Prediction of HIV Status from Demographic Data Using Neural Networks

Brain Leke-Betechuoh, Member, IEEE, Tshilidzi Marwala, Member, IEEE, Taryn Tim, Monica Lagazio

Abstract—Neural Networks are used as pattern recognition tools in data mining to classify HIV status of individuals based on demographic and socio-economic characteristics. The data consists of seroprevalence survey information and contains variables such as age, education, location, race, parity and gravidity. The radial basis function (RBF) neural network architecture was used for this study since as preliminary design showed this architecture to be the most optimal. The Bayesian method of training used was approximated with the evidence framework. The design of classifiers involves the assessment of classification performance, and this is based on the accuracy of the prediction using the confusion matrix. An accuracy of 84.24% was obtained in this design. This thus implies that the HIV status of an individual can be predicted using demographic data to 84.24% accuracy. A network comprising of 9 primary RBF, and MLP networks of structure 1-3-1 (input-hidden node-output node) and one secondary MLP network of structure 9-77-1, was used with a prior of 0.24693 and 144 training cycles which was found as the optimal training cycles.

Keywords: Bayesian, Classification, Neural Networks, Multi Layer Perceptron, Confusion Matrix, Genetic Algorithms, AIDS

I. INTRODUCTION

ACQUIRED Immunodeficiency Syndrome (AIDS) was first defined in 1982 to describe the first cases of unusual immune system failure that were identified in the previous year. The Human Immunodeficiency Virus (HIV) was later identified as the cause of AIDS. Since the identification of the virus and the disease, very little has been effective in stopping the spread. AIDS is now an epidemic, which at the end of 2003 had claimed an estimated 2.9 million lives. Epidemiology examines the role of host, agent and environment to explain the incidence and transmission of disease. Risk factor epidemiology examines the individual (demographic and social) characteristics of individuals and attempts to determine the factors that place an individual at risk of acquiring a disease [1]. In this study, the demographic and social characteristics of individuals and their behaviour are used to determine the risk of HIV infection; this is referred to as "biomedical individualism" [1, 2]. The prevalence of infectious diseases is dependant on the nature of the disease transmission. HIV is primarily transmitted sexually, hence the HIV status in one person is dependant on that of others as well as exposures to other individuals. Social factors therefore affect the risk of exposure as well as the probability of transmission of the disease and are necessary to understand and model the disease. By identifying the individual risk factors that lead to the disease, it is possible to modify social conditions which give rise to these factors, and thus design effective HIV intervention policies [1].

Artificial intelligence or machine learning techniques have been used successfully in medical informatics for decision making, clinical diagnosis, prognosis, and prediction of outcomes. Neural networks are capable of non-linear pattern recognition without the need for an exact model. When applied to classification, neural networks are, firstly, used to discover which characteristics or combinations of characteristics are useful for distinguishing between classes. In order to do this, domain knowledge of the application is required. The second objective of a pattern classification system is to find a separator that will divide the classes, placing as many samples into the correct classes as possible [3]. The final objective is to use the pattern classification system to classify new unseen cases. Based on the work that has been previously carried out in the field of HIV classification (Section II and III), the objective of this study is as follows; Multilayer Perceptron (MLP) neural networks are trained to classify the HIV status of individuals by providing them with demographic factors, social and behavioral characteristics of the individuals. Preliminary work carried out by the authors of this paper, showed that compared to the Radial Basis Functions (RBF) neural networks, the MLP outperformed the RBF with respect to accuracy. The experimental data was obtained from antenatal seroprevalence surveys conducted in South Africa.

Section II provides a background on Biomedical Individualism; meanwhile Section III gives an overview on the use of Artificial Intelligence in HIV/AIDS predictions. Section IV provides a background on the neural network architecture used in this study, MLP. Classification performance measures are explained in Section V. The methodology used to train and assess the networks is explained in Section VII and the results obtained are presented in Section VIII.

II. BIOMEDICAL INDIVIDUALISM

Biomedical individualism [1, 2] is defined as the basis of risk factor epidemiology and this is different from social
epidemiology in that, in social epidemiology, social conditions are considered as the fundamental causes of diseases meanwhile in biomedical individualism, demographic and behavioral characteristics are considered. Poundstone et al. [1] related the demographic properties and their effect on HIV. They related demographic and social characteristics such as structural violence and discrimination, Race/ethnicity and racism [4], stigma and collective denial, legal structures, demographic change and policy environment, to the spread of the HIV. This justifies the use of such socio-demographic parameters in creating a model to predict the HIV status of individuals. The risk of contracting the virus can hence be obtained from the socio-demographic characteristics. Wolters distinguishes biomedical and developmental paradigms. In the paper, the risk is conceptualized and the significant increase in the use of the biomedical paradigm is also commented on. The use of the biomedical paradigm has stemmed mainly from the reports about effective treatments of AIDS and hope of development of vaccines [5]. Epidemiology has however been criticized for focusing on individual disease risk factors. Shy [6] suggested that academic epidemiology has served clinical medicine due to the narrow biomedical perspective, dealing with risk factors and disease associations, rather than contributing to the population understanding the disease patterns. The lack of social, economic, environmental and political analysis has been used as the criticism for biomedical individualism and this criticism of its individualism has led to calls for a new public health that sees understanding the social, environmental, and economic determinants of health as crucial. Takano et al [7] looks beyond individual risk factors to features of environments and locations that affect health. This was done through a study of longevity in senior citizens that found a correlation between longevity (probability of five year survival), whether they reported they had space for taking a stroll near their residence, parks, and tree lined streets near their residence. Their findings demonstrate how epidemiological methods can be adapted to research the structural factors that affect people’s health; they suggest that exercise patterns reflect the environments in which people live; and they contribute to an evidence base for health promotion initiatives based on settings. Rabbisc comments on the need for social scientists to study social constructions of HIV/AIDS and identify the social and cultural determinants of HIV/AIDS occurrence [8]. Our study presents a methodology which identifies such determinants by using the demographic properties to create a model. Rhodes and Simic [9], in their paper, comment on the implications of HIV prevention. They mention that strategies focus on ameliorating conditions underpinning increased HIV risk, such as policy change. Srisawat and Kijsirikul [10] proposed a technique using autoassociative classification for predicting HIV-1 drug resistance. The best model provided an accuracy between 84.11% and 92.64%. Rauner and Brandeau [11] discuss key issues relevant to developing AIDS policy models – and thus an improved AIDS policy making process and more informed decisions. In their paper the necessity of target groups and intervention is commented on. Our methodology aims at creating models that analyze demography thus analyzing groups, hence more effective control can be put for the different demography. They also state that, in order to design effective HIV prevention programs, it is important to understand how HIV spreads in different geographical areas.

Our model is focused on understanding the spread from demographic properties of which area is one. The novelty in this design is the application of genetic algorithms in the HIV prediction from demographic data. Literature review showed that genetic algorithms have not been applied in HIV modeling, even though they have been applied to the field of data mining. Our aim is thus to investigate whether genetic algorithms will increase the prediction accuracy.

### III. ARTIFICIAL INTELLIGENCE IN HIV/AIDS PREDICTIONS

Artificial neural networks (ANNs) have been used to classify and predict the status of HIV/AIDS patients from symptoms [12]. The data used were all the complete entries from a publicly available AIDS Cost and Services Utilization Survey performed in the United States of America. Multilayer perceptron architecture, with 15 linear inputs and 3 hidden logistic nodes and one output, being the HIV status or AIDS status, was trained using 200 epochs with a learning rate of 0.1 and momentum of 0.1. 1026 cases were used for training and 667 HIV cases were used for testing. The best accuracy obtained was 587 correct. A study was also performed to predict the functional health status of HIV and AIDS patients defined as well or not well, using neural networks [13]. The other applications of neural networks in AIDS research have been in bioinformatics where modeling of the virus has been done on a molecular level, such as the prediction of HIV-1 Protease Cleavage Sites. The above models concluded that ANN performed well in pattern recognition and signal processing. The methodology presented here aims at using other demographic and social factors, to predict the status of an individual.

### IV. MULTI LAYER PERCEPTRON

Multilayer perceptrons (MLPs) are feedforward neural networks. They are supervised networks, so they require a desired response to be trained. They learn how to transform input data into a desired response, so they are widely used for pattern classification. With one or more hidden layers, they can approximate virtually any input-output map. MLP’s are probably the most widely used architecture for practical applications. [14]

The network can be described as follows:

\[
y_k = f_{outer} \left( \sum_{j=1}^{M} w_{kj}^{(2)} f_{inner} \left( \sum_{i=1}^{d} w_{ij}^{(1)} x_i + w_{i0}^{(1)} \right) + w_{k0}^{(2)} \right).
\]

Where \( y_k \) represents the k-th output, \( f_{outer} \) represents the output layer transfer function, \( f_{inner} \) represents the input layer transfer function, \( w \) represents the weights and biases, \( w_{i0}^{(1)} \) represent the i-th layer.

The linear activation function will be used for the output and the hyperbolic tangent function will be used in the hidden layers [14]. Because of its efficiency, the scaled conjugate
gradient method is used as the optimization technique used to train the networks. For a two class classifier one output node is sufficient, so there is only one activation at the second layer. The outputs from the hidden layer are connected via weighted connections to the output node and biased to form the second layer activation:

\[ a_j^{(2)} = \sum_{j=1}^{n} w_j^{(2)} \phi_j + b^{(2)} \]  

(2)

This second layer activation is transformed by the logistic output activation function, as it operates in the range 0 to 1, and it allows the output to be given a probabilistic interpretation, since it is derived using Bayes theorem to represent the posterior probabilities of membership to classes. Additionally, the sigmoidal function is able to represent both non-linear functions as well as linear functions.

\[ y = \frac{1}{1 + e^{-a(2)}} \]  

(3)

The output \( y \), is a continuous scalar bounded between 0 and 1, thus to use \( y \) as the indicator of class membership it needs to be converted to binary values using a threshold. Since the resultant model is non-linear, when applied to classification, the decision boundary between the classes produced by the network is also non-linear. This is an advantage over most other classification methods such as trees which have linear decision boundaries. Non-linearity allows for highly flexible decision surface shapes, but since non-linear estimation of the parameters is not straightforward, iterative techniques are used (training). The MLP network is then trained to find the weights, biases.

V. QUALITY OF CLASSIFICATION

A. The confusion matrix

The mean square error (MSE) is insufficient as a classification accuracy measure, as it indicates only the total number of correct classifications. In medical diagnosis in particular, it is necessary for a more detailed accuracy analysis, including the number of false positives, false negatives, true positives and true negatives. The confusion matrix shows the cross-classification of the predicted class against the true class. By splitting misclassifications into the different cells of the matrix, it is possible to assign a cost of making that particular misclassification [15]. The confusion matrix is given in (4).

\[ C = \begin{pmatrix} TN & FP \\ FN & TP \end{pmatrix} \]  

(4)

Where: \( TN \) = True Negatives (where network predicts an HIV negative person as negative), \( FP \) = False Positives (where network predicts an HIV negative person as positive), \( FN \) = False Negatives (where network predicts an HIV positive person as negative) and \( TP \) = True Positives (where network predicts an HIV positive person as positive).

The rows represent the true classes and the columns represent the predicted classes. The ideal solution has no false positives, nor false negatives, so the diagonal entries are at a maximum. Usually, as is true for this case, the cost of misclassification is difficult to determine. Using the quantities in the confusion matrix, it is possible to derive the Receiver Operating Characteristic or ROC curve. It is also possible to get the accuracy for the measurements from the confusion matrix which will be used to qualify the network and the results obtained.

B. Receiver Operating Characteristic and Accuracy

The neural network classifiers produce a continuous output indicative of the probability that the element belongs to a class. A threshold is applied to convert this output to predict class membership, and the value of the threshold affects performance. For an instance and a classifier there are four possible outcomes: true positive, where the instance is positive and is classified as positive; true negative, where the instance is negative and is classified as negative; false positive, where the instance is negative but is classified as positive; and false negative, where the instance is positive but is classified as negative. These outcomes are often summarized in a confusion matrix, where the entries along the major diagonal represent correct decisions, and the entries off the diagonal are the errors. Other quantities are derived from the possible outcomes. The True Positive Rate (hit rate or sensitivity) is defined in (6), and the False Positive Rate (false alarm rate), or specificity is defined in (7).

\[ TruePositiveRate = \frac{TP}{TP + FN} \]  

(6)

\[ FalsePositiveRate = \frac{FP}{FP + TN} \]  

(7)

The True Positive Rate [15, 16] is plotted against the False Positive Ratio for different threshold values. The accuracy in general is the number of correctly classified out of the total number of cases. The accuracy is obtained as follows:

\[ Accuracy = \frac{TN + TP}{TN + FN + TP + FP} \]  

(8)

1) Model Selection: Plotting the True Positive Ratio on the y-axis against the False Positive Ratio on the x-axis results in the ROC curve, which depicts the trade-offs between true positives and the costs (false positives). Perfect classification occurs at the point (0,1) on the ROC space, while (0,0) indicates that the classifier never issues positive classifications, and (1,1) represents a classifier always issuing positive classification [10]. The threshold can therefore be selected according to the misclassification costs: if a classification should only be made if there is strong evidence, then a classifier in the lower left hand side should be selected. Conversely, if the aim is for the classifier to be sensitive to possible positive cases, the upper right hand corner shows the classifiers that make positive classifications even if evidence is low.

VI. IMPLEMENTATION

A. Data Processing

1) Data Source: Demographic and medical data came from the South African antenatal seroprevalence survey of 2001. This is a national survey, and any pregnant women attending selected public health care clinics participating for the first
time in the survey were eligible to participate. Anonymity is guaranteed. The antenatal seroprevalence surveys are used as the main source of HIV prevalence data worldwide, reasons for this are that antenatal clinics are found throughout the world, and pregnant women are ideal candidates for the study as they are sexually active.

2) Missing Data: Out of the total data set cases, 12945 complete cases were selected, out of 13087 cases (98.91%) and the incomplete entries (142 cases – 1.09%) were discarded.

3) Variables: The variables obtained in the study are: race, region, age of the mother, age of the father, education level of the mother, gravidity, parity, province of origin, race, region of origin and HIV status [17]. The qualitative variables such as race and region are converted to integer values. The age of mother and father are represented in years. The integer value representing education level represents the highest grade successfully completed, with 13 representing tertiary education. Gravidity is the number of pregnancies, complete or incomplete, experienced by a female, and this variable is represented by an integer between 0 and 11. Parity is the number of times the individual has given birth, (for example, multiple births are counted as one) and this is not the same as gravidity. Both these quantities are important, as they show the reproductive activity as well as the reproductive health state of the women. The HIV status is binary coded; a 1 represents positive status, while a 0 represents negative status. Thus the final number of input variables is 10, shown in Table I. There is one output.

4) Outliers: Age is the only variable with outliers. The standard age bracket used in demographic studies relating to female fertility is 14-50 in African countries, and this was used to extract outliers in mother’s age.

5) Dataset Used: The dataset was divided into three sets; training, validation and testing sets. The sets were created by dividing the huge dataset into three equivalent small datasets of 1988 entries each. The inputs used were; age of female, age gap, educational level of female, gravidity, parity, province of origin, race, and region of origin. The training set is balanced to consist of an equal number of positive outcomes as negatives, by duplicating the positive entries. An alternative to oversampling the minority class is to assign distinct costs to training examples, or by undersampling the majority class [18]. Due to the limited size of the dataset, over sampling the positive cases was used rather than undersampling the negative cases. The original training set consisted of more negatives than positives with a ratio of 3:1. If the neural network had been trained on this biased dataset, the predicted outcome would always have been negative. This data was randomized and the inputs were scaled between 0 and 1. This accounted for the biasing of the data set.

B. Neural Network Architecture Design

1) Training: The scaled conjugate gradient optimization technique is used in error back-propagation to train the networks.

2) Number of Hidden Nodes: With the number of input nodes for the primary RBF and MLP networks fixed at 1 and the input nodes for the secondary MLP network fixed at 9.

VII. MLP NETWORK IMPLEMENTATION

The NETLAB toolbox [20] was used to create and train the MLP architecture. Three data sets were used for training, validation and testing. Also, since the determination of network architecture involves the optimization of the number of hidden nodes, it is incorrect to use the testing set to compare results before determining the final number of hidden nodes, and thus the final architecture. Genetic algorithm (GA) was used to optimize the network (hidden nodes, and training cycles). The network implemented comprised of 9 primary RBF neural networks each accepting a different demographic input and mapping that to the output (HIV). The outputs of these 9 primary RBF networks are then

<table>
<thead>
<tr>
<th>Variable</th>
<th>Type</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age Group</td>
<td>Integer</td>
<td>14-50</td>
</tr>
<tr>
<td>Age Gap</td>
<td>Integer</td>
<td>1-7</td>
</tr>
<tr>
<td>Education</td>
<td>Integer</td>
<td>0-13</td>
</tr>
<tr>
<td>Gravidity</td>
<td>Integer</td>
<td>0-11</td>
</tr>
<tr>
<td>Parity</td>
<td>Integer</td>
<td>0-40</td>
</tr>
<tr>
<td>Province</td>
<td>Integer</td>
<td>0-19</td>
</tr>
<tr>
<td>Race</td>
<td>Integer</td>
<td>1-5</td>
</tr>
<tr>
<td>Region</td>
<td>Integer</td>
<td>1-36</td>
</tr>
<tr>
<td>RPR</td>
<td>Integer</td>
<td>0-2</td>
</tr>
<tr>
<td>WT.REV</td>
<td>Continuous</td>
<td>0.638 - 1.2743</td>
</tr>
<tr>
<td>Output Variable</td>
<td>Continuous</td>
<td>[0,1]</td>
</tr>
</tbody>
</table>

The upper limit on the number of hidden nodes was 110, using a factor of 2 for the weights to data point’s ratio. Several networks of differing complexity between 2 and 110 hidden nodes were trained with a regularization coefficient of between 0.001 and 1. The number training cycles ranged from 100 to 1000. Genetic algorithm [19] was then used to optimize the number of hidden nodes, the regularization coefficient and the number of training cycles. The fitness criteria used was the validation accuracy. The optimized network from the genetic algorithm had 77 hidden nodes, a regularization coefficient of 0.24693 and required 144 training cycles.

C Threshold adjustment

The output from the neural network is a continuous value between 0 and 1. A threshold was thus needed to convert the network output to a binary value representing the predicted HIV status. This was achieved by rounding the output to 1 if greater than a threshold and rounding it to 0 otherwise. The acceptance of the threshold was based on the accuracy obtained from the training and validation data sets. This threshold value was adjusted until an optimal value for the accuracy was obtained. Threshold values of between 0 and 1 were tested and the optimal value yielded by the genetic algorithm was 0.80107. The best performing MLP network had a testing accuracy of 84.24%.
fed into a secondary MLP network and they are related again to the output. The primary RBF networks thus contained one input node, three hidden nodes and one output node. The fitness function utilized by the GA was the accuracy obtained by the validation data set. The secondary MLP network was composed of 9 input nodes, 77 hidden nodes (optimal) and 1 output node. A threshold value of 0.80107 was used since as this was found as the most optimal threshold during training and validation. This threshold was used subsequently used for the test data set. Our work focused on using GA as the network parameter optimization method rather than line search methods since as preliminary research showed that the GA outperformed the line search methods.

VIII. RESULTS

The performance analysis is based on classification accuracy and training times. The most optimal network was the MLP network with 10 primary MLP networks, and 1 secondary MLP network. The optimal number of hidden nodes for the primary networks was 3; hence the structure was a 1-3-1 structure. The optimal number of hidden nodes for the secondary network was 20 hence the structure was 9-77-1. This network combination gave an accuracy of 88.66% during training time and an accuracy of 84.24% on the test data sets. The accepted threshold value which gave the optimal value was 0.80107. The training time was 352.8s. The confusion matrix obtained for the above network is as shown in Table II. The accuracy of the above network for the data sets calculated using equation (8) is as stated above, that is, 88.66% for the training data set and 84.24% for the test data set. The ROC curve for the classifier is shown in Fig. 1.

Table II

<table>
<thead>
<tr>
<th>Confusion Matrix</th>
<th>Predicted Pos</th>
<th>Predicted Neg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training Data Set</td>
<td>993</td>
<td>0</td>
</tr>
<tr>
<td>Test Data Set</td>
<td>424</td>
<td>570</td>
</tr>
</tbody>
</table>

Table III

<table>
<thead>
<tr>
<th>No. Of Hidden Nodes</th>
<th>No. Of Training Cycles</th>
<th>Accuracy</th>
<th>Training Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>100</td>
<td>0.78561</td>
<td>4.156</td>
</tr>
<tr>
<td>30</td>
<td>100</td>
<td>0.78661</td>
<td>6.609</td>
</tr>
<tr>
<td>40</td>
<td>100</td>
<td>0.7846</td>
<td>8.312</td>
</tr>
<tr>
<td>50</td>
<td>100</td>
<td>0.78661</td>
<td>10.655</td>
</tr>
<tr>
<td>60</td>
<td>100</td>
<td>0.7846</td>
<td>12.515</td>
</tr>
<tr>
<td>70</td>
<td>100</td>
<td>0.7846</td>
<td>14.485</td>
</tr>
<tr>
<td>77</td>
<td>300</td>
<td>0.78661</td>
<td>43.941</td>
</tr>
<tr>
<td>77</td>
<td>500</td>
<td>0.78661</td>
<td>78.032</td>
</tr>
<tr>
<td>77</td>
<td>700</td>
<td>0.78661</td>
<td>110.69</td>
</tr>
<tr>
<td>77</td>
<td>900</td>
<td>0.78661</td>
<td>131.63</td>
</tr>
</tbody>
</table>

A. Bayesian training of neural networks

Different permutations of the number of hidden nodes and training cycles were simulated and the error assessed to determine the optimal parameters. A summary of the average for 5 runs of each configuration is presented in Table III. The accuracy is used as a means of comparison for the different networks. For a calibrated classifier, the threshold is 1. Calibration is done by calculating the accuracy for each point and returning the threshold with maximum accuracy.

IX. CONCLUSION

Artificial intelligence methods can be used for classification. In this paper, supervised learning was used to train neural networks to classify the HIV status of an individual, given certain demographic factors. The preliminary stage of the design proved that the multi layer perceptron networks were more accurate than the other possible architecture, the radial basis function, thus this was used as the optimal design architecture. In order to avoid over-fitting, a validation set is used during training to optimize the design of the network. Optimizing the design involves setting the regularization parameters and the complexity of the network, set by the number of hidden nodes.

The best network is chosen according to the lowest standard error between targets and predicted outputs. The design process is however time consuming, since there are many permutations of parameters. To cater for this optimization stage, genetic algorithms were used in the design phase. The number of hidden nodes was chosen arbitrarily and was then optimized. It was found that the optimal number of hidden nodes for the primary RBF and MLP networks was 3. The optimal number of hidden neurons obtained from the genetic algorithm was 77 neurons for the secondary MLP network thus forming a structure 9-77-1 (with 9 inputs, 77 hidden nodes and 1 output node). The optimal number of training cycles was also optimized. This
value was found as 144. The optimal regularization parameter was 0.24693.

Performance metrics such as the accuracy are used to analyze the classification. Design of classifiers involves setting a threshold value to convert a probabilistic output to an actual classification. This value was determined by using genetic algorithms, and in this study, was set to the value that gives the best accuracy value, 0.80107. The results show that the MLP neural network architecture was quite efficient in predicting the HIV status from demographic properties to 84.24% accuracy. Demographic data are thus sufficient to accurately predict HIV status.

Draghici and Potter [22] did a study on predicting HIV drug resistance using neural networks for classification and correctly classified unseen data with an accuracy of between 60% and 70%. The method employed in this design thus significantly increased on the accuracy.

A preliminary methodology developed by the authors of this paper using multilayer perceptron networks with gradient optimization methods yielded accuracies of between 70% and 74%. Genetic algorithms thus increase the accuracy of the classification. Further parameters such as; the financial standing of individuals, a more elaborate data set which spans the entire population and the inter-mobility level of individuals, will enhance the prediction and classification and possibly increase the accuracy significantly. It is recommended that different input features be tested, as well as automatic relevance detection to assess which inputs contribute to the output.

REFERENCES