luis Potosí, San Luis Potosí 78210, Mexico

Corresponding author: Christian A. García-Sepúlveda (christian.garcia@uaslp.mx)

This work was supported in part by the Mexican Science and Technology Council (CONACYT) under Grant CB-2011-01, #167374, Grant SSA/IMSS/ISSSTE-CONACYT-2009-01-115226, and Grant CONACYT-SALUD-2008-87340. The work of Jose S. Altamirano-Flores was supported by the Mexico's National Science and Technology Council (CONACYT) under Grant #436028.

ABSTRACT The Human Immunodeficiency Virus (HIV) Viral Infectivity Factor (Vif) is a 192-amino acid accessory protein essential to viral replication which counteracts host APOBEC3 proteins. APOBEC3 proteins interfere with the replication of HIV, hepatitis C virus, hepatitis B virus and retrotransposons. Vif is a recent candidate target for therapeutic and preventative interventions in HIV/AIDS yet little is known about its clinical relevance. We describe the results of applying different machine learning algorithms (Apriori, Multifactor Dimensionality Reductor, C4.5, Artificial Neural Networks and ID3) to the search of associations between HIV-1 Vif protein attributes and clinical endpoints. Final iterations showed that the presence of mutations in BC Boxes, APOBEC motifs and Cullin5 binding motifs were together associated with higher initial CD4 T cells while mutations of specific APOBEC motifs coupled with the conservation of other APOBEC motifs were associated with lower historic CD4 T cells. Conservation of specific APOBEC motifs and BC boxes were linked to lower initial viral loads while different combinations of mutations in the Nuclear Localisation Inhibition Signal and BC Boxes were associated with higher historic viral loads. Further scrutiny of these combinations through traditional statistical methods revealed striking differences in both CD4 T cells and viral loads in patients stratified into those having the previous combinations. While artificial intelligence algorithms do not phase out traditional statistical methods, our Artificial Intelligence (AI)-based approach highlights their use at reducing the dimensionality of large and complex datasets and at proposing novel, unimaginable, associations of biological patterns with functional relevance or clinical roles.

INDEX TERMS Artificial intelligence, bioinformatics, genomics, machine learning, medicine.

I. INTRODUCTION

The Human Immunodeficiency Virus (HIV) Viral Infectivity Factor (Vif) is a 192-amino acid (23 kDa) accessory protein essential to viral replication. Vif proteins counteract host proteins exhibiting anti-viral activity of the APOlipoprotein Bmessenger RNA Editing enzyme, Catalytic polypeptide-like (APOBEC3) family. APOBEC3 proteins are zinc-dependent deaminases responsible for nucleic acid editing (mutating cytidine to uridine in both viral DNA and RNA molecules). The APOBEC3 family has seven members (APOBEC3A, APOBEC3B, APOBEC3C, APOBEC3D, APOBEC3F, APOBEC3G and APOBEC3H). APOBEC3 proteins interfere with the

The associate editor coordinating the review of this manuscript and approving it for publication was Hiram Ponce.

replication and propagation of HIV, hepatitis C virus, hepatitis B virus and retrotransposons in humans [1]. APOBEC3G, discovered in 2002, is the best characterised member of the family, it forms stable complexes with the viral core which are ultimately encapsidated into budding virions [2], [3]. APOBEC3G hypermutates HIV DNA during the second round of viral replication leading to non-functional virions. Additional APOBEC3 antiviral activities include plus-strand transfer interference, reverse transcription blockage, as well as inhibition of viral DNA replication and priming, inhibition of viral DNA elongation and inhibition of proviral integration [3]. Vif binding of APOBEC3G recruits elonginB (EloB)-elonginC (EloC)-Cullin5 (Cul5) E3 ligase complex which in turn induces proteasomal degradation of the complex [3]. In addition, Vif blocks APOBEC3G catalytic activity, inhibits APOBEC3G incorporation into budding virions and interferes with APOBEC3G translation [4]. Moreover, recent evidence has demonstrated that certain Vif alleles derived from specific HIV-1 strains can modulate the host cell cycle to induce G2/M cell cycle arrest. While the exact way in which Vif proteins hijack the cell cycle has not been elucidated, both Vif-induced cell cycle arrest and APOBEC4 degradation seem to involve the same Vif functional regions: cullin-5 (CUL5) E3 ubiquitin ligase, elongin B and C as well as the core binding factor beta (CBF-β) [5]–[7]. As such, Vif allows HIV to evade host innate mechanisms that would otherwise protect cells. In recent years, the Vif accessory protein has become a candidate target for both therapeutic and preventative interventions in HIV/AIDS. Nonetheless, little is known about the clinical relevance of Vif protein features and diversity.

Current strategies of exploring the effect that viral polymorphisms have on clinical endpoints rely on either hypothesis-driven techniques (whereby attributes inferred to have functional implications are tested) or on exploratory studies searching for statistical associations which might be indicative of true interactions (which must then be confirmed). When information on the biological role of viral attributes (phenotypic or genotypic in nature) is scarce or unclear, exploratory studies allow novel or interesting associations to be discovered. However, the use of traditional statistical strategies in these exploratory studies (i.e., through contingency tables and χ^2 or Fisher's exact test) involves testing for the effect of numerous attributes which imply the need for statistical corrections for multiple testing. This is particularly important in genome-wide association studies, where the number of variables to be tested give rise to multiple opportunities for spurious associations to arise [8].

Machine learning is a sub-field of AI for use in classification or regression problems very specially adapted to the detection of complex non-lineal interactions in datasets having multiple independent input variables (e.g. attributes) as well as dependant outputs (e.g. clinical endpoint classes) [9], [10]. Some of the most important machine learning algorithms for use in classification problems include Artificial Neural Network (ANN), Support Vector Machine (SVM), Bayesian methods, Decision trees, Apriori and Multifactor Dimensionality Reduction (MDR) algorithms, among others [10]–[18]. ANNs are currently regarded as state-of-the-art algorithms for multi-dimensional dataset explorations and classification of cases. Unfortunately, both ANN and SVM are characterised by a ''black-box'' behaviour in which the underlying patterns of interactions remain invisible and largely unexplorable to the operator. However, Apriori, MDR and Decision trees include mechanisms that help assess how informative the attributes are. In this study we describe our results at applying four different artificial intelligence algorithms to the search of genetic associations among HIV-1 Vif protein attributes and clinical endpoints.

II. MATERIALS AND METHODS

A. STUDY COHORT

Seventy-seven proviral DNA Vif sequences derived from an archived cohort of HIV-infected antiretroviral therapy (ARV)-naive Mexican mestizo patients were included in this study, the protein features of which have been described previously [19]. Patient samples were referred to our laboratory by the state's public HIV/AIDS clinic ''Centro Ambulatorio de Prevención y Atención en SIDA e ITS'' from 2009 to 2014, no RNA samples were available for these patients. CD4 T cell numbers were assessed using a FACScan flow cytometer (Becton, Dickinson and Company, Franklin Lakes, NJ, USA) while HIV viral loads were determined with a COBAS Amplicor HIV-1 Monitor assay (version 1.5 Ultrasensitive, F. Hoffmann-La Roche Ltd. Basel, Switzerland) by the state reference laboratory (Departamento Estatal de Prevención y Control de VIH/SIDA, Servicios de Salud del Estado de San Luis Potosí). An in-depth description of the clinical features of this study cohort is provided in a previous previous publication [20]. Ethics approval for the study was granted by the corresponding Institutional Review Boards (Facultad de Medicina UASLP and the state's public health authority ''Servicios de Salud del Estado de San Luis Potosí'').

B. VIF SEQUENCE ATTRIBUTES INCLUDED

Figure 1 summarises the different protein substitutions present in the 77 sequences, $n = 77$. Analysis of Vif protein substitutions focused on functionally relevant regions and domains known for interacting with other proteins relevant to the biological role of Vif. Substitutions outside of these regions were not considered so as to facilitate interpretation of results and avoid inferences which have not been substantiated through functional or site-directed mutagenesis studies. In Figure 1 Vif protein domains and functionally relevant regions included in our analysis are shown in bold type. These include 17 attributes ($a = 17$): eight APOBEC3-protein binding domains (APOBEC-1 through -8), the nuclear localization inhibitory signal (NLIS), the two Core Binding Factor interaction sites (CBFβ-1 and -2), the three Cullin-5 binding domains (Cul5-1, -2 and -3) and the three Elongin B/C box sites (BCBox-1, -2 and -3). Other sites including the

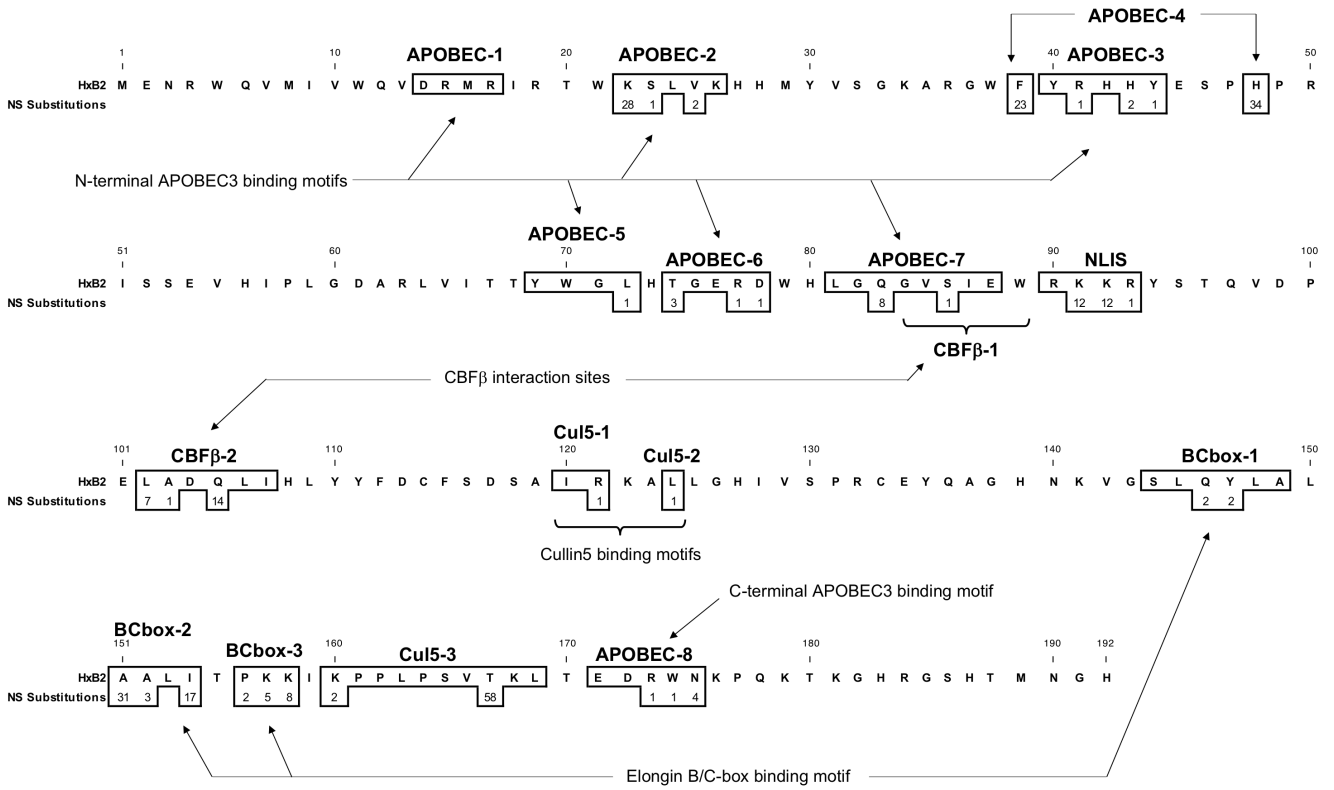

FIGURE 1. Vif protein attributes. Vif protein attributes ($a = 17$) included in the study are shown in bold type and include APOBEC-1 through -8, the nuclear localisation inhibitory signal (NLIS), the Core Binding Factor interaction sites (CBFβ-1 and -2), the Cullin-5 binding domains (Cul5-1, -2 and -3) and the Elongin B/C box sites (BCbox-1, -2 and -3). The number of sequences bearing non-synonymous substitutions at each of these sites is shown below the HXB2 reference sequence. Substitutions observed outside of these functional domains and regions are not shown for clarity.

tryptophans (W) involved in APOBEC3G binding, the MAPK phosphorylation sites, the Zn++-binding motifs, protease processing site, additional phosphorylation sites and the dimerization sites were not included in our analysis. The 17 Vif protein attributes present in each sequence were arbitrarily encoded as either *Conserved* (Cons) or *Mutant* (Mut) after comparing the physico-chemical nature of the substitution to the HXB2 Vif reference sequence. These *Conserved* or *Mutant* attribute status were encoded as 0 and 1 in our working database, respectively. *Conserved* status was assigned when none of the sites within a region had a non-conservative substitution (with regards to HXB2) while *Mutant* status was assigned when at least one non-conservative substitution was present in that region.

C. ATTRIBUTES AND DATABASE COMPILATION

Clinical information for each of the patients included in this study along with their corresponding Vif protein attributes were compiled into a database. The patient's CD4 T cell numbers and viral loads (VL) assessed at the time of initial medical examination are designated herein as initial CD4 and VL. Median values of each patient's CD4 T cell numbers and viral loads assessed on a trimestral basis during the patient's follow-up were calculated (after proving their non-parametric distribution) and designated herein as historic CD4 and VL. Patient derived sequences (S1 through S77 in Table 1) were stratified into <500 or $\geq$ 500 CD4 T cells/μL and $\geq$10,000 or $<$10,000 cp/mL of viral load groups for each of the four different categorical clinical endpoint classes (Initial CD4, historic CD4, initial VL and historic VL) based on established criteria [21]. CD4 and VL classes were encoded as 1 if CD4 T cells $<$500 cells/μL and VL $>$10,000 cp/mL or otherwise as 0 (see Table 1).

D. ARTIFICIAL INTELLIGENCE ALGORITHMS

Here, we introduce our AI-based approach for the identification of the best Vif attribute combinations associated with each of the four clinical classes (see Figure 2). Individual clinical endpoint class databases (Initial CD4, historic CD4, initial VL and historic VL) were screened through three AI algorithms (Apriori, MDR and C4.5) to enhance our identification of Vif attributes repeatedly associated to a clinical class. The Apriori, MDR and C4.5 algorithms identified rules, models or decision trees, respectively [15], [22], [23]. The Apriori algorithm was implemented on the Waikato Environment for Knowledge Analysis (WEKA) workbench v3.6 to generate rules associated with each clinical class [24]. Rules include a body (a string of Vif attributes) associated

TABLE 1. The conserved or mutated state of the 17 Vif protein attributes present in regions of interest (APOBEC-1 through BCbox-3) of the 77 patient (Pt) sequences were encoded as 0 or 1, respectively. Clinical endpoint classes given in far-right columns were encoded as 1 when <500 CD4 T cells/μL or >10,000 cp/mL, or 0 if otherwise.

	Attributes																Classes			
Pt	APOBEC-1	APOBEC-2	APOBEC-3	APOBEC-4	APOBEC-5	APOBEC-6	APOBEC-7	APOBEC-8	CBFβ-1	CBFβ-2	NLIS	Cul5-1	Cul5-2	Cul5-3	BCbox-1	BCbox-2	BCbox-3	Initial CD4	Historic CD4	Initial VL	Historic VL
S1	0	1	0	0	0	0	1	0	0	1	0	0	0	1	0	1	0	1	1	1	0
S2	0	1	0	1	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1	1	1
S3	0	1	0	1	0	0	0	0	0	0	1	0	0	0	0	1	1	0	1	1	1
...	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
S77	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1	0	1	1	0	0

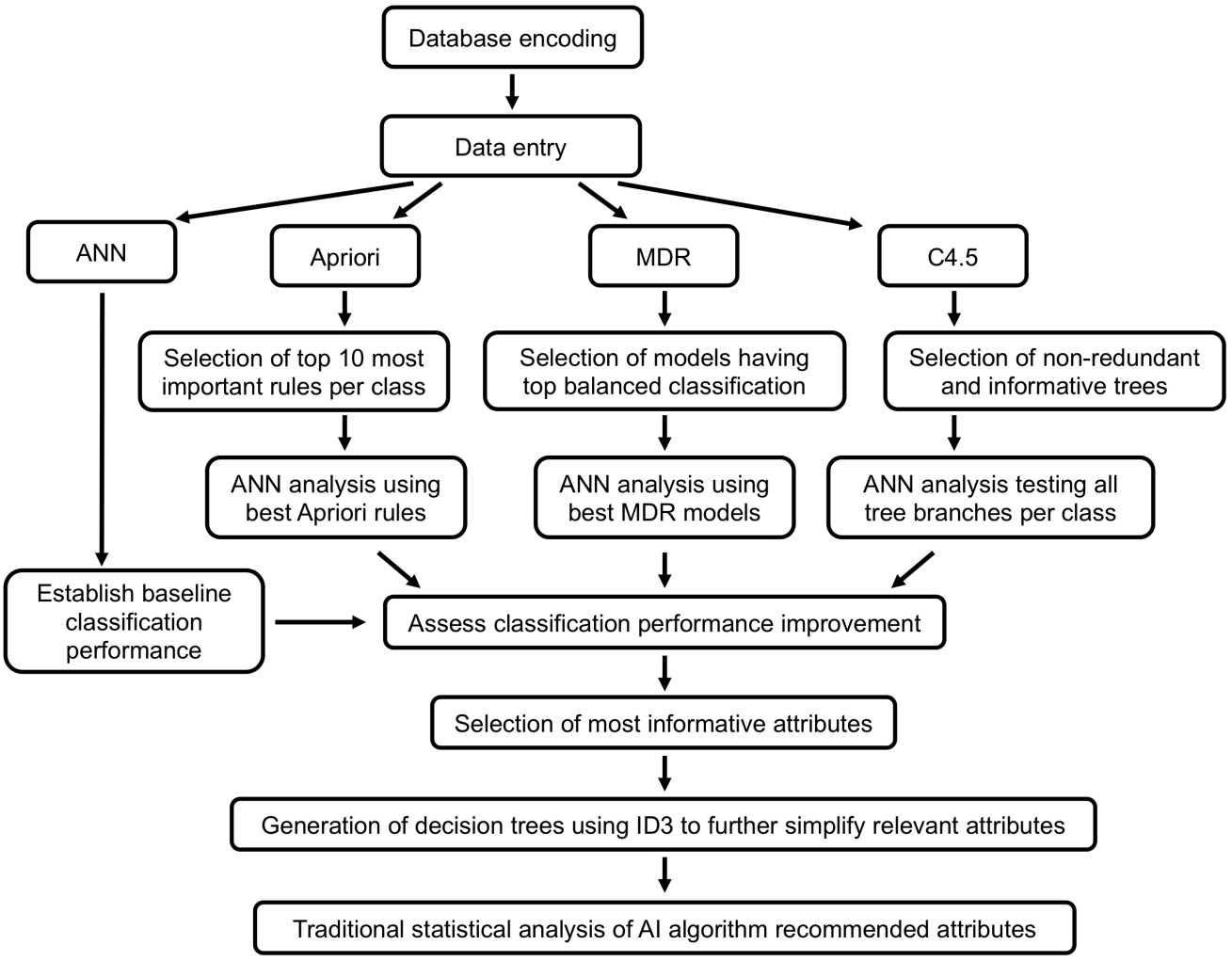

FIGURE 2. AI-based approach. Selection of Vif protein attributes through artificial intelligence algorithms first required establishing a baseline classification using ANN and subsequently selecting for most informative attributes using the Apriori algorithm, MDR and C4.5 to determine which of these improved classification performance of a second-round analysis with ANN. Vif protein attributes selected through this procedure were then used as input for inducing decision trees with ID3 to further select Vif attributes and their status for final testing through traditional statistical tests.

to a head (clinical endpoint class). Apriori is very computationally expensive and is not apt for work with high dimensional datasets. The inclusion of only 17 attributes in Apriori produces around 1.8 million rules. The MDR algorithm detects and allows the user to visualise non-additive combinations and interactions of attributes influencing a clinical class. MDR is currently regarded as a non-parametric model-free alternative to traditional statistical

techniques [25], [26]. MDR is also very computationally expensive and was therefore limited to the generation of only six models having from 1 to a maximum of 6 Vif protein attributes. As model overfitting is common to most AI algorithms, estimation of a model's suitability for generalisation through 10-fold Cross-Validation (CV) was used. Accuracy is a measure of a model's capacity to correctly identify true-positive and true-negative cases against the total number of cases available. However, Balanced Accuracy (BA) is calculated by adding the fractions of correctly identified cases per class divided by the number of classes, and therefore is less affected by data imbalance. The best MDR models were therefore selected on the basis of CV and BA. The C4.5 algorithm used for decision tree induction was also implemented on WEKA, where it is designated J48. This algorithm produces decision trees in which different Vif attributes are hierarchically arranged and related to their clinical endpoint class. Trees are generated by modifying the algorithms pruning parameter (also known as Confidence Level (CF) in 0.01 increments from 0.01 to 0.51, thereby producing 51 possible trees. The hierarchical relevance of each attribute is automatically established based on the information gain ratio during tree induction. Non-redundant and informative trees (i.e. those not having a single branch) were considered for further analysis. ANN (Multi-Layer Perceptron) was also implemented on WEKA and baseline classification accuracy was determined by including all 17 Vif protein attributes and clinical classes during the training process [27]. Once top-ranking rules, models and trees had been generated by the individual AI algorithms (Apriori, MDR and C4.5), ANN classification performance was assessed using the attributes from these as input. Vif protein attributes that were consistently present in the best Apriori-ANN, MDR-ANN and C4.5-ANN results were then selected as input for further processing using the ID3 algorithm. ID3, also implemented on WEKA, generates un-pruned decision trees displaying the hierarchy of attributes and their status (conserved or mutated) as well as their relationship with each clinical class [23], [28]. Finally, Vif protein attribute-status thus identified by AI algorithms were tested through traditional statistical tests (see Figure 2). Full nucleotide and amino acid sequences for each patient were not used directly in our analysis given the computational expense implied and to avert making inferences of polymorphisms which have to date not been shown to be relevant to the biological role of Vif.

E. TRADITIONAL STATISTICAL ANALYSIS

Frequencies of Vif protein mutations were calculated by direct counting of attributes (mutant/conserved sites) and expressed as the percentage of the sequences bearing each. Statistical significance of attribute frequency differences between clinical endpoint groups relied on two-sided Fisher's exact test and binary logistic regression tests for independence using IBM SPSS Statistics (version 21, IBM Corporation, USA). Covariates used in logistic regression analysis included all attributes found statistically significant. Significance was established at $p < 0.05$. Correction for multiple tests employed the Benjamini–Hochberg step-up procedure [29]. Comparison of non-stratified (real) CD4+ T cell numbers and viral loads present in groups having or lacking attributes, attribute status or attribute combinations relied on either t-test with Welsh's correction or Kolmogorov-Smirnov t-test with 2-tailed p values using GraphPad Prism 6 depending on the normality of their distribution (GraphPad Software, Inc. USA).

III. RESULTS

The N-terminal APOBEC3 binding site (^{14}DRMR^{17} in Figure 1) was highly conserved and therefore excluded from subsequent analysis. Premature stop codons prevented two patient sequences from providing information for some Vif attributes. Apriori identified a total of 511,552 rules associated with clinical classes: 135,598 for initial CD4 T cell numbers (129,903 associated with <500 CD4 T cells/μL), 138,062 for historic CD4s (133,151 associated with <500 CD4 T cells/μL), 112,926 for initial VLs (85,255 associated with >10,000 cp/mL) and 124,966 for historic VLs (124,966 associated with >10,000 cp/mL), see example of output shown in Table 4 in Appendix for detailed results.

MDR produced 6 different models having combinations from one to a maximum of six Vif attributes. The three top models associating Vif attributes to initial CD4 T lymphocyte numbers had 6, 2 and 1 attributes, exhibited BAs of 56.45, 56.45 and 59.58 and CV consistencies of 5/10, 6/10 and 8/10, respectively. MDR models for historic CD4 T lymphocyte numbers had 3, 2 and 1 attributes, exhibited BAs of 55.17, 55.84 and 56.66 and CV consistencies of 5/10, 6/10 and 6/10, respectively. For initial VL's these had 5, 3 and 4 attributes, BAs of 57.65, 60.46 and 61.73 and CV consistencies of 5/10, 8/10 and 6/10, respectively. For historic VL's, only two models were considered as the third best did not exceed a 50% BA minimum. The models had only 2 and 1 attributes, BAs of 57.58 and 62.88 and CV consistencies of 7/10 and 10/10, respectively. See Table 5 in Appendix for detailed results.

Of the 51 possible trees generated by C4.5 only four unique trees were identified which associated Vif attributes with initial CD4 class, two trees for historical CD4's, four for initial VL's and two for historic VL's. The remaining trees were either non-informative or redundant, see Table 6 in Appendix for detailed results.

ANN baseline accuracy using the 17 Vif protein attributes for the classification of patients into each of the four clinical classes (initial and historic CD4 T lymphocytes and initial and historic VLs) was of 76.62, 71.43, 64.94 and 55.84, respectively. ANN classification performance exceeded the baseline classification threshold produced by ANN alone in eight Apriori rules, three MDR models and ten C4.5 trees in the initial CD4 T lymphocyte analysis; in 19 rules, three models and 13 trees for the historic CD4 T lymphocytes analysis; in one rule, three models and ten trees for the initial VL analysis as

TABLE 2. Contribution of specific Vif protein attributes to the results produced by each of the four AI algorithms used.

		Apriori		MDR		C4.5				Occurrence in ANN's			Weighed contribution	
		Occurrence in rules		Occurrence in cells		Occurrence in branches		Occurrence in	Average rank					
	Attribute	≥500 cells/μL	<500 cells/μL	≥500 cells/μL	<500 cells/μL	≥500 cells/μL	<500 cells/μL	distinct trees	in trees	Apriori	MDR	C4.5	≥500 cells/μL	<500 cells/μL
Initial CD4 T lymphocytes	APOBEC-2	-	10	-	-	-	-	0/4	-	8/8	0/3	0/10	20	40
	APOBEC-3	8	1	-	-	-	-	0/4	-	1/8	0/3	0/10	19	5
	APOBEC-4	-	10	-	6	1	-	3/4	4	8/8	1/3	5/10	60	122
	APOBEC-5	2	1	-	-	-	-	0/4	-	1/8	0/3	0/10	7	5
	APOBEC-6	1	-	-	-	-	-	0/4	-	0/8	0/3	0/10	2	-
	APOBEC-7	-	-	-	6	-	1	3/4	4	0/8	1/3	8/10	43	97
	APOBEC-8	3	-	-	-	-	-	0/4	-	0/8	0/3	0/10	6	-
	CBFβ-1	1	4	-	-	-	-	0/4	-	3/8	0/3	0/10	10	16
	CBFβ-2	-	-	-	6	-	2	1/4	5	0/8	1/3	1/10	16	75
	NLIS-total	-	-	-	-	-	1	1/4	6	0/8	0/3	0/10	4	10
	Cul5-1	1	4	-	-	-	-	0/4	-	3/8	0/3	0/10	10	16
	Cul5-2	1	4	1	1	-	-	0/4	-	2/8	0/3	0/10	15	21
	Cul5-3	-	-	-	6	-	5	3/4	3	0/8	1/3	4/10	33	111
	BCbox-1	1	-	-	-	-	2	1/4	4	0/8	0/3	1/10	13	23
	BCbox-2	-	-	-	6	-	3	1/4	3	0/8	1/3	1/10	19	85
	BCbox-3	10	-	2	8	1	8	4/4	1	0/8	1/3	10/10	99	169
Historic CD4 T lymphocytes	APOBEC-2	10	10	-	8	-	5	2/2	1	19/19	3/3	13/13	110	204
	APOBEC-3	10	10	-	6	-	2	1/2	3	19/19	2/3	7/13	82	142
	APOBEC-4	-	-	-	-	-	-	0/2	-	0/19	0/3	0/13	-	-
	APOBEC-5	5	-	-	-	-	-	0/2	-	5/19	0/3	0/13	15	5
	APOBEC-6	8	-	-	-	-	2	1/2	4	7/19	0/3	6/13	48	44
	APOBEC-7	-	10	-	-	-	3	2/2	2	10/19	0/3	11/13	55	93
	APOBEC-8	-	-	-	-	-	-	0/2	-	0/19	0/3	0/13	-	-
	CBFβ-1	4	4	-	-	-	-	0/2	-	8/19	0/3	0/13	16	16
	CBFβ-2	-	-	-	-	-	2	1/2	6	0/19	0/3	4/13	18	30
	NLIS-total	-	-	-	-	-	-	2/2	6.5	0/19	0/3	3/13	21	21
	Cul5-1	4	4	-	-	-	-	0/2	-	8/19	0/3	0/13	16	16
	Cul5-2	-	3	-	-	-	-	0/2	-	3/19	0/3	0/13	3	9
	Cul5-3	10	-	-	4	-	-	1/2	8	9/19	1/3	2/13	47	59
	BCbox-1	-	3	-	-	-	-	0/2	-	3/19	0/3	0/13	3	9
	BCbox-2	10	-	-	-	-	1	1/2	7	9/19	0/3	3/13	44	30
	BCbox-3	-	-	-	-	-	3	2/2	4	0/19	0/3	7/13	32	50
		Apriori		MDR		C4.5				Occurrence in ANN's			Weighed contribution	
		Occurrence in rules		Occurrence in cells		Occurrence in branches		Occurrence in	Average rank					
	Attribute	≥10,000 cp/mL	<10,000 cp/mL	≥10,000 cp/mL	<10,000 cp/mL	≥10,000 cp/mL	<10,000 cp/mL	distinct trees	in trees	Apriori	MDR	C4.5	≥10,000 cp/mL	<10,000 cp/mL
Initial viral load	APOBEC-2	10	-	3	13	4	10	4/4	1	1/1	0/3	10/10	138	234
	APOBEC-3	5	10	-	-	-	-	0/4	-	1/1	0/3	0/10	30	40
	APOBEC-4	-	-	-	-	-	-	1/4	6	0/1	0/3	0/10	4	4
	APOBEC-5	2	4	-	-	-	-	0/4	-	1/1	0/3	0/10	24	28
	APOBEC-6	2	2	-	-	-	-	0/4	-	0/1	0/3	0/10	4	4
	APOBEC-7	-	10	2	8	-	-	0/4	-	0/1	0/3	0/10	16	84
	APOBEC-8	-	-	-	-	-	-	0/4	-	0/1	0/3	0/10	-	-
	CBFβ-1	-	2	-	-	-	-	0/4	-	0/1	0/3	0/10	-	4
	CBFβ-2	-	10	-	-	4	6	4/4	2.75	0/1	0/3	6/10	60	92
	NLIS-total	10	-	3	13	-	-	0/4	-	1/1	0/3	0/10	64	124
	Cul5-1	2	2	-	-	-	-	0/4	-	0/1	0/3	0/10	4	4
	Cul5-2	2	1	-	-	-	-	0/4	-	0/1	0/3	0/10	4	2
	Cul5-3	10	-	-	-	4	2	4/4	3.75	1/1	0/3	6/10	99	67
	BCbox-1	10	1	1	4	3	8	4/4	2.75	1/1	0/3	9/10	111	147
	BCbox-2	10	-	3	13	3	2	3/4	4.67	1/1	0/3	3/10	104	158
	BCbox-3	-	10	-	-	-	-	0/4	-	0/1	0/3	0/10	-	20
Historic viral load	APOBEC-2	10	10	-	-	3	-	2/2	4	12/12	0/2	5/11	87	69
	APOBEC-3	4	5	-	-	-	-	0/2	-	1/12	0/2	0/11	10	12
	APOBEC-4	-	10	-	-	-	-	1/2	6	6/12	0/2	1/11	19	39
	APOBEC-5	2	2	-	-	-	-	0/2	-	3/12	0/2	0/11	9	9
	APOBEC-6	-	-	-	-	-	-	0/2	-	0/12	0/2	0/11	-	-
	APOBEC-7	6	-	-	-	-	-	0/2	-	4/12	0/2	0/11	19	7
	APOBEC-8	-	10	-	-	2	-	1/2	3	6/12	0/2	4/11	45	53
	CBFβ-1	4	2	-	-	-	-	0/2	-	4/12	0/2	0/11	15	11
	CBFβ-2	5	-	-	-	-	-	0/2	-	3/12	0/2	0/11	15	5
	NLIS-total	10	-	2	2	5	-	2/2	2	6/12	2/2	9/11	139	89
	Cul5-1	-	2	1	1	-	-	0/2	-	1/12	1/2	0/11	20	24
	Cul5-2	-	2	-	-	-	-	0/2	-	1/12	0/2	0/11	2	6
	Cul5-3	5	-	-	-	-	-	1/2	5	3/12	0/2	2/11	29	19
	BCbox-1	-	10	-	-	5	-	2/2	1	6/12	0/2	11/11	90	80
	BCbox-2	10	-	-	-	2	-	2/2	5	6/12	0/2	5/11	69	37
	BCbox-3	-	-	-	-	-	-	0/2	-	0/12	0/2	0/11	-	-

well as in 12 rules, two models and 11 trees for the historic VL analysis, respectively. The specific contribution of each Vif protein attribute to the results of each of these results is shown in Table 2. Average ANN classification performance using the attributes suggested by the Apriori, MDR and C4.5 algorithms on the initial CD4 T cell analysis was 77.3, 75.8 and 78, respectively. Those for the historic CD4 T cell group were 77.4, 79.2 and 79.3; those of the initial viral load group of 60.8, 60.6 and 65.42 and those of the historic viral load group were of 56.9, 63.6 and 62.5, respectively. Vif attribute occurrence in each of the individual algorithms results as well as their contribution to the ANN classification performance was weighed arbitrarily to compensate for the fact that most rules produced for Apriori were ignored (as only the top 10 for each class were used). As such, the attribute's occurrence was increased 2-fold for those of Apriori, 6-fold for MDR and 8-fold for C4.5 (far-right columns in Table 2). Ultimately, the three attributes with the highest-ranking weighted contribution were selected for inclusion as input attributes for ID3. Only the two highest-ranking attributes in the historic CD4 T lymphocyte group were selected as the third-highest weighted contribution (that of APOBEC-7) was below half of that of the highest (APOBEC-2).

ID3 produced single trees for each of the clinical endpoints using these AI-suggested Vif protein attributes. The tree produced for the initial CD4 T lymphocyte group had three levels, six nodes and seven branches (see Figure 3a). That for the historic CD4 T lymphocyte group had two levels, three nodes and four branches (see Figure 3b), whereas the trees produced for both the initial and historic viral load group had three levels, five nodes and six branches (see Figure 3c and Figure 3d, respectively). Each of the branches (attribute

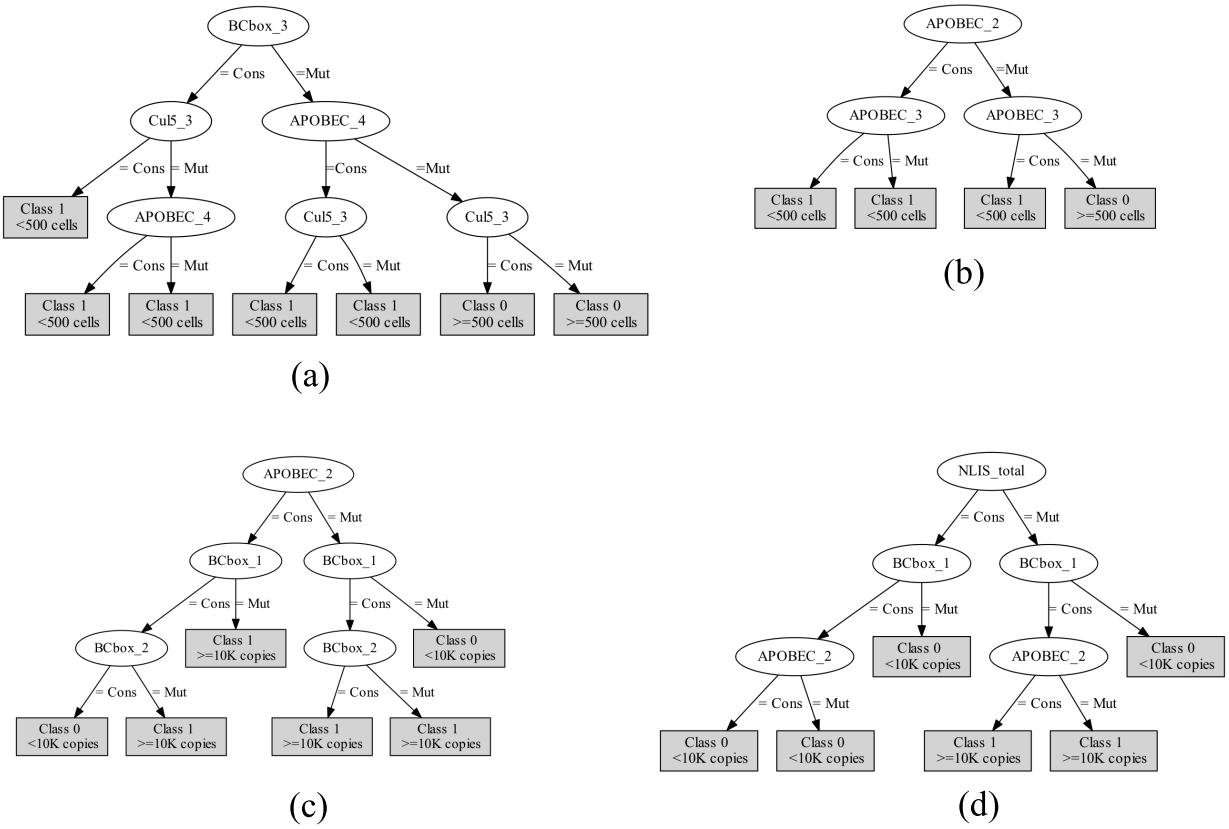

FIGURE 3. ID3 decision trees. Decision trees produced from the final iteration of AI analysis showing the different attributes serving as nodes (ovals) their status (Cons = Conserved, Mut = Mutated) and their relationship with clinical endpoint classes for: (a) the initial CD4 T cell analysis, (b) historic CD4 T cell analysis, (c) initial viral load analysis and (d) historic viral load analysis.

TABLE 3. The most relevant Vif protein attribute combinations associated with clinical endpoints out of the 23 identified by artificial intelligence algorithms. Vif protein regions can either be conserved (Cons) or mutated (Mut) and associated with protection (*prot*) or *risk* to either <500 cells/μL CD4 T cells or ≥10,000 cp/mL of viral load.

Output	Vif attribute combination		Contingency tables		Classification		p-value^{effect}
			≥500 cells/μL	<500 cells/μL	Accuracy	Error	
Initial CD4	BCbox-3^{Mut}, APOBEC-4^{Mut}, Cul5-3^{Cons}	absent	13	59	82.7%	17.3%	0.0083^{prot}
		present	3	0	(62/75)	(13/75)	
	BCbox-3^{Mut}, APOBEC-4^{Mut}, Cul5-3^{Mut}	absent	12	56	80.0%	20.0%	0.0338^{prot}
		present	4	3	(60/75)	(15/75)	
Historic CD4	APOBEC-2^{Mut}, APOBEC-3^{Cons}	absent	14	34	53.3%	46.7%	0.0074^{risk}
		present	1	26	(40/75)	(35/75)	
	APOBEC-2^{Cons}, APOBEC-3^{Cons}	absent	2	29	56%	44.0%	0.0182^{prot}
		present	13	31	(42/75)	(33/75)	
			<10,000 cp/mL	≥10,000 cp/mL			
Initial VL	APOBEC-2^{Cons}, BCbox-1^{Cons}, BCbox-2^{Cons}	absent	15	40	68%	32.0%	0.0318^{prot}
		present	11	9	(51/75)	(24/75)	
	APOBEC-2^{Mut}, BCbox-1^{Cons}, BCbox-2^{Mut}	absent	24	35	50.6%	49.3%	0.0415^{risk}
		present	2	14	(38/75)	(37/75)	
Historic VL	NLIS^{Mut}, BCbox-1^{Cons}, APOBEC-2^{Mut}	absent	41	27	62.6%	37.3%	0.0392^{risk}
		present	1	6	(47/75)	(28/75)	

status combinations) indicated by these trees were then used to manually re-encode our original database to stratify each of the patient's sequences into groups having or not-having these combinations for further traditional statistical analysis. Table 3 summarises those attribute status combinations which proved to be statistically significant for each of the clinical endpoints. Two combinations were found to be significant for the initial CD4 T lymphocyte group, both suggesting a protective effect from having less than 500 CD4 cells/μL. Two combinations were significant for the historic CD4 T

lymphocyte group, one acting as a possible risk factor for having less than 500 CD4 T cells/μL and the other one showing a protective effect. Two combinations resulted significant for initial VLs, the first one protecting from viral loads in excess of 10,000 cp/mL and the second one acting as a risk factor for these high viral titres. The single combination having statistical significance in the historic VL group was found to act as a risk factor for high viral loads.

When the real (un-stratified) CD4 T lymphocyte numbers and viral loads for each of the patients were analysed after grouping them into those having these attribute combinations and those lacking them, striking differences were observed for both initial and historic CD4 T lymphocyte numbers and initial VL but not for historic VL (see Figure 4). Mean initial CD4 T lymphocyte numbers among patients having BCbox-3^{Mut}, APOBEC-4^{Mut}, Cul5-3^{Mut} ($n = 4$) was of 649.8 $\pm$ 35.46 Standard Error of the Mean (SEM) and $\pm$ 70.91 Standard Deviation (SD) cells/μL while those lacking this attribute combination ($n = 56$) had 256.6 $\pm$ 17.73 SEM $\pm$ 135.0 SD cells/μL, $p = 0.0003$ (see Figure 4a). Mean historic CD4 T lymphocyte numbers among patients having APOBEC-2^{Cons}, APOBEC-3^{Cons} ($n = 13$) was of 617.5 $\pm$ 27.28 SEM and $\pm$ 94.52 SD cells/μL while those lacking this attribute combination ($n = 29$) had 335.5 $\pm$ 23.6 SEM $\pm$ 127.1 SD cells/μL, $p < 0.0001$ (see Figure 4b). Likewise, mean initial viral loads among patients having APOBEC-2^{Cons}, BCbox-1^{Cons}, BCbox-2^{Cons} ($n = 11$) were of 1368 $\pm$ 616.8 SEM and $\pm$ 1950 SD cp/mL while those lacking this attribute combination ($n = 40$) had 193,666 $\pm$ 43,341 SEM $\pm$ 274,115 SD cp/mL, $p < 0.0001$ (see Figure 4c, note that is in logarithmic scale for visualization). No Vif protein attribute alone proved to be associated with significant differences in CD4 T cell numbers, nor with VL on either initial or historic groups. For comparison's sake, the effect that each of the 17 Vif attributes had on the four clinical endpoints was tested through traditional statistical methods. The frequency of BCbox-2 mutations was higher among patients having $\geq$500 initial CD4 T cells/μL (81.3%) than in patients having <500 cells/μL (49.2%), $p = 0.025$, suggesting a protective effect. This difference became even more contrasting when all BCbox mutations were considered (BCbox-1 through -3) as a single attribute (93.8% versus 59.3%, respectively, $p = 0.014$). This last effect remained significant after logistic regression (odds ratio OR = 0.097, 95% CI 0.012 - 0.786, $p = 0.029$). With regards to historic CD4 T cells, only APOBEC-2 mutations were found to be detrimental. The frequency of APOBEC-2 mutations was higher among patients with <500 cells/μL (42.6%) in comparison to those having $\geq$500 cells/μL (12.5%), OR = 5.21 (95% CI 1.08 – 25.0, $p = 0.039$). Interestingly, APOBEC-2 mutations were also associated with higher initial viral loads as it was more frequent among patients having $\geq$10,000 cp/mL (44.9%) versus patients having <10,000 cp/mL (21.4%), OR = 2.988 (95% CI 1.031 – 8.657), $p = 0.049$. Mutations of the NLIS were the only attributes associated with greater historic viral loads. Mutated NLIS were present in 48.5% of patients having $\geq$10,000 cp/mL historic VL and in only 23.3% of those having <10,000 cp/mL, OR = 3.106 (95% CI 1.162 – 8.302),

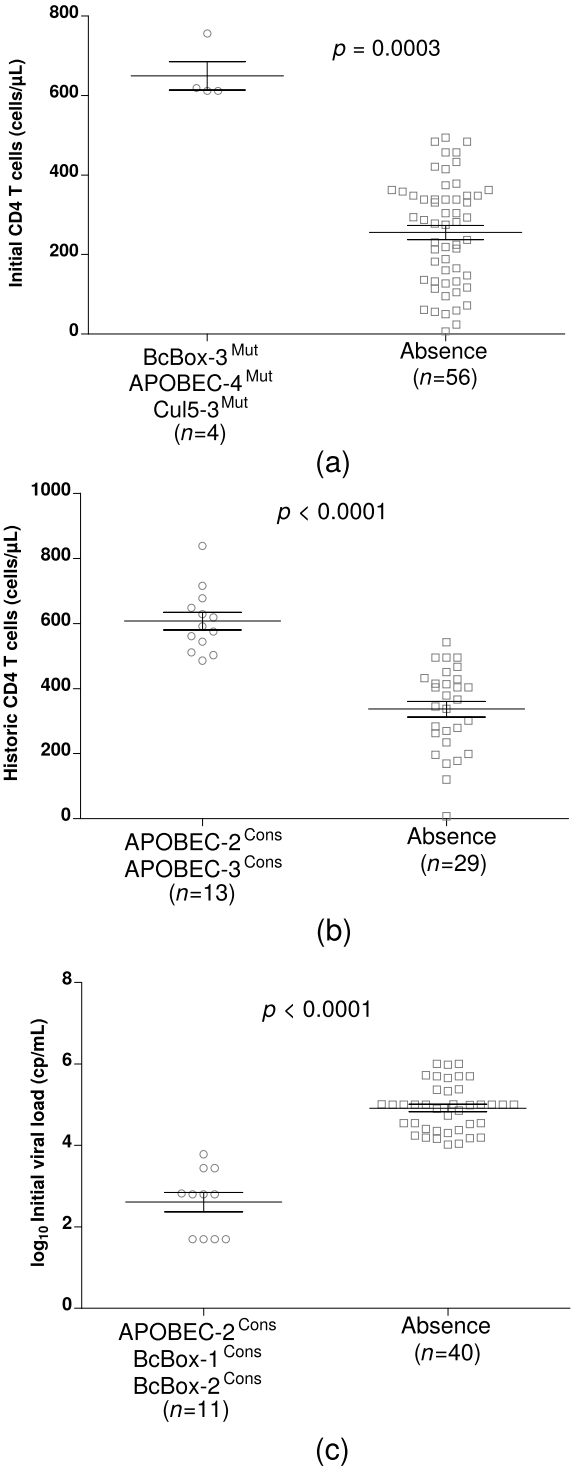

FIGURE 4. Distribution of initial and historic CD4 T lymphocyte numbers and initial viral loads in HIV-1 infected patients having or lacking the different AI suggested Vif protein attribute combinations. Bars depict mean ± standard error of the mean. (a) Initial CD4 T cells. (b) Historic CD4 T cells. (c) Initial viral load – in logarithmic scale.

$p = 0.029$. All of the previously mentioned statistically significant findings lost power after correcting for multiple tests, which lowered the alpha level from 0.05 to 0.003 and 0.002, respectively.

IV. DISCUSSION

Initial evaluations of HIV sequences, such as the screening for antiretroviral drug resistance mutations in a treatment naive patient, are best performed using plasma-derived viral RNA at a time where the patient exhibits high viral loads and before initiating antiretroviral therapy. As of 2013, all Mexican HIV-infected patients are immediately prescribed antiretroviral drugs on diagnosis in accordance with World Health Organisation recommendations and irrespective of their CD4 T cell counts or clinical features (active tuberculosis, hepatitis B infection and/or pregnancy) [30]. Whereas plasma-derived viral RNA sequences provide information on the most replication-competent viral species present at the time of sampling, the use of proviral DNA provides information on archived viral sequences which have been present since the time of HIV integration. As such, the presence of premature stop codons in our Vif sequences highlights the fact that some features may represent archived and genetically defective, replication-incompetent genomes. Nevertheless, proviral DNA has been shown to be an alternative source of viral nucleic acids for molecular studies such as genotyping, genotypic tropism testing, and phylogenetic studies in patients having low to undetectable viral loads [31]–[35].

Our analysis focused on assessing the clinical relevance of Vif substitutions at two distinct clinical phases of HIV infection, 1) at the time of initial medical evaluation, a point in which patients had not yet been subjected to antiretroviral therapy, and 2) during follow-up and after being exposed to the effect of antiretroviral drugs. Our identification of Vif substitutions associated with clinical variables at time of initial evaluation suggests that Vif polymorphism might prove to have an effect on the progression of unchecked HIV infections. On the other hand, the identification of associations further-down in the medical follow-up of patients suggests that these effects might still be present in spite of current antiretroviral therapy. We developed an AI-based attribute combination discovery approach which when combined with traditional statistical methods is capable of identifying associations between sequence traits and clinical endpoints in HIV/AIDS. Our results highlight the capacity of AI algorithms to guide traditional statistical methods for the study of the biological role and clinical relevance of factors for which hypothesis-driven techniques would be otherwise unsuccessful or laborious. Our AI-based approach is a complex analysis pipeline which allows us to identify Vif attributes repeatedly identified by different AI algorithms as important, so as to further enhance the selection of those attributes that would later be explored through traditional statistics. Examination of the results generated by individual AI algorithms (see Tables 4–6 in Appendix) provides evidence that no single algorithm was capable of identifying all attributes found to be significant on the final iteration. In addition, this application demonstrates the way AI algorithms can condense data with little human intervention. To our knowledge, this represents the first report associating complex multi-dimensional combinations of Vif protein attributes with two of the most important clinical follow-up parameters in HIV/AIDS: CD4 T lymphocyte numbers and viral loads. Strict adherence to the principles of testing and correction for this simple dataset comprising 77 combinations of 17 attributes and 4 different outputs would have limited the statistical power of most findings. The results produced by AI algorithms alone were congruent with those produced through traditional methods. BCbox mutations were associated with high initial CD4 T cells in univariate analysis but were also part of the final AI attribute combination associated with these. These results are in agreement with those published previously describing the epistatic effects of some pairs of amino acids encompassing the BCbox regions of Vif proteins with low CD4 T cell counts [36]. The importance of BCbox attributes in Vif's function highlighted in our results is in agreement with previous findings regarding its functional role. Vif hijacks the E3 ligase using the BCbox region that interacts with ElonginC and a zinc finger motif that interacts with Cullin5. Vif recognition and binding of APOBEC3 through Cul5 involves forming a complex with EloB and EloC, which in turn recruits CBFβ. The interaction between Vif and EloC is mediated by the ^{144}SLQ(Y/F)LA^{149} motif (BCbox-1) present in the viral Elongin B/C-box [37], [38]. This domain is perhaps the most critical Vif region determining APOBEC3 protein suppression. Previous reports have shown that the short side chain of Ala^{149} plays a crucial role in EloC-binding. As both the Vif-induced G2/M arrest and APOBEC3G degradation effects involve interactions with virtually the same host ubiquitin ligase machinery including Cul5, EloB and EloC as well as CBFβ, the exact biological role through which Vif exerts the observed effects in our study cohort can not be ascertained. Nevertheless, Vif's capacity to induce G2/M arrest has been observed in HIV viruses derived from clinical samples and this capacity has been shown to be associated with increased viral replication in vitro T cells cultures [6], [39].

Similarly, APOBEC-2 mutations were associated with the risk of low historic CD4 T cells and high initial viral loads in univariate analysis but also formed part of the AI combinations associated with the risk of low historic CD4 T cells and high historic viral loads. Interestingly, the conservation of APOBEC-2 was shown to be associated with lower initial viral loads by AI algorithms. That APOBEC3 binding sites are among the most repeatedly encountered Vif protein regions associated with both CD4 T cell numbers and VL is not surprising. Different motifs are used selectively by Vif to bind different APOBEC3 family members [37]. For Vif to exert its action it must bind APOBEC3 to subsequently act as the substrate binding subunit of a cullin RING ligase-5 (CRL5) E3 ligase complex [37]. Previous authors have demonstrated that the single most important factor

TABLE 4. Top 10 rules produced by the Apriori algorithm associating Vif protein attributes and their status (conserved or mutated) with clinical classes.

Class	Vif protein attribute and status	n (%)	Confidence
Initial CD4 T lymphocytes <500 cells/μL	APOBEC-2^{Mut} APOBEC-4^{Cons}	15 (20%)	1
	APOBEC-2^{Mut} APOBEC-4^{Cons} CBFβ-1^{Cons}	15 (20%)	1
	APOBEC-2^{Mut} APOBEC-4^{Cons} Cul5-1^{Cons}	15 (20%)	1
	APOBEC-2^{Mut} APOBEC-4^{Cons} Cul5-2^{Cons}	15 (20%)	1
	APOBEC-2^{Mut} APOBEC-4^{Cons} CBFβ-1^{Cons} Cul5-1^{Cons}	15 (20%)	1
	APOBEC-2^{Mut} APOBEC-4^{Cons} CBFβ-1^{Cons} Cul5-2^{Cons}	15 (20%)	1
	APOBEC-2^{Mut} APOBEC-4^{Cons} Cul5-1^{Cons} Cul5-2^{Cons}	15 (20%)	1
	APOBEC-2^{Mut} APOBEC-4^{Cons} CBFβ-1^{Cons} Cul5-1^{Cons} Cul5-2^{Cons}	15 (20%)	1
	APOBEC-2^{Mut} APOBEC-3^{Cons} APOBEC-4^{Cons}	14 (18.7%)	1
	APOBEC-2^{Mut} APOBEC-4^{Cons} APOBEC-5^{Cons}	14 (18.7%)	1
Initial CD4 T lymphocytes $\geq$500 cells/μL	APOBEC-3^{Cons} BCbox-3^{Mut}	8 (10.7%)	0.57
	APOBEC-8^{Cons} BCbox-3^{Mut}	8 (10.7%)	0.57
	APOBEC-3^{Cons} APOBEC-5^{Cons} BCbox-3^{Mut}	8 (10.7%)	0.57
	APOBEC-3^{Cons} APOBEC-6^{Cons} BCbox-3^{Mut}	8 (10.7%)	0.57
	APOBEC-3^{Cons} APOBEC-8^{Cons} BCbox-3^{Mut}	8 (10.7%)	0.57
	APOBEC-3^{Cons} CBFβ-1^{Cons} BCbox-3^{Mut}	8 (10.7%)	0.57
	APOBEC-3^{Cons} Cul5-1^{Cons} BCbox-3^{Mut}	8 (10.7%)	0.57
	APOBEC-3^{Cons} Cul5-2^{Cons} BCbox-3^{Mut}	8 (10.7%)	0.57
	APOBEC-3^{Cons} BCbox-1^{Cons} BCbox-3^{Mut}	8 (10.7%)	0.57
	APOBEC-5^{Cons} APOBEC-8^{Cons} BCbox-3^{Mut}	8 (10.7%)	0.57
Historic CD4 T lymphocytes <500 cells/μL	APOBEC-2^{Mut} APOBEC-3^{Cons} APOBEC-7^{Cons} CBFβ-1^{Cons}	21 (28%)	1
	APOBEC-2^{Mut} APOBEC-3^{Cons} APOBEC-7^{Cons} Cul5-1^{Cons}	21 (28%)	1
	APOBEC-2^{Mut} APOBEC-3^{Cons} APOBEC-7^{Cons} Cul5-2^{Cons}	21 (28%)	1
	APOBEC-2^{Mut} APOBEC-3^{Cons} APOBEC-7^{Cons} BCbox-1^{Cons}	21 (28%)	1
	APOBEC-2^{Mut} APOBEC-3^{Cons} APOBEC-7^{Cons} CBFβ-1^{Cons} Cul5-1^{Cons}	21 (28%)	1
	APOBEC-2^{Mut} APOBEC-3^{Cons} APOBEC-7^{Cons} CBFβ-1^{Cons} Cul5-2^{Cons}	21 (28%)	1
	APOBEC-2^{Mut} APOBEC-3^{Cons} APOBEC-7^{Cons} CBFβ-1^{Cons} BCbox-1^{Cons}	21 (28%)	1
	APOBEC-2^{Mut} APOBEC-3^{Cons} APOBEC-7^{Cons} Cul5-1^{Cons} Cul5-2^{Cons}	21 (28%)	1
	APOBEC-2^{Mut} APOBEC-3^{Cons} APOBEC-7^{Cons} Cul5-1^{Cons} BCbox-1^{Cons}	21 (28%)	1
	APOBEC-2^{Mut} APOBEC-3^{Cons} APOBEC-7^{Cons} CBFβ-1^{Cons}	21 (28%)	1
Historic CD4 T lymphocytes $\geq$500 cells/μL	APOBEC-2^{Cons} APOBEC-3^{Cons} APOBEC-6^{Cons} Cul5-3^{Mut} BCbox-2^{Mut}	8 (10.7%)	0.47
	APOBEC-2^{Cons} APOBEC-3^{Cons} APOBEC-5^{Cons} APOBEC-6^{Cons} Cul5-3^{Mut} BCbox-2^{Mut}	8 (10.7%)	0.47
	APOBEC-2^{Cons} APOBEC-3^{Cons} APOBEC-6^{Cons} CBFβ-1^{Cons} Cul5-3^{Mut} BCbox-2^{Mut}	8 (10.7%)	0.47
	APOBEC-2^{Cons} APOBEC-3^{Cons} APOBEC-6^{Cons} Cul5-1^{Cons} Cul5-3^{Mut} BCbox-2^{Mut}	8 (10.7%)	0.47
	APOBEC-2^{Cons} APOBEC-3^{Cons} APOBEC-5^{Cons} APOBEC-6^{Cons} CBFβ-1^{Cons} Cul5-3^{Mut} BCbox-2^{Mut}	8 (10.7%)	0.47
	APOBEC-2^{Cons} APOBEC-3^{Cons} APOBEC-5^{Cons} APOBEC-6^{Cons} Cul5-1^{Cons} Cul5-3^{Mut} BCbox-2^{Mut}	8 (10.7%)	0.47
	APOBEC-2^{Cons} APOBEC-3^{Cons} APOBEC-6^{Cons} CBFβ-1^{Cons} Cul5-1^{Cons} Cul5-3^{Mut} BCbox-2^{Mut}	8 (10.7%)	0.47
	APOBEC-2^{Cons} APOBEC-3^{Cons} APOBEC-5^{Cons} APOBEC-6^{Cons} CBFβ-1^{Cons} Cul5-1^{Cons} Cul5-3^{Mut} BCbox-2^{Mut}	8 (10.7%)	0.47
	APOBEC-2^{Cons} APOBEC-3^{Cons} Cul5-3^{Mut} BCbox-2^{Mut}	8 (10.7%)	0.44
	APOBEC-2^{Cons} APOBEC-3^{Cons} APOBEC-5^{Cons} Cul5-3^{Mut} BCbox-2^{Mut}	8 (10.7%)	0.44
Initial viral load <10,000 cp/mL	APOBEC-3^{Cons} APOBEC-7^{Cons} CBFβ-2^{Mut} BCbox-3^{Cons}	8 (10.7%)	1
	APOBEC-3^{Cons} APOBEC-5^{Cons} APOBEC-7^{Cons} CBFβ-2^{Mut} BCbox-3^{Cons}	8 (10.7%)	1
	APOBEC-3^{Cons} APOBEC-6^{Cons} APOBEC-7^{Cons} CBFβ-2^{Mut} BCbox-3^{Cons}	8 (10.7%)	1
	APOBEC-3^{Cons} APOBEC-7^{Cons} CBFβ-1^{Cons} CBFβ-2^{Mut} BCbox-3^{Cons}	8 (10.7%)	1
	APOBEC-3^{Cons} APOBEC-7^{Cons} CBFβ-2^{Mut} Cul5-1^{Cons} BCbox-3^{Cons}	8 (10.7%)	1
	APOBEC-3^{Cons} APOBEC-7^{Cons} CBFβ-2^{Mut} Cul5-2^{Cons} BCbox-3^{Cons}	8 (10.7%)	1
	APOBEC-3^{Cons} APOBEC-7^{Cons} CBFβ-2^{Mut} BCbox-1^{Cons} BCbox-3^{Cons}	8 (10.7%)	1
	APOBEC-3^{Cons} APOBEC-5^{Cons} APOBEC-6^{Cons} APOBEC-7^{Cons} CBFβ-2^{Mut} BCbox-3^{Cons}	8 (10.7%)	1
	APOBEC-3^{Cons} APOBEC-5^{Cons} APOBEC-7^{Cons} CBFβ-1^{Cons} CBFβ-2^{Mut} BCbox-3^{Cons}	8 (10.7%)	1
	APOBEC-3^{Cons} APOBEC-5^{Cons} APOBEC-7^{Cons} CBFβ-2^{Mut} Cul5-1^{Cons} BCbox-3^{Cons}	8 (10.7%)	1
Initial viral load $\geq$10,000 cp/mL	APOBEC-2^{Cons} NLIS-total^{Cons} Cul5-3^{Mut} BCbox-1^{Cons} BCbox-2^{Cons}	10 (13.3%)	0.77
	APOBEC-2^{Cons} APOBEC-3^{Cons} NLIS-total^{Cons} Cul5-3^{Mut} BCbox-1^{Cons} BCbox-2^{Cons}	10 (13.3%)	0.77
	APOBEC-2^{Cons} APOBEC-5^{Cons} NLIS-total^{Cons} Cul5-3^{Mut} BCbox-1^{Cons} BCbox-2^{Cons}	10 (13.3%)	0.77
	APOBEC-2^{Cons} APOBEC-6^{Cons} NLIS-total^{Cons} Cul5-3^{Mut} BCbox-1^{Cons} BCbox-2^{Cons}	10 (13.3%)	0.77
	APOBEC-2^{Cons} NLIS-total^{Cons} Cul5-1^{Cons} Cul5-3^{Mut} BCbox-1^{Cons} BCbox-2^{Cons}	10 (13.3%)	0.77
	APOBEC-2^{Cons} NLIS-total^{Cons} Cul5-2^{Cons} Cul5-3^{Mut} BCbox-1^{Cons} BCbox-2^{Cons}	10 (13.3%)	0.77
	APOBEC-2^{Cons} APOBEC-3^{Cons} APOBEC-5^{Cons} NLIS-total^{Cons} Cul5-3^{Mut} BCbox-1^{Cons} BCbox-2^{Cons}	10 (13.3%)	0.77
	APOBEC-2^{Cons} APOBEC-3^{Cons} APOBEC-6^{Cons} NLIS-total^{Cons} Cul5-3^{Mut} BCbox-1^{Cons} BCbox-2^{Cons}	10 (13.3%)	0.77
	APOBEC-2^{Cons} APOBEC-3^{Cons} NLIS-total^{Cons} Cul5-1^{Cons} Cul5-3^{Mut} BCbox-1^{Cons} BCbox-2^{Cons}	10 (13.3%)	0.77
	APOBEC-2^{Cons} APOBEC-3^{Cons} NLIS-total^{Cons} Cul5-2^{Cons} Cul5-3^{Mut} BCbox-1^{Cons} BCbox-2^{Cons}	10 (13.3%)	0.77
Historic viral load <10,000 cp/mL	APOBEC-2^{Cons} APOBEC-7^{Cons} CBFβ-2^{Cons} NLIS-total^{Cons} BCbox-2^{Cons}	11 (14.7%)	0.85
	APOBEC-2^{Cons} APOBEC-7^{Cons} NLIS-total^{Cons} Cul5-3^{Mut} BCbox-2^{Cons}	11 (14.7%)	0.85
	APOBEC-2^{Cons} CBFβ-1=^{Cons} CBFβ-2^{Cons} NLIS-total^{Cons} BCbox-2^{Cons}	11 (14.7%)	0.85
	APOBEC-2^{Cons} CBFβ-1=^{Cons} NLIS-total^{Cons} Cul5-3^{Mut} BCbox-2^{Cons}	11 (14.7%)	0.85
	APOBEC-2^{Cons} APOBEC-3^{Cons} APOBEC-7^{Cons} CBFβ-2^{Cons} NLIS-total^{Cons} BCbox-2^{Cons}	11 (14.7%)	0.85
	APOBEC-2^{Cons} APOBEC-3^{Cons} APOBEC-7^{Cons} NLIS-total^{Cons} Cul5-3^{Mut} BCbox-2^{Cons}	11 (14.7%)	0.85
	APOBEC-2^{Cons} APOBEC-3^{Cons} CBFβ-1^{Cons} CBFβ-2^{Cons} NLIS-total^{Cons} BCbox-2^{Cons}	11 (14.7%)	0.85
	APOBEC-2^{Cons} APOBEC-3^{Cons} CBFβ-1^{Cons} NLIS-total^{Cons} Cul5-3^{Mut} BCbox-2^{Cons}	11 (14.7%)	0.85
	APOBEC-2^{Cons} APOBEC-5^{Cons} APOBEC-7^{Cons} CBFβ-2^{Cons} NLIS-total^{Cons} BCbox-2^{Cons}	11 (14.7%)	0.85

TABLE 4. *(Continued.)* **Top 10 rules produced by the Apriori algorithm associating Vif protein attributes and their status (conserved or mutated) with clinical classes.**

Class	Vif protein attribute and status	n (%)	Confidence
Historic viral load $\geq$ 10,000 cp/mL	APOBEC-2^{Mut} APOBEC-4^{Mut} APOBEC-8^{Cons} BCbox-1^{Cons}	8 (10.7%)	0.73
	APOBEC-2^{Mut} APOBEC-3^{Cons} APOBEC-4^{Mut} APOBEC-8^{Cons} BCbox-1^{Cons}	8 (10.7%)	0.73
	APOBEC-2^{Mut} APOBEC-4^{Mut} APOBEC-5^{Cons} APOBEC-8^{Cons} BCbox-1^{Cons}	8 (10.7%)	0.73
	APOBEC-2^{Mut} APOBEC-4^{Mut} APOBEC-8^{Cons} CBFβ-1^{Cons} BCbox-1^{Cons}	8 (10.7%)	0.73
	APOBEC-2^{Mut} APOBEC-4^{Mut} APOBEC-8^{Cons} Cul5-1^{Cons} BCbox-1^{Cons}	8 (10.7%)	0.73
	APOBEC-2^{Mut} APOBEC-4^{Mut} APOBEC-8^{Cons} Cul5-2^{Cons} BCbox-1^{Cons}	8 (10.7%)	0.73
	APOBEC-2^{Mut} APOBEC-3^{Cons} APOBEC-4^{Mut} APOBEC-5^{Cons} APOBEC-8^{Cons} BCbox-1^{Cons}	8 (10.7%)	0.73
	APOBEC-2^{Mut} APOBEC-3^{Cons} APOBEC-4^{Mut} APOBEC-8^{Cons} CBFβ-1^{Cons} BCbox-1^{Cons}	8 (10.7%)	0.73
	APOBEC-2^{Mut} APOBEC-3^{Cons} APOBEC-4^{Mut} APOBEC-8^{Cons} Cul5-1^{Cons} BCbox-1^{Cons}	8 (10.7%)	0.73
	APOBEC-2^{Mut} APOBEC-3^{Cons} APOBEC-4^{Mut} APOBEC-8^{Cons} Cul5-2^{Cons} BCbox-1^{Cons}	8 (10.7%)	0.73

TABLE 5. **Best models produced by MDR associating Vif protein attributes with clinical classes.**

Class	Model	Training data BA	Test data BA	CV consistency
Initial CD4 T lymphocytes	BCbox-3	69.62	59.58	8/10
	Cul5-2, BCbox-3	73	56.45	6/10
	Cul5-3, BCbox-2, BCbox-3	76.54	52.05	5/10
	APOBEC-4, Cul5-3, BCbox-2, BCbox-3	80.34	54.66	4/10
	APOBEC-7, CBFβ-2, Cul5-3, BCbox-2, BCbox-3	84.85	56.3	6/10
	APOBEC-4, APOBEC-7, CBFβ-2, Cul5-3, BCbox-2, BCbox-3	87.25	56.45	5/10
Historic CD4 T lymphocytes	APOBEC-2	65.38	56.66	6/10
	APOBEC-2, APOBEC-3	71.02	55.84	6/10
	APOBEC-2, APOBEC-3, Cul5-3	75.36	55.17	5/10
	APOBEC-2, CBFβ-2, Cul5-3, BCbox-2	79.92	44.98	3/10
	APOBEC-2, APOBEC-4, Cul5-3, BCbox-2, BCbox-3	84.63	47.75	6/10
	APOBEC-2, APOBEC-7, CBFβ-2, Cul5-3, BCbox-2, BCbox-3	87.95	52.51	3/10
Initial viral load	APOBEC-2	62.39	47.19	7/10
	APOBEC-2, BCbox-1	67.66	53.32	8/10
	APOBEC-2, NLIS-total, BCbox-2	71.88	60.46	8/10
	APOBEC-2, APOBEC-7, NLIS-total, BCbox-2	75.85	61.73	6/10
	APOBEC-2, APOBEC-7, NLIS-total, BCbox-1, BCbox-2	78.71	57.65	5/10
	APOBEC-2, APOBEC-4, NLIS-total, Cul5-3, BCbox-1, BCbox-2	81.07	52.55	4/10
Historic viral load	NLIS-total	62.88	62.88	10/10
	NLIS-total, Cul5-1	65.7	57.58	7/10
	APOBEC-7, NLIS-total, BCbox-3	68.77	50	7/10
	APOBEC-2, APOBEC-7, APOBEC-8, CBFβ-2	71.93	45.08	3/10
	APOBEC-2, APOBEC-7, APOBEC-8, NLIS-total, BCbox-2	74.96	44.7	3/10
	APOBEC-2, APOBEC-4, CBFβ-2, NLIS-total, BCbox-2, BCbox-3	77.95	46.97	4/10

governing Vif functionality, and therefore clinical relevance, corresponds to the APOBEC3 binding regions. The importance of APOBEC-2 motif ^{22}KSLVK^{26} was first established in a cohort of Brazilian Brazilian treatment-naïve patients, where the K22H mutation was shown to be associated with lower CD4 T cells and higher viral loads [40]. With regards to the relevance of the APOBEC-4 motif, previous studies have also highlighted the importance of positions 39 and 48 for Vif to counteract the influence of APOBEC3H proteins [41], [42]. It is also worth noting that both APOBEC-2 and -3 motifs, both of which were determined to be clinically relevant in our study, are located in the N-terminal region of Vif protein and in sites that have been previously shown to be under positive selection whereas the C-terminal APOBEC-8 motif did not seem to be associated with clinical classes [36]. Contrastingly, NLIS status was the only attribute exhibiting contradicting associations between traditional and AI methods. Artificial intelligence is the capability for machines to imitate intelligent human behaviour once trained through mathematical and statistical techniques to enable prediction of previously unseen patterns without having been explicitly programmed to do so. The ability of AI to analyse datasets and detect patterns in an n-dimensional feature space provides them with the capability of suggesting attribute combinations which would seem intractable to humans. This capacity has been illustrated in our results and further substantiated through traditional statistical techniques. While AI does not completely phase out traditional statistical methods, the AI-based approach proposed herein highlights their use at reducing the dimensionality of large and complex datasets and at proposing novel, unimaginable, associations of biological patterns with functional relevance or clinical roles. Determining the overall generalisation of AI applications to real-world patient management is critical to the development of a truly successful implementation strategy.

V. CONCLUSIONS

This paper proposes an AI-based approach, which was shown to be capable of identifying associations between HIV

TABLE 6. Top-rated decision trees produced by the C4.5 algorithm associating Vif protein attributes and their status (conserved or mutated) with clinical classes.

Class		CF Range	Accuracy	Leaves	Tree size
Initial CD4 T lymphocytes	Uninformative trees	0.01-0.02	79.22	1	1
		0.03-0.09	77.92	1	1
	First unique tree	0.1-0.13	77.92	3	5
	Second unique tree	0.14-0.14	77.92	4	7
		0.15-0.24	79.22	4	7
	Third unique tree	0.25-0.25	79.22	5	9
		0.26-0.5	80.52	5	9
	Fourth unique tree	0.51-0.51	76.62	11	21
Historic CD4 T lymphocytes	Uninformative trees	0.01-0.15	79.22	1	1
		0.16-0.25	77.92	1	1
		0.26-0.31	76.62	1	1
		0.32-0.36	75.32	1	1
		0.37-0.38	74.03	1	1
	First unique tree	0.39-0.47	74.03	5	9
		0.48-0.5	72.73	5	9
	Second unique tree	0.51-0.51	72.73	11	21
Initial viral load	Uninformative trees	0.01-0.04	61.04	1	1
		0.05-0.06	58.44	1	1
		0.07-0.14	55.84	1	1
		0.15-0.17	57.14	1	1
	First unique tree	0.18-0.21	57.14	6	11
		0.22-0.24	58.44	6	11
	Second unique tree	0.25-0.25	58.44	6	11
	Third unique tree	0.26-0.27	59.74	7	13
		0.28-0.5	61.04	7	13
	Fourth unique tree	0.51-0.51	62.34	8	15
Historic viral load	First unique tree	0.01-0.15	57.14	2	3
		0.16-0.24	58.44	2	3
		0.25-0.25	57.14	2	3
		0.26-0.28	55.84	2	3
		0.29-0.29	53.25	2	3
	Second unique tree	0.3-0.3	53.25	5	9
		0.31-0.31	54.55	5	9
		0.32-0.32	55.84	5	9

sequence traits and clinical endpoints in AIDS. These results highlight the capacity of AI algorithms at guiding traditional statistical methods in the search for novel interactions of virus and host genes and proteins in the absence of hypothesis-driven techniques. To the best of our knowledge, this represents the first report describing such novel and complex multi-dimensional associations of Vif protein attributes with CD4 T lymphocyte numbers and viral loads in HIV/AIDS. While this study focused on the genetically distinct Mexican mestizo human population, we envisage that future applications of this AI-based approach are very likely to prove beneficial and informative for other HIV-infected human groups. These results open the possibility of incorporating the study of novel genetic marker combinations into routine clinical management algorithms currently in use, further complementing the molecular arsenal of tools available for people living with HIV and AIDS.

Although most of our findings have been previously independently reported, the way in which their combinations interact and modulate clinical endpoints has not been previously explored. While aware of the limitations imposed by the use of proviral DNA and by the size of our dataset, our sequential approach employing both artificial intelligence algorithms along with traditional statistical methods was capable of identifying Vif sequence attributes associated with clinical endpoints. As is well known, the low number of patients enrolled for this pilot study underpowers the clinical significance of the discovered associations. This does not, however, undermine the need to further investigate these findings in larger and even different study cohorts. Our discovery supports the notion that similar approaches have great utility at guiding conventional association-discovery approaches in biomedical sciences.

APPENDIX
ADDITIONAL TABLES AND FIGURES

Tables 4–6 correspond to supplementary material which complements Tables 2–3 and Figures 3–4, see Section III. Table 4 summarises the results obtained through the use of the Apriori algorithm, Table 5 summarises those obtained by MDR and Table 6 those corresponding to the C4.5 algorithm.

ACKNOWLEDGMENT

The authors wish to thank the Universidad Autónoma de San Luis Potosí for all support as well as the patients, nursing, and medical personnel at Centro Ambulatorio para la Prevención y Atención del SIDA e Infecciones de Transmisión Sexual (CAPASITS), San Luis Potosí, Mexico, for making this study possible.

REFERENCES

[1] A. Moris, S. Murray, and S. Cardinaud, "AID and APOBECs span the gap between innate and adaptive immunity," *Frontiers Microbiol.*, vol. 5, p. 534, Oct. 2014, doi: 10.3389/fmicb.2014.00534.

[2] A. M. Sheehy, N. C. Gaddis, J. D. Choi, and M. H. Malim, "Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral vif protein," *Nature*, vol. 418, no. 6898, pp. 646–650, Jul. 2002, doi: 10.1038/nature00939.

[3] V. B. Soros and W. C. Greene, "APOBEC3g and HIV-1: Strike and counterstrike," *Current HIV/AIDS Rep.*, vol. 4, no. 1, pp. 3–9, Feb. 2007, doi: 10.1007s11904-007-0001-1.

[4] M. Santa-Marta, P. M. de Brito, A. Godinho-Santos, and J. Goncalves, "Host factors and HIV-1 replication: Clinical evidence and potential therapeutic approaches," *Frontiers Immunol.*, vol. 4, p. 343, Oct. 2013, doi: 10.3389fimmu.2013.00343.

[5] J. L. DeHart, A. Bosque, R. S. Harris, and V. Planelles, "Human immunodeficiency virus type 1 Vif induces cell cycle delay via recruitment of the same E3 ubiquitin ligase complex that targets APOBEC3 proteins for degradation," *J. Virol.*, vol. 82, no. 18, pp. 9265–9272, Jul. 2008, doi: 10.1128/jvi.00377-08.

[6] T. Izumi, K. Io, M. Matsui, K. Shirakawa, M. Shinohara, Y. Nagai, M. Kawahara, M. Kobayashi, H. Kondoh, N. Misawa, Y. Koyanagi, T. Uchiyama, and A. Takaori-Kondo, "HIV-1 viral infectivity factor interacts with TP53 to induce G2 cell cycle arrest and positively regulate viral replication," *Proc. Nat. Acad. Sci. USA*, vol. 107, no. 48, pp. 20798–20803, Nov. 2010, doi: 10.1073/pnas.1008076107.

[7] E. L. Evans, J. T. Becker, S. L. Fricke, K. Patel, and N. M. Sherer, "HIV-1 Vif's capacity to manipulate the cell cycle is species specific," *J. Virol.*, vol. 92, no. 7, Mar. 2018, Art. no. e02102. [Online]. Available: https://jvi.asm.org/content/92/7/e02102-17

[8] W. S. Noble, "How does multiple testing correction work?" *Nature Biotechnol.*, vol. 27, no. 12, pp. 1135–1137, Dec. 2009, doi: 10.1038/nbt1209-1135.

[9] A. L. Beam, A. Motsinger-Reif, and J. Doyle, "Bayesian neural networks for detecting epistasis in genetic association studies," *BMC Bioinf.*, vol. 15, no. 1, p. 368, Nov. 2014, doi: 10.1186/s12859-014-0368-0.

[10] R. Jiang, W. Tang, X. Wu, and W. Fu, "A random forest approach to the detection of epistatic interactions in case-control studies," *BMC Bioinf.*, vol. 10, no. S1, p. S65, Jan. 2009, doi: 10.1186/1471-2105-10-s1-s65.

[11] K. J. Cios and N. Liu, "A machine learning method for generation of a neural network architecture: A continuous ID3 algorithm," *IEEE Trans. Neural Netw.*, vol. 3, no. 2, pp. 280–291, Mar. 1992, doi: 10.1109/72.125869.

[12] J. C. Cuevas Tello, D. Hernández-Ramírez, and C. A. García-Sepúlveda, "Support vector machine algorithms in the search of KIR gene associations with disease," *Comput. Biol. Med.*, vol. 43, no. 12, pp. 2053–2062, Dec. 2013, doi: 10.1016/j.compbiomed.2013.09.027.

[13] J. Forsström, P. Nuutila, and K. Irjala, "Using the ID3 algorithm to find discrepant diagnoses from laboratory databases of thyroid patients," *Med. Decis. Making*, vol. 11, no. 3, pp. 171–175, Aug. 1991, doi: 10.1177/0272989x9101100305.

[14] B. Han, X. wen Chen, Z. Talebizadeh, and H. Xu, "Genetic studies of complex human diseases: Characterizing SNP-disease associations using Bayesian networks," *BMC Syst. Biol.*, vol. 6, no. 3, p. S14, 2012, doi: 10.1186/1752-0509-6-s3-s14.

[15] M. D. Ritchie, L. W. Hahn, N. Roodi, L. R. Bailey, W. D. Dupont, F. F. Parl, and J. H. Moore, "Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer," *Amer. J. Hum. Genet.*, vol. 69, no. 1, pp. 138–147, Jul. 2001, doi: 10.1086/321276.

[16] M. Somek and M. Hercigonja-Szekeres, "Decision support systems in health care–velocity of Apriori algorithm," *Stud Health Technol Inf.*, vol. 244, pp. 53–57, Oct. 2017.

[17] D. L. Tong, D. J. Boocock, G. K. R. Dhondalay, C. Lemetre, and G. R. Ball, "Artificial neural network inference (ANNI): A study on gene-gene interaction for biomarkers in childhood sarcomas," *PLoS ONE*, vol. 9, no. 7, Jul. 2014, Art. no. e102483, doi: 10.1371/journal.pone.0102483.

[18] D. L. Tong and A. C. Schierz, "Hybrid genetic algorithm-neural network: Feature extraction for unpreprocessed microarray data," *Artif. Intell. Med.*, vol. 53, no. 1, pp. 47–56, Sep. 2011, doi: 10.1016/j.artmed.2011.06.008.

[19] S. E. Guerra-Palomares, P. G. Hernandez-Sanchez, M. A. Esparza-Perez, J. R. Arguello, D. E. Noyola, and C. A. Garcia-Sepulveda, "Molecular characterization of mexican HIV-1 vif sequences," *AIDS Res. Hum. Retroviruses*, vol. 32, no. 3, pp. 290–295, Mar. 2016. [Online]. Available: http://online.liebertpub.com/doi/10.1089/aid.2015.0290

[20] D. Hernández-Ramírez, M. A. Esparza-Pérez, J. L. Ramirez-Garcialuna, J. R. Arguello, P. B. Mandeville, D. E. Noyola, and C. A. García-Sepúlveda, "Association of KIR3dl1/s1 and HLA-bw4 with CD4 T cell counts in HIV-infected mexican mestizos," *Immunogenetics*, vol. 67, no. 8, pp. 413–424, Jun. 2015, doi: 10.1007/s00251-015-0848-z.

[21] Y. Jiang, O. Chen, C. Cui, B. Zhao, X. Han, Z. Zhang, J. Liu, J. Xu, Q. Hu, C. Liao, and H. Shang, "KIR3ds1/l1 and HLA-bw4-80i are associated with HIV disease progression among HIV typical progressors and long-term nonprogressors," *BMC Infectious Diseases*, vol. 13, no. 1, Sep. 2013, doi: 10.1186/1471-2334-13-405.

[22] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," *ACM SIGMOD Rec.*, vol. 22, no. 2, pp. 207–216, Jun. 1993, doi: 10.1145/170036.170072.

[23] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo, CA, USA: Morgan Kaufmann, 1993.

[24] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *SIGKDD Explor*, vol. 11, no. 1, pp. 10–18, 2009.

[25] L. W. Hahn, M. D. Ritchie, and J. H. Moore, "Multifactor dimensionality reduction software for detecting gene–gene and gene–environment interactions," *Bioinformatics*, vol. 19, no. 3, pp. 376–382, Feb. 2003, doi: 10.1093/bioinformatics/btf869.

[26] A. Jakulin and I. Bratko, "Analyzing attribute dependencies," in *Knowledge Discovery in Databases—PKDD*. Berlin, Germany: Springer, 2003, pp. 229–240, doi: 10.1007/978-3-540-39804-2_22.

[27] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986. [Online]. Available: http://www.nature.com/doifinder/10.1038/323533a0

[28] J. R. Quinlan, "Learning efficient classification procedures and their application to chess end games," *Mach. Learn. Artif. Intell. Approach*, pp. 463–482, 1984.

[29] Y. Benjamini, D. Drai, G. Elmer, N. Kafkafi, and I. Golani, "Controlling the false discovery rate in behavior genetics research," *Behavioural Brain Res.*, vol. 125, nos. 1–2, pp. 279–284, Nov. 2001. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0166432801002972

[30] *Consolidated Guidelines Use Antiretroviral Drugs for Treating Preventing HIV Infection*, WHO, Geneva, Switzerland, 2013.

[31] L. P. R. Vandekerckhove, A. M. J. Wensing, R. Kaiser, F. Brun-Vézinet, B. Clotet, A. De Luca, S. Dressler, "European guidelines on the clinical management of HIV-1 tropism testing," *Lancet Infectious Diseases*, vol. 11, no. 5, pp. 394–407, May 2011, doi: 10.1016/S1473-3099(10)70319-4.

[32] C. Soulié, S. Fourati, S. Lambert-Niclot, I. Malet, M. Wirden, R. Tubiana, M.-A. Valantin, C. Katlama, V. Calvez, and A.-G. Marcelin, "Factors associated with proviral DNA HIV-1 tropism in antiretroviral therapy-treated patients with fully suppressed plasma HIV viral load: Implications for the clinical use of CCR5 antagonists," *J. Antimicrobial Chemotherapy*, vol. 65, no. 4, pp. 749–751, Feb. 2010, doi: 10.1093/jac/dkq029.

[33] P. Frange, J. Galimand, C. Goujard, C. Deveau, J. Ghosn, C. Rouzioux, L. Meyer, M.-L. Chaix, "High frequency of x4/DM-tropic viruses in PBMC samples from patients with primary HIV-1 subtype-b infection in 1996–2007: The French ANRS CO06 PRIMO cohort study," *J. Antimicrobial Chemotherapy*, vol. 64, no. 1, pp. 135–141, May 2009, doi: 10.1093/jac/dkp151.

[34] C. Verhofstede, D. Brudney, J. Reynaerts, D. Vaira, K. Fransen, A. De Bel, C. Seguin-Devaux, S. De Wit, L. Vandekerckhove, and A.-M. Geretti, "Concordance between HIV-1 genotypic coreceptor tropism predictions based on plasma RNA and proviral DNA," *HIV Med.*, vol. 12, no. 9, pp. 544–552, Apr. 2011, doi: 10.1111/j.1468-1293.2011.00922.x.

[35] K. Huruy, A. Mulu, U. G. Liebert, and M. Melanie, "HIV-1c proviral DNA for detection of drug resistance mutations," *PLOS ONE*, vol. 13, no. 10, Oct. 2018, Art. no. e0205119, doi: 10.1371/journal.pone.0205119.

[36] M. C. Bizinoto, S. Yabe, É. Leal, H. Kishino, L. de O. Martins, M. L. de Lima, E. R. Morais, R. S. Diaz, and L. M. Janini, "Codon pairs of the HIV-1 vif gene correlate with CD4+ T cell count," *BMC Infectious Diseases*, vol. 13, no. 1, p. 173, Apr. 2013, doi: 10.1186/1471-2334-13-173.

[37] Y. Feng, T. T. Baig, R. P. Love, and L. Chelico, "Suppression of APOBEC3-mediated restriction of HIV-1 by Vif," *Frontiers Microbiol.*, vol. 5, p. 450, Aug. 2014, doi: 10.3389/fmicb.2014.00450.

[38] Y. Yu, Z. Xiao, E. S. Ehrlich1, X. Yu, and X.-F. Yu, "Selective assembly of HIV-1 vif-cul5-ElonginB-ElonginC e3 ubiquitin ligase complex through a novel SOCS box and upstream cysteines," *Genes Develop.*, vol. 18, no. 23, pp. 2867–2872, 2004, doi: 10.1101/gad.1250204.

[39] K. Zhao, J. Du, Y. Rui, W. Zheng, J. Kang, J. Hou, K. Wang, W. Zhang, V. A. Simon, and X.-F. Yu, "Evolutionarily conserved pressure for the existence of distinct G2/M cell cycle arrest and A3H inactivation functions in HIV-1 Vif," *Cell Cycle*, vol. 14, no. 6, pp. 838–847, Jan. 2015, doi: 10.1080/15384101.2014.1000212.

[40] F. Villanova, M. Barreiros, L. M. Janini, R. S. Diaz, and E. Leal, "Genetic diversity of HIV-1 gene vif among treatment-naive Brazilians," *AIDS Res. Hum. Retroviruses*, vol. 33, no. 9, pp. 952–959, Sep. 2017, doi: 10.1089/aid.2016.0230.

[41] M. Ooms, M. Letko, and V. Simon, "The structural interface between HIV-1 Vif and human APOBEC3H," *J. Virol.*, vol. 91, no. 5, Feb. 2017.

[42] M. Binka, M. Ooms, M. Steward, and V. Simon, "The activity spectrum of vif from multiple HIV-1 subtypes against APOBEC3g, APOBEC3f, and APOBEC3h," *J. Virol.*, vol. 86, no. 1, pp. 49–59, Oct. 2011, doi: 10.1128/jvi.06082-11.

JOSE S. ALTAMIRANO-FLORES received the M.Sc. degree in computer science (artificial intelligence) from the Instituto Tecnológico de León (ITL), in 2016. He is currently pursuing the Ph.D. degree in computer science with the Universidad Autónoma de San Luis Potosí (UASLP), Mexico. His main research interests include data mining, machine learning, pattern recognition, artificial intelligence, evolutionary computation, and bioinformatics.

SANDRA E. GUERRA-PALOMARES was born in San Luis Potosí, Mexico, in 1986. She received the bachelor's degree in chemistry from the Chemical Sciences School, Universidad Autónoma de San Luis Potosí (UASLP), San Luis Potosí, in 2008, and the master's and Ph.D. degrees in basic biomedical sciences from the Faculty of Medicine, UASLP, in 2010 and 2016, respectively. She is currently appointed as an Associate Researcher at the Viral and Human Genomics Laboratory, Faculty of Medicine, UASLP. She has authored six articles on HIV genomics and blood borne pathogens. She continues to be involved in work with HIV and the hepatitis B virus.

PEDRO G. HERNANDEZ-SANCHEZ was born in Mexico City, in 1984. He received the bachelor's degree in biochemistry engineering from the Instituto Tecnológico de La Paz, Mexico, in 2007, and the master's degree and the Ph.D. degree in basic biomedical sciences from the Faculty of Medicine, Universidad Autónoma de San Luis Potosí (UASLP), San Luis Potosí, Mexico, in 2011. He is currently working at the Centro de Investigación en Ciencias de la Salud y Biomedicina (CICSaB), UASLP. He has participated in the study of antiretroviral resistance mutations in HIV-1 and HIV subtype diversity and polymorphisms. He has author manuscripts related to Mexican HIV-1 protease sequence diversity and with the prevalence of drug resistance mutations among Mexican HIV Isolates.

JOSE L. RAMIREZ-GARCIALUNA was born in Mexico City, Mexico, in 1985. He received the Medical degree and the M.Sc. degree from the Universidad Autónoma de San Luis Potosí (UASLP), Mexico, in 2011 and 2014, respectively, and the Ph.D. degree in experimental surgery from McGill University, Montreal, QC, Canada, in 2019. Since 2020, he has been affiliated to the McGill University Health Centre and Swift Medical Inc., as Postdoctoral Fellow. He has authored over 20 articles on the immunological response to biomaterial implantation, the immunological response to wound healing, and the characterisation of novel approaches to promote wound healing in pre-clinical animal models. He is also involved in clinical research through the development of diagnostic and prognostic algorithms using machine learning and artificial intelligence.

J. RAFAEL ARGÜELLO-ASTORGA was born in Gomez Palacio, Mexico, in 1963. He received the Medical degree from the Faculty of Medicine, Universidad Autónoma de Coahuila (UAdeC), Torreón, Mexico, in 1987, and the Ph.D. degree in molecular biology from the University College London and the Royal Free Hospital School of Medicine, London, U.K., in 1999. He is currently the Head of the Biomedical Research Center, Department of Molecular Immunobiology, Faculty of Medicine, UAdeC, and the General Director of the Institute of Science and Genomic Medicine, Mexico. Over the last years, he has been working on the development of new molecular methods and bioinformatics platforms for the analysis of genetic diversity and its impact in monogenic and complex diseases in Latin American populations. He has published more than 50 scientific articles in prestigious journals. He received the Overseas Research Students Award, the Dynal Literature Prize in Oslo, Norway, and the Shirley Nolan Prize in London, England.

DANIEL E. NOYOLA was born in San Luis Potosí, Mexico, in 1968. He received the M.D. and Ph.D. degrees from the Universidad Autónoma de San Luis Potosí (UASLP). He carried out a Residency in pediatrics at the University of Connecticut and the Postdoctoral Fellowship in pediatric infectious diseases at the Baylor College of Medicine. Since 2000, he has been working at the Microbiology Department, Medical Faculty, UASLP, where he is currently a Professor and a Researcher. He is the author/coauthor of more than 75 publications in the international literature. His research interests focus on viral infections with special emphasis on the respiratory syncytial virus, influenza, and cytomegalovirus.

JUAN C. CUEVAS-TELLO (Member, IEEE) received the M.Sc. degree in computer science (artificial intelligence) from the Universidad Nacional Autónoma de Mexico (UNAM), in 2001, and the Ph.D. degree in computer science and artificial intelligence from the University of Birmingham, U.K., in 2007. He is currently a full-time Research Professor with the Engineering Faculty, Universidad Autónoma de San Luis Potosí (UASLP), Mexico. His main research areas include data mining, machine learning, pattern recognition, artificial neural networks, evolutionary computation, computer vision, deep learning, high-performance computing, and bioinformatics. He is a member of the Mexican Society on Artificial Intelligence and a Professional Member of the Association for Computing Machinery (ACM). He is a Paper Reviewer of *Neural Networks* and *Pattern Recognition* journals (Elsevier).

CHRISTIAN A. GARCÍA-SEPÚLVEDA was born in Santiago, Chile, in 1973. He received the Medical degree from the Faculty of Medicine, Universidad Autónoma de Coahuila (UAdeC), Torreón, Mexico, in 1999, and the Ph.D. degree in haematology and immunogenetics from the University College London and the Royal Free Hospital School of Medicine, in 2005. He worked as a Scientific Advisor at the Hospital Angeles Lomas Haemopoietic Stem Cell Transplant Unit. He is currently appointed as a Principal Investigator at the Faculty of Medicine, Viral and Human Genomics Laboratory, Universidad Autónoma de San Luis Potosí (UASLP), San Luis Potosí, Mexico. He has authored more than 40 articles on natural killer-cell immunogenetics, HIV genomics, and artificial intelligence applications on biological data. He continues to be involved in work with HIV, hepatitis B virus, and emerging viral infectious diseases.