An Investigation of Classification Algorithms for Predicting HIV Drug Resistance without Genotype Resistance Testing

Pascal Brandt[1,2], Deshendran Moodley[1], Anban W. Pillay[1], Christopher J. Seebregts[1,2], and Tulio de Oliveira[3]

[1] Centre for Artificial Intelligence Research and Health Architecture Laboratory, University of KwaZulu-Natal, Durban/CSIR Meraka, Pretoria, South Africa
[2] Jembi Health Systems, Cape Town and Durban, South Africa
[3] Africa Centre for Health and Population Studies, Nelson R. Mandela School of Medicine, University of KwaZulu-Natal, Durban, South Africa

Abstract. The development of drug resistance is a major factor impeding the efficacy of antiretroviral treatment of South Africa's HIV infected population. While genotype resistance testing is the standard method to determine resistance, access to these tests is limited in low-resource settings. In this paper we investigate machine learning techniques for drug resistance prediction from routine treatment and laboratory data to help clinicians select patients for confirmatory genotype testing. The techniques, including binary relevance, HOMER, MLkNN, predictive clustering trees (PCT), RAkEL and ensemble of classifier chains were tested on a dataset of 252 medical records of patients enrolled in an HIV treatment failure clinic in rural KwaZulu-Natal in South Africa. The PCT method performed best with a discriminant power of 1.56 for two drugs, above 1.0 for three others and a mean true positive rate of 0.68. These methods show potential for application where access to genotyping is limited.

Keywords: HIV, treatment failure, machine learning, multi-label classification, clinical decision support.

1 Introduction

South Africa has one of the highest HIV infection rates in the world with more than 5.6 million infected people[1]. Consequently, the country has the largest antiretroviral treatment program in the world with more than one and a half million people on treatment [1]. The recommended treatment for HIV/AIDS (known as highly active antiretroviral therapy (HAART)) is a combination of three drugs from two or more different drug groups. HIV treatment failure occurs when the antiretroviral drugs (ARVs) no longer controls the infection and is due to, amongst other reasons, the development of drug resistance [2].

[1] `http://www.unaids.org/en/regionscountries/countries/southafrica/`

J. Gibbons and W. MacCaull (Eds.): FHIES 2013, LNCS 8315, pp. 236–253, 2014.

The standard method to identify resistance to specific drugs is the genotype resistance test (GRT) [3]. The GRT is a biochemical test conducted on a sample of the HIV population in the blood of an infected patient. Resistance algorithms such as Rega [4] or Stanford [3] are used to interpret the results and predict actual resistance from the viral genetic data. While some studies conclude that the cost of including GRT into treatment guidelines might be cost neutral [5], the current South African guidelines do not include GRT for every patient and it is therefore considered a limited resource.

Our aim in this work was to investigate the extent to which six multi-label classification techniques could predict the resistance level without a genotype test. We used data from a comprehensive HIV-1 ART treatment program with access to GRT in South Africa. The performance of the techniques was evaluated by comparing the predictions produced against the results obtained using GRT. The predictions produced by these classification techniques could help care providers to decide whether or not to refer a patient for GRT or to help select an optimal therapy if the regimen is to be changed.

The rest of the paper is organised as follows. In section 2 we describe previous work on factors contributing to HIV drug resistance and the use of machine learning techniques for HIV drug resistance prediction. Section 3 provides a description of the data and the machine learning techniques used in this study. The results are described in section 4 and an analysis is given in section 5. Conclusions are drawn and directions for future work are given in section 6.

2 Previous Work

Studies have shown that poor adherence to treatment regimen is an important factor influencing the development of drug resistance [6,7,8]. Patients who start treatment with a high viral load are more likely to develop resistance [8]. Other factors that influence resistance include exposure to more and greater variation of drugs as well as drug regimens that only partially suppress the virus population [9].

Computerised predictive models have been found to be useful for clinicians in practice and can sometimes even outperform human experts [10,11]. Machine learning techniques such as artificial neural networks, random forests and support vector machines have used genotype data to provide useful predictions about patient outcomes [12,13,14]. Using such methods to select new regimens has also been shown to be viable [15,16,17]. Furthermore, predictions have been shown to be more accurate when clinical data is included in training [15]. However, limited research is available on the efficacy of machine learning techniques for prediction without genotype data. A recent study in this regard is [16]. At a population level, this could potentially help optimise utilization of a scarce resource, such as resistance genotyping.

3 Methods

3.1 Study Data

The training dataset was constructed from anonymised records of adult (age > 16) patients attending the treatment failure management clinic at the Africa Centre in Mtubatuba, KwaZulu-Natal, South Africa[2].

Table 1 summarises the dataset attributes. Patients in the dataset also have a viral isolate associated with their clinical record, which is used to determine the resistance profile of the patient. This resistance profile is then used to construct the label sets associated with each training example. Resistance to a drug is determined using the HIVDB 6.0.5 algorithm [3] and a patient is considered resistant to a drug if the algorithm returns a susceptibility value of ≤ 0.5.

To store patient data, we used RegaDB, an open source patient-centric clinical data management system that stores data related to HIV treatment [18]. The data was stored longitudinally in a relational database. A software utility[3] was developed that uses the RegaDB API to extract patient data in the ARFF[4] format.

3.2 Multi-label Classification

Since each patient may develop resistance to multiple ARVs, multi-label classification is required. Multi-label classification involves associating each input example with multiple labels, rather than a single label [19]. In this study each patient record is associated with a set of 11 binary labels indicating resistance (presence of label) or susceptibility (absence of label) to each ARV drug.

There are three solution groups for solving the problem of multi-label classification. Problem transformation (PT) methods divide the problem into a set of multi-class classification problems and combine the result to form the multi-label prediction. Algorithm adaptation (AA) methods construct specialized algorithms, often by modifying existing multi-class prediction algorithms, to produce a prediction. Ensemble methods (EM) are developed on top of common problem transformation or algorithm adaptation methods [20].

3.3 Stratification

In order for the performance of each cross validation cycle to be representative of the performance if the full dataset was used for training, it is necessary to construct each fold so that the label distribution in the fold is the same as for the complete dataset [21]. In the case of binary classification, the task of stratifying

[2] The study and drug resistance analysis of the data was approved by the Biomedical Research Ethics Committee of the University of KwaZulu-Natal (BF052/010) and the Provincial Health Research Committee of the KwaZulu-Natal Department of Health (HRKM176/10).

[3] Source code available at `https://github.com/psbrandt/dsm`

[4] `http://weka.wikispaces.com/ARFF+(book+version)`

Table 1. Summary of features used to construct the predictive models

Category	Features	Types	Count
Demographic	Age, Weight, Geographic Location, Ethnicity, Gender, Province, Country of Origin	Numeric, Categorical	7
Clinical	Drug Exposure (Tenofovir, Lopinavir/r, Atazanavir, Zidovudine, Ritonavir, Efavirenz, Abacavir, Nevirapine, Raltegravir, Stavudine, Didanosine, Lamivudine), Recent Blood HB, Recent Blood ALT, Recent Blood Creatinine Clearance, Other Drug Exposure (three features), Tuberculosis Therapy (Prior, During, Post), HTLV-1 Status, HBV Status	Numeric, Categorical	23
Adherence	Treatment Break, Patient Estimated Adherence, Missed, Buddy, Remember, Counseling, Side Effects, Worst Stop, Disclosure, Names, Stop	Categorical	11
Other	Transmission Group, Other Co-morbidities, Partner On Treatment, Exposure to Single Dose NVP, Identified Virological Failure Reason, Traditional Medicine, Alcohol Consumption, TB Treatment Starting Soon, Diarrhea or Vomiting	Categorical	9
Derived	Baseline Viral Load, Time on Failing Regimen, Drug Exposure Count, Pre-Resistance Testing Viral Load, Median Viral Load, Recent CD4 Count Gradient, Post-Treatment CD4 Count, Baseline CD4 Gradient, Pre-Resistance Testing CD4 Count, Mean Viral Load, Pre-Resistance Testing Immunological Failure, Virus Ever Suppressed	Numeric, Categorical	12
Total			**62**

the dataset is simple, since there is only one target label whose distribution must be maintained. However, in multi-label datasets there are multiple distributions that must be maintained.

Since the dataset used in the work is small, the Meka[5] implementation of the iterative stratification algorithm [22] was used to generate 10 folds to be used for cross validation.

3.4 Model Development

Seven multi-label classification models were trained and tested for their ability to predict a known resistance result from demographic and treatment data. The models were selected as a representative sample of those available in the multi-label classification domain [20]. In each case, 10 fold cross validation was done using folds generated by the iterative stratification algorithm described in section 3.3.

Binary Relevance with Support Vector Machine (SVM) base classifier (BR-SVM). The radial basis function kernel was selected because it is able to model non-linear relationships. Model parameters were optimized using the technique in [20,23].

[5] `http://meka.sourceforge.net/`

Binary Relevance with naive Bayes (NB) base classifier (BR-NB). The second experiment conducted replaces the SVM base classifier with a naive Bayes classifier. A naive Bayes classifier models the probability of the class variable using the simplifying assumption that each feature in the feature vector is independent [24].

HOMER. The third experiment used the Hierarchy Of Multi-label classifiERs (HOMER) method [25]. This problem transformation method trains classifiers in a hierarchy on subsets of the labels. The MULAN[6] implementation of this algorithm was used.

MLkNN. The first algorithm adaptation method tested was the Multi-Label k Nearest Neighbours (MLkNN) algorithm [26]. This algorithm, based on the traditional k nearest neighbours method, works by first identifying the k nearest neighbours of an unseen instance and making a prediction based on the labels associated with these neighbours. The MLkNN implementation provided by MULAN was used for this experiment.

Predictive Clustering Trees (PCT). The fifth experiment used the predictive clustering framework algorithm adaptation method [27]. This method builds a clustering tree using top-down induction. The Clus[7] implementation was used.

RAkEL. The first ensemble method tested was RAndom k-labELsets (RAkEL). This method trains ensemble members on small random subsets of labels [28]. The MULAN implementation of the RAkEL algorithm was used and a naive Bayes base classifier was selected.

Ensemble of Classifier Chains (ECC). The MULAN implementation of the ensemble of classifier chains (ECC) method (with a naive Bayes base classifier) was used as the second ensemble method experiment. This method builds on the binary relevance idea by extending the attribute space by adding a binary feature to represent the label relevances of all previous classifiers, thereby forming a classifier chain [29].

3.5 Evaluation Metrics

There are many possible evaluation metrics that can be used to measure the performance of a classifier [30] and an attempt was made select a representative sample of those available. Most metrics are calculated from the confusion matrix generated by a classification or cross validation run. Since the dataset used in this study was highly unbalanced, it was necessary to choose evaluation metrics that are informative in the presence of label imbalance. For this reason, we chose to use true positive rate (TPR, often called sensitivity) and true negative rate (TNR, often called specificity). Along with TPR and TNR, we chose to use Matthew's correlation coefficient, discriminant power and area under the receiver operating characteristic curve (AUC).

[6] `http://mulan.sourceforge.net/`
[7] `http://dtai.cs.kuleuven.be/clus/`

Matthew's correlation coefficient (MCC) is a correlation coefficient between the observed and predicted classes. Its value is in the range $[-1; 1]$ with 1 indicating perfect prediction, 0 indicating no better than random prediction and -1 indicating total disagreement between prediction and observation [31]. It is defined in equation 1.

$$MCC = \frac{tp \cdot fn - fp \cdot fn}{\sqrt{(tp + fp) \cdot (tp + fn) \cdot (tn + fp) \cdot (tn + fn)}} \tag{1}$$

where tp is the number of true positives, fn the number of false negatives, tn the number of true negatives and fp the number of false positives from the confusion matrix.

Discriminant power (DP) is a measure that summarizes sensitivity and specificity and is a measure of how well a classifier distinguishes between positive and negative examples [32]. It is defined in equation 2.

$$DP = \frac{\sqrt{3}}{\pi} ln \left(\frac{tp}{fp} \cdot \frac{tn}{fn} \right) \tag{2}$$

A classifier is a poor discriminant if $DP < 1$, limited if $DP < 2$, fair if $DP < 3$ and good otherwise.

Receiver operating characteristic (ROC) graphs depict relative tradeoffs between benefits (true positives) and costs (false negatives) and are insensitive to changes in class distribution. Area under the receiver operating characteristic curve (AUC) is a single scalar value that represents expected ROC performance [33]. Note that ROC graphs can only be plotted for classifiers that produce a probability estimate and hence no such graphs are plotted for the BR-SVM and PCT methods.

3.6 Validation

In order to ensure valid and robost results, model performance was averaged over the 10 cross validation cycles. Further, the use of stratification during the construction of the folds helps ensure that all training and test cycles are performed using data that is representative of the complete dataset.

4 Results

4.1 Dataset Characteristics

The size of the dataset is 252 patients, with a mean age of 37.49. There are 160 females in the dataset (75.4%). Table 2 shows the number of patients with demonstrated resistance to each number of drugs.

Table 2. Number of patients resistant to each number of drugs

Drugs	0	1	2	3	4	5	6	7	8	9	10	11
Patients	19	0	3	11	5	18	72	5	11	38	45	25
Percent (%)	7.54	0.00	1.19	4.37	1.98	7.14	28.57	1.98	4.37	15.08	17.86	9.92

Table 3. Resistance and susceptibility counts for each drug

	efavirenz	**didanosine**	**emtricitabine**	**delavirdine**	**stavudine**	**nevirapine**
Resistant	229 (90.87%)	103 (40.87%)	217 (86.11%)	226 (89.68%)	93 (36.90%)	229 (90.87%)
Susceptible	23 (9.13%)	149 (59.13%)	35 (13.89%)	26 (10.32%)	159 (63.10%)	23 (9.13%)
	tenofovir	**etravirine**	**abacavir**	**lamivudine**	**zidovudine**	
Resistant	61 (24.21%)	196 (77.78%)	123 (48.81%)	217 (86.11%)	77 (30.56%)	
Susceptible	191 (75.79%)	56 (22.22%)	129 (51.19%)	35 (13.89%)	175 (69.44%)	

The final training dataset consists of 62 features per patient (see table 1). The average feature completeness is 76.84%. 36 of the features (58.06%) are over 90% complete and 24 features (30.71%) are 100% complete. Eight features (12.90%) are over 78% complete. 14 features are between 22% and 78% complete. Four features (6.45%) are less than 10% complete. Completeness here is defined as the number of patients having a value for the specific feature divided by the total number of patients.

It's important to note that in this dataset there are many more patients with resistance than without. As seen in table 2, only 19 (7.54%) showed no resistance. This results in a label imbalance in the training data. Table 3 shows, for each label, how many examples are associated with the label. Since we define label presence as indicating resistance, we can see that resistance heavily dominates that dataset. Tenofovir, zidovudine and stavudine are the only cases where non-resistance dominates over resistance. Didanosine and abacavir are relatively well balanced.

4.2 Model Evaluation

The numeric values for each evaluation metric are given in tables 4–6. Figures 1 - 3 show the ROC curves for the performance of each method that produces a probability estimate to which threshold variation can be applied. The ROC curves are vertically averaged across the 10 folds as put forward in [33]. The *AUC* along with the standard error is given in the legend for each method in each graph.

Problem Transformation Methods. The BR-SVM method produces blanket positive predictions (which occur when $TPR = 1.0$ and $TNR = 0.0$) for five labels (efavirenz, emtricitabine, delavirdine, nevirapine and lamivudine). These five labels correspond to the five most unbalanced labels in the dataset ($> 86\%$ of examples are resistant). For all other labels except tenofovir the method does not

discriminate well, as seen by the low DP values in table 4. Over all, the BR-SVM method performs poorly, with only the performance of predicting tenofovir moderately above the performance of a random classifier based on the MCC value.

When we switched the base classifier to naive Bayes, we saw that BR-NB performed drastically better. There were no blanket positive predictions and the mean DP value increased by over 20%. The mean TPR and TNR values also increased. BR-NB had an MCC value of just over 0.3 for tenofovir, meaning that it is significantly better than a random classifier at predicting, in this case, absence of resistance to this drug. The mean AUC of all the drugs for the BR-NB method was 10% above that of a random classifier (0.5) at 0.6.

HOMER performed slightly worse than the BR-NB method, with the mean of each metric less than BR-NB. The exception was the mean TNR value, which didn't change. HOMER outperforms BR-NB at predicting cases of lamivudine resistance, with a TPR of 0.94. The MCC value for the zidovudine indicates that HOMER performs better than a random classifier for this label. The mean AUC for HOMER is 0.57.

Algorithm Adaptation Methods. MLkNN appears to be the worst performing of all the methods. It has the lowest mean value for each statistic and blanket positive predictions for six labels (efavirenz, emtricitabine, delavirdine, nevirapine, etravirine and lamivudine). Further, the mean AUC is 0.49, which indicates worse than random performance. This is confirmed by multiple MCC and DP values less than zero.

The PCT method does not suffer from blanket positive predictions and has the highest mean DP, MCC and TNR values. It has three MCC values above 0.35 and numerous DP values above 1. These values indicate that PCT is substantially better than a random classifier at predicting resistance to efavirenz, delavirdine and nevirapine. PCT performs especially well relative to the other methods for efavirens and nevirapine with a TPR of 0.97 and DP value of 1.56 in both cases. Unlike the other methods, it does not appear to be a good predictor for tenofovir.

Ensemble Methods. The RAkEL method suffers from no blanket positive predictions, but does have MCC and DP values below zero for two drugs, indicating very poor performance in these cases. RAkEL appears at first glance to perform relatively well for lamivudine, with a DP value of 1.06. However, the TNR for this drug is only 0.11, indicating that the method is not good at detecting negative examples. The RAkEL method has an average AUC of 0.58, which puts it between BR-NB and HOMER in terms of this metric.

The ECC method performs worse than random for efavirenz, nevirapine and etravirine, but has relatively high DP values (greater than 1) for emtricitabine, tenofovir and lamivudine. Of the latter three, only tenofovir also shows an MCC value significantly greater than zero (0.38). The average AUC for the ECC method is 0.63 with an AUC value of 0.72 for tenofovir, making it the best performer for this drug and in terms of the ROC analysis in general.

Table 4. Results of the problem transformation methods

	BR-SVM				BR-NB					HOMER				
	TPR	**TNR**	**MCC**	**DP**	**TPR**	**TNR**	**MCC**	**DP**	**AUC**	**TPR**	**TNR**	**MCC**	**DP**	**AUC**
efavirenz	1.00	0.00	-	-	0.86	0.26	0.10	0.43	0.54	0.89	0.13	0.02	0.09	0.51
didanosine	0.27	0.80	0.08	0.22	0.54	0.62	0.16	0.36	0.62	0.56	0.59	0.15	0.34	0.62
emtricitabine	1.00	0.00	-	-	0.93	0.20	0.15	0.63	0.61	0.94	0.17	0.16	0.70	0.57
delavirdine	1.00	0.00	-	-	0.83	0.23	0.05	0.22	0.56	0.84	0.19	0.02	0.11	0.51
stavudine	0.22	0.85	0.08	0.24	0.38	0.70	0.08	0.18	0.58	0.38	0.81	0.20	0.50	0.60
nevirapine	1.00	0.00	-	-	0.86	0.26	0.10	0.43	0.54	0.89	0.13	0.02	0.09	0.51
tenofovir	0.26	0.94	0.27	0.92	0.43	0.86	0.31	0.85	0.65	0.44	0.84	0.29	0.80	0.66
etravirine	0.99	0.02	0.03	0.31	0.88	0.16	0.05	0.17	0.60	0.52	0.52	0.03	0.08	0.50
abacavir	0.50	0.61	0.12	0.26	0.54	0.64	0.18	0.41	0.63	0.63	0.57	0.19	0.43	0.66
lamivudine	1.00	0.00	-	-	0.93	0.20	0.15	0.63	0.61	0.94	0.17	0.16	0.70	0.57
zidovudine	0.05	0.96	0.03	0.15	0.38	0.79	0.17	0.45	0.63	0.38	0.82	0.22	0.57	0.62
mean	0.66	0.38	0.10	0.35	0.69	0.45	0.14	0.43	0.60	0.67	0.45	0.13	0.40	0.57
std dev	0.40	0.44	0.09	0.28	0.23	0.27	0.07	0.21	0.04	0.23	0.30	0.10	0.28	0.06
best	1.00	0.96	0.27	0.92	0.93	0.86	0.31	0.85	0.65	0.94	0.84	0.29	0.80	0.66

BR-SVM - Binary relevance with support vector machine base classifier, **BR-NB** - Binary relevance with naive Bayes base classifier, **HOMER** - Hierarchy of multi-label classifiers, **TPR** - True positive rate, **TNR** - True negative rate, **MCC** - Matthew's correlation coefficient, **DP** - Discriminative power, **AUC** - Area under the receiver operating characteristic curve

5 Discussion

The multi-label classifiers were evaluated for the ability to predict known resistance to a set of ARVs based only on prior patient biographical and treatment history data excluding the result of a genotype resistance test. Some of the methods were found to be good predictors for specific drugs. For example, PCT predicts resistance to nevirapine with a true positive rate of 0.97.

The mean AUC for the ECC method (0.63) is comparable to the results obtained in the recent study by Revell in [16], which used data from a Southern African dataset on a model trained with a number of international datasets. Revell predicted virological response, while we predicted resistance to specific drugs. Revell achieves a TPR of 0.60 compared to our 0.71 (mean over all drugs) and TNR of 0.62 compared to our 0.38 (mean over all drugs) for ECC. For PCT we have a mean TPR of 0.68 and mean TNR of 0.53, which is comparable to Revell's result. This could imply that if an equally large training set were used in our models, results may improve.

Models that perform well on the highly imbalanced labels (such as BR-NB and PCT) perform less well on the relatively balanced labels (didanosine and abacavir). This supports the idea that no single model should be used to assess resistance to all drugs and the results of multiple models should be combined into an ensemble prediction to produce the best results.

Table 5. Results of the algorithm adaptation methods

	MLkNN					PCT			
	TPR	**TNR**	**MCC**	**DP**	**AUC**	**TPR**	**TNR**	**MCC**	**DP**
efavirenz	1.00	0.00	-	-	0.44	0.97	0.35	0.39	1.56
didanosine	0.08	0.83	-0.14	-0.51	0.45	0.47	0.78	0.26	0.62
emtricitabine	1.00	0.00	-	-	0.50	0.94	0.31	0.30	1.09
delavirdine	1.00	0.00	-0.02	-	0.52	0.97	0.31	0.36	1.45
stavudine	0.04	0.97	0.03	0.18	0.48	0.31	0.77	0.09	0.22
nevirapine	1.00	0.00	-	-	0.44	0.97	0.35	0.39	1.56
tenofovir	0.00	0.98	-0.06	-	0.52	0.28	0.87	0.18	0.55
etravirine	1.00	0.00	-	-	0.47	0.91	0.23	0.18	0.60
abacavir	0.39	0.63	0.02	0.04	0.54	0.56	0.70	0.26	0.60
lamivudine	1.00	0.00	-	-	0.50	0.94	0.31	0.30	1.09
zidovudine	0.01	0.99	0.04	0.46	0.53	0.19	0.83	0.04	0.11
mean	0.59	0.40	-0.02	0.04	0.49	0.68	0.53	0.25	0.86
std dev	0.48	0.47	0.07	0.41	0.04	0.32	0.26	0.12	0.52
best	1.00	0.99	0.04	0.46	0.54	0.97	0.87	0.39	1.56

Table 6. Results of the ensemble methods

	RAkEL					ECC				
	TPR	**TNR**	**MCC**	**DP**	**AUC**	**TPR**	**TNR**	**MCC**	**DP**	**AUC**
efavirenz	0.92	0.22	0.13	0.62	0.59	1.00	0.00	-0.02	-	0.68
didanosine	0.52	0.58	0.10	0.23	0.58	0.41	0.73	0.15	0.35	0.64
emtricitabine	0.93	0.20	0.16	0.67	0.60	1.00	0.03	0.09	1.02	0.62
delavirdine	0.96	0.04	0.01	0.05	0.60	1.00	0.04	0.19	-	0.69
stavudine	0.49	0.62	0.11	0.26	0.54	0.25	0.82	0.09	0.24	0.54
nevirapine	0.91	0.22	0.12	0.56	0.60	1.00	0.00	-0.02	-	0.61
tenofovir	0.38	0.88	0.28	0.82	0.66	0.39	0.93	0.38	1.16	0.72
etravirine	0.85	0.11	-0.05	-0.20	0.50	0.97	0.02	-0.02	-0.20	0.56
abacavir	0.42	0.74	0.18	0.42	0.60	0.59	0.70	0.28	0.65	0.68
lamivudine	0.98	0.11	0.19	1.06	0.60	1.00	0.03	0.09	1.02	0.60
zidovudine	0.26	0.82	0.09	0.27	0.55	0.26	0.89	0.19	0.58	0.63
mean	0.69	0.41	0.12	0.43	0.58	0.71	0.38	0.13	0.60	0.63
std dev	0.28	0.32	0.09	0.36	0.04	0.33	0.42	0.13	0.46	0.06
best	0.98	0.88	0.28	1.06	0.66	1.00	0.93	0.38	1.16	0.72

MLkNN - Multi-label k nearest neighbours, **PCT** - Predictive clustering trees, **RAkEL** - Random k-labelsets, **ECC** - Ensemble of classifier chains, **TPR** - True positive rate, **TNR** - True negative rate, **MCC** - Matthew's correlation coefficient, **DP** - Discriminative power, **AUC** - Area under the receiver operating characteristic curve

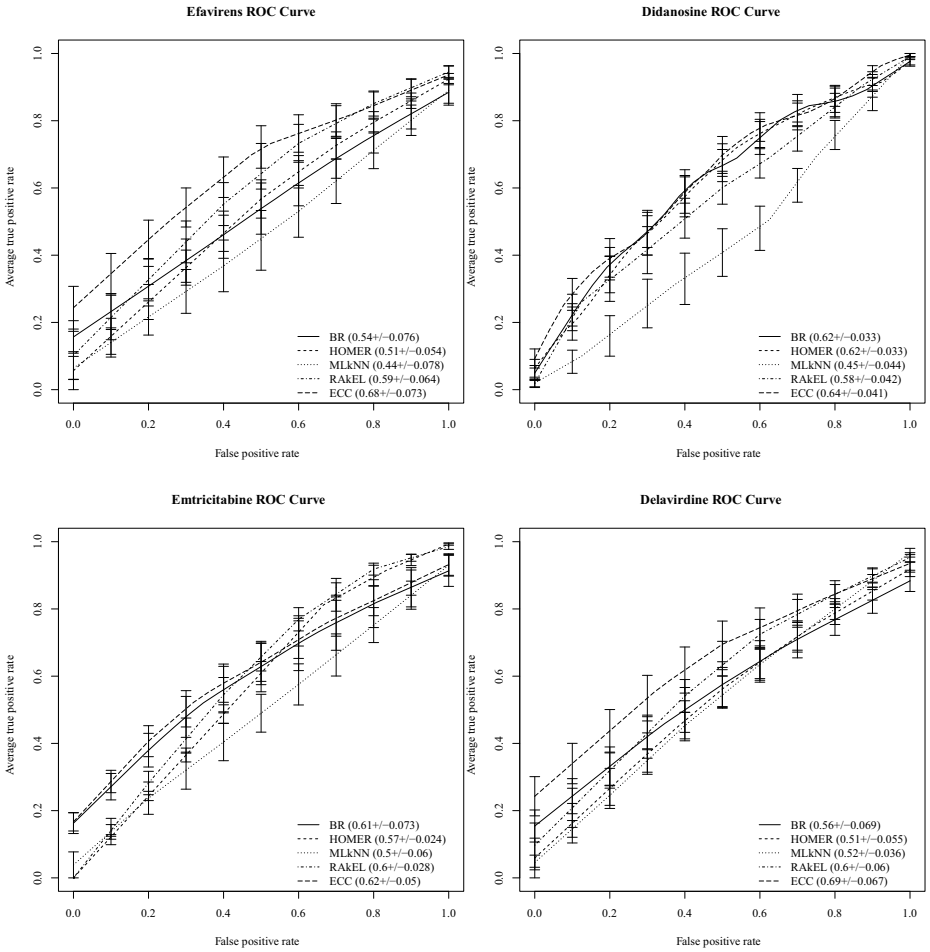

Fig. 1. ROC curves for efavirens, didanosine, emtricitabine and delavirdine. *AUC* and standard error are given for each method in parentheses

ROC - Receiver operating characteristic, **AUC** - Area under ROC curve, **BR** - Binary relevance (with naive Bayes base classifier), **HOMER** - Hierarchy of multi-label classifiers, **MLkNN** - Multi-label k nearest neighbours, **RAkEL** - Random k-labelsets, **ECC** - Ensemble of classifier chains

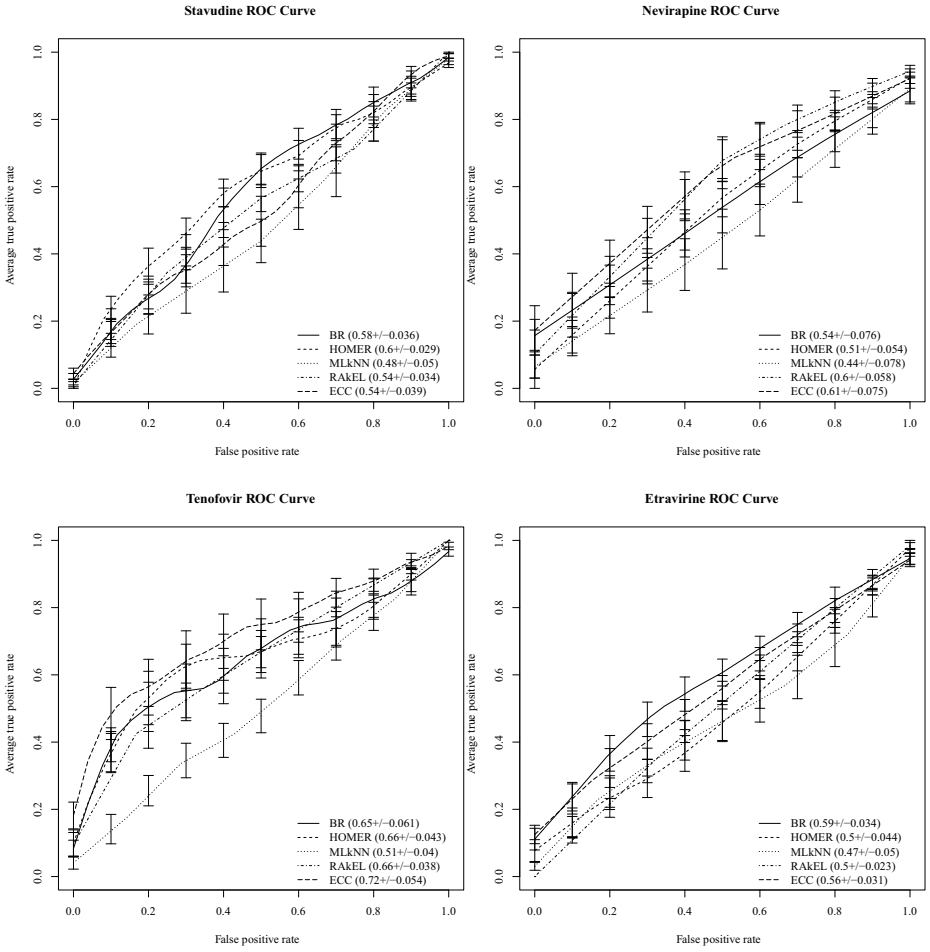

Fig. 2. ROC curves for stavudine, nevirapine tenofovir and etravirine. *AUC* and standard error are given for each method in parentheses

ROC - Receiver operating characteristic, **AUC** - Area under ROC curve, **BR** - Binary relevance (with naive Bayes base classifier), **HOMER** - Hierarchy of multi-label classifiers, **MLkNN** - Multi-label k nearest neighbours, **RAkEL** - Random k-labelsets, **ECC** - Ensemble of classifier chains

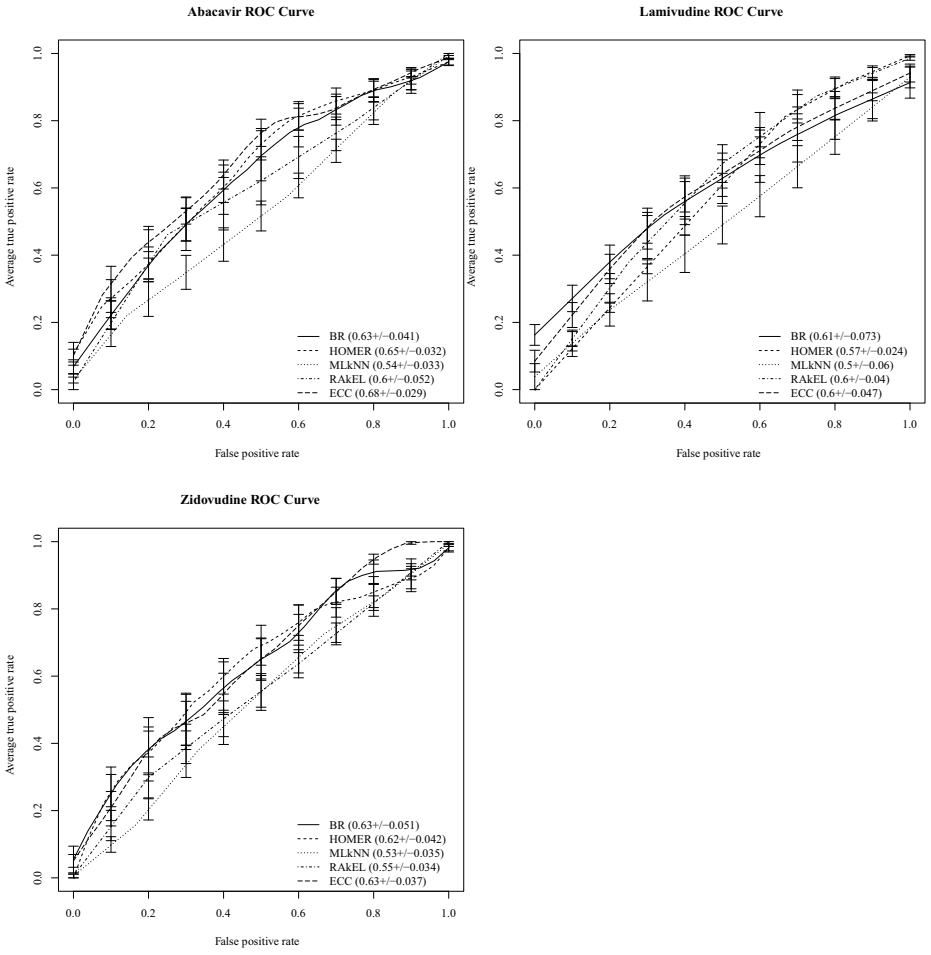

Fig. 3. ROC curves for abacavir, lamivudine and zidovudine. *AUC* and standard error are given for each method in parentheses

ROC - Receiver operating characteristic, **AUC** - Area under ROC curve, **BR** - Binary relevance (with naive Bayes base classifier), **HOMER** - Hierarchy of multi-label classifiers, **MLkNN** - Multi-label k nearest neighbours, **RAkEL** - Random k-labelsets, **ECC** - Ensemble of classifier chains

Techniques that have high true positive rates often have a relatively low true negative rate, which means that the rate of false positives is high. This could result in an increased number of genotype tests being requested. However, if GRT is not available and/or if the care provider decides that a regimen change is necessary, the false positives can be seen as the algorithms making conservative predictions. It is important to try and retain patients as long as possible on first line treatment regimens since second line therapy costs as much as 2.4 times more than first line therapy and compromises treatment outcomes [5], but if it is decided to switch regimen, conservative prediction results could be used to help design a treatment regimen to which the virus is susceptible.

Algorithms like PCT, BR-NB and ECC have high values for all of the metrics for the drugs that they perform well on. This should give us confidence that the decision support information provided is in fact good and not just the result of a single outlier metric value. Conversely, we should have less confidence in the performance of a technique on a drug if only one of the metrics show good performance, for example, the MLkNN technique has a high TNR value (0.97) for stavudine, but all other metrics indicate very poor performance. Techniques that produce blanket positive predictions provide no information that can be used to support decisions.

The results of this analysis need to be interpreted in light of certain limitations. The most important limitation was the lack of any objective or proxy measures of adherence, such as pharmacy refill data or medication possession ratios, as this information was not available routinely in the program. Adherence is one of the most important determinants of drug resistance and is an essential variable in a predictive dataset. It is also important to consider that certain attributes captured for each patient are self-reported (such as the adherence attributes) and that the reliability of such data may be questionable.

The small number of training examples ($n = 252$) could be partly responsible for the limitations in performance. Another factor that could influence the performance is the presence of features in the dataset that contain incomplete data. Unfortunately, incomplete and inaccurate data are common features of public health ART treatment program data in South Africa and other countries. Patients miss clinic visits for a number of health and socio-economic reasons and it is often the case that laboratory test results are not available when needed by the clinician. Further investigation should be done on the use of feature selection to ensure the optimal subset of features is being used for training and prediction, since irrelevant and redundant features can reduce classification accuracy [34].

Resource-limited public health treatment programs are usually characterised by limited availability of GRT, second line therapy and highly skilled clinicians. In these settings, the predictive power demonstrated by the classifiers may be sufficient to facilitate and improve clinical decision-making. If the primary goal is to optimize the use of GRT, then the method with the highest mean TNR should be selected, which is PCT. If the primary goal is new therapy selection, then

BR-NB should be used, since it has the highest mean TPR. If it is of particular importance to know about the presence of resistance to a specific drug, then the method with the highest TNR for that drug should be selected, since this minimises false positives.

6 Conclusion and Future Work

We have demonstrated that the machine learning techniques examined in this work can be used to a limited degree to predict HIV drug resistance and mostly perform better than a random classifier and in some cases, substantially better. Even though some results show promise there is insufficient evidence to support the conclusion that machine learning prediction models using only clinical, adherence and demographic details, can replace GRT. However, these techniques may be useful in resource-limited public health settings where decisions such as whether to remain on the same therapy, and if not, which new drugs to select, need to be made in the absence of a GRT result or a specialized clinician. While none of the seven methods stands out as a significantly good predictor for resistance to all drugs, some methods perform relatively well on some drugs. For example, PCT is good at identifying resistance to nevirapine. A future area of work could be to construct individual predictive models per drug, using different underlying methods and then combine these results into an ensemble to produce one prediction. The advent of limited resistance testing in the public ART program in South Africa and development of national surveillance data sets [35] will also allow the construction of larger datasets and potentially increase the accuracy of machine learning techniques. The existing techniques will be tested on these larger datasets as they become available. We also plan to extend the analysis and investigate the predictive potential at different stages of treatment failure. Time spent on a failing regimen could lead to accumulation of resistance mutations and this should be taken into consideration when care providers need to decide on a course of action. The software developed could easily be integrated into existing ART programs that use RegaDB and could act as a passive early warning system to alert providers to patients for whom the classifiers predict high levels of resistance.

Acknowledgements. PB received a scholarship for this study from the Health Architecture Laboratory (HeAL) project funded by grants from the Rockefeller Foundation (Establishing a Health Enterprise Architecture Lab, a research laboratory focused on the application of enterprise architecture and health informatics to resource-limited settings, Grant Number: 2010 THS 347) and the International Development Research Centre (IDRC) (Health Enterprise Architecture Laboratory (HeAL), Grant Number: 106452-001). CS, TdO and the Southern African Treatment and Research Network (SATuRN) network were funded by grants from the Delegation of the European Union to South Africa (SANTE 2007 147-790; Drug

resistance surveillance and treatment monitoring network for the public sector HIV antiretroviral treatment programme in the Free State) and CDC/PEPFAR via a grant to the Centre for the AIDS Programme of Research in South Africa (CAPRISA) (project title: Health System Strengthening and the HIV Treatment Failure Clinic System (HIV-TFC)). TdO is supported by the Wellcome Trust (grant number 082384/Z/07/Z). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Statistics, S.A.: Statistical release Mid-year population estimates (July 2011), http://www.statssa.gov.za/Publications/statsdownload.asp?PPN=P0302
2. Rossouw, T., Tulio, O., Lessels, R.J.: HIV & TB Drug Resistance & Clinical Management Case Book. South African Medical Research Council Press (2013)
3. Liu, T.F., Shafer, R.W.: Web resources for HIV type 1 genotypic-resistance test interpretation. Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America 42(11), 1608–1618 (2006)
4. Van Laethem, K., De Luca, A., Antinori, A., et al.: A genotypic drug resistance interpretation algorithm that significantly predicts therapy response in HIV-1-infected patients. Antiviral Therapy 7(2), 123–129 (2002)
5. Rosen, S., Long, L., Sanne, I., et al.: The net cost of incorporating resistance testing into HIV/AIDS treatment in South Africa: a Markov model with primary data. Journal of the International AIDS Society 14(1), 24 (2011)
6. Robbins, G.K., Daniels, B., Zheng, H., et al.: Predictors of antiretroviral treatment failure in an urban HIV clinic. Journal of Acquired Immune Deficiency Syndromes 44(1), 30–37 (1999)
7. Parienti, J.J., Massari, V., Descamps, D., et al.: Predictors of virologic failure and resistance in HIV-infected patients treated with nevirapine- or efavirenz-based antiretroviral therapy. Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America 38(9), 1311–1316 (2004)
8. Harrigan, P.R., Hogg, R.S., Dong, W.W.Y., et al.: Predictors of HIV drug-resistance mutations in a large antiretroviral-naive cohort initiating triple antiretroviral therapy. The Journal of Infectious Diseases 191(3), 339–347 (2005)
9. Di Giambenedetto, S., Zazzi, M., Corsi, P., et al.: Evolution and predictors of HIV type-1 drug resistance in patients failing combination antiretroviral therapy in Italy. Antiviral Therapy 14(3), 359–369 (2009)
10. Larder, B., Revell, A., Mican, J.M., et al.: Clinical evaluation of the potential utility of computational modeling as an HIV treatment selection tool by physicians with considerable HIV experience. AIDS Patient Care and STDs 25(1), 29–36 (2011)
11. Zazzi, M., Kaiser, R., Sönnerborg, A., et al.: Prediction of response to antiretroviral therapy by human experts and by the EuResist data-driven expert system (the EVE study). HIV Medicine 12(4), 211–218 (2011)
12. Larder, B., Wang, D., Revell, A., et al.: The development of artificial neural networks to predict virological response to combination HIV therapy. Antiviral Therapy 12(1), 15–24 (2007)

13. Prosperi, M.C.F., Altmann, A., Rosen-Zvi, M., et al.: Investigation of expert rule bases, logistic regression, and non-linear machine learning techniques for predicting response to antiretroviral treatment. Antiviral Therapy 14(3), 433–442 (2009)
14. Altmann, A., Rosen-Zvi, M., Prosperi, M., et al.: Comparison of classifier fusion methods for predicting response to anti HIV-1 therapy. PloS One 3(10), e3470 (2008)
15. Rosen-Zvi, M., Altmann, A., Prosperi, M., et al.: Selecting anti-HIV therapies based on a variety of genomic and clinical factors. Bioinformatics 24(13), 399–406 (2008)
16. Revell, A.D., Wang, D., Wood, R., et al.: Computational models can predict response to HIV therapy without a genotype and may reduce treatment failure in different resource-limited settings. Journal of Antimicrobial Chemotherapy (March 2013)
17. Prosperi, M.C.F., Rosen-Zvi, M., Altmann, A., et al.: Antiretroviral therapy optimisation without genotype resistance testing: a perspective on treatment history based models. PloS One 5(10), e13753 (2010)
18. Libin, P., Beheydt, G., Deforche, K., et al.: RegaDB: Community-driven data management and analysis for infectious diseases. Bioinformatics, 1–5 (April 2013)
19. Tsoumakas, G., Katakis, I., Vlahavas, I.: Mining multi-label data. Data Mining and Knowledge..., 1–20 (2010)
20. Madjarov, G., Kocev, D., Gjorgjevikj, D., et al.: An extensive experimental comparison of methods for multi-label learning. An Extensive Experimental Comparison of Methods for Multi-label Learning 45, 3084–3104 (2012)
21. Kohavi, R.: A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. International Joint Conference on Artificial Intelligence 14(12), 1137–1143 (1995)
22. Sechidis, K., Tsoumakas, G., Vlahavas, I.: On the stratification of multi-label data. In: Machine Learning and Knowledge ... (2011)
23. Hsu, C.W., Chang, C.C., Lin, C.J.: A Practical Guide to Support Vector Classification. Bioinformatics 1(1), 1–16 (2010)
24. Rish, I.: An empirical study of the naive Bayes classifier. IJCAI 2001 Workshop on Empirical Methods in Artificial... (2001)
25. Tsoumakas, G.: Effective and efficient multilabel classification in domains with large number of labels. In: Proc. ECML/PKDD 2008 Workshop on Mining Multidimensional Data, MMD 2008 (2008)
26. Zhang, M.L., Zhou, Z.H.: ML-KNN: A lazy learning approach to multi-label learning. Pattern Recognition 40(7), 2038–2048 (2007)
27. Blockeel, H., De Raedt, L.: Top-down induction of first-order logical decision trees. Artificial Intelligence 101(1-2), 285–297 (1998)
28. Tsoumakas, G., Vlahavas, I.P.: Random *k*-labelsets: An ensemble method for multilabel classification. In: Kok, J.N., Koronacki, J., Lopez de Mantaras, R., Matwin, S., Mladenič, D., Skowron, A. (eds.) ECML 2007. LNCS (LNAI), vol. 4701, pp. 406–417. Springer, Heidelberg (2007)
29. Read, J., Pfahringer, B., Holmes, G., et al.: Classifier chains for multi-label classification. Machine Learning 85(3), 333–359 (2011)
30. Sokolova, M., Lapalme, G.: A systematic analysis of performance measures for classification tasks. Information Processing & Management 45(4), 427–437 (2009)
31. Baldi, P., Brunak, S.R., Chauvin, Y., et al.: Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 16(5), 412–424 (2000)

32. Sokolova, M.V., Japkowicz, N., Szpakowicz, S.: Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation. In: Sattar, A., Kang, B.-H. (eds.) AI 2006. LNCS (LNAI), vol. 4304, pp. 1015–1021. Springer, Heidelberg (2006)
33. Fawcett, T.: An introduction to ROC analysis. Pattern Recognition Letters 27(8), 861–874 (2006)
34. Okun, O.: Introduction to Feature and Gene Selection. In: Feature Selection and Ensemble Methods for Bioinformatics: Algorithmic Classification and Implementations, pp. 117–122. IGI Global, Hershey (2011)
35. Conradie, F., Wilson, D., Basson, A., et al.: The 2012 southern African ARV drug resistance testing guidelines by the Southern African HIV Clinicians Society. Southern African Journal of HIV Medicine 13(4), 162–167 (2012)