Autoencoder networks for HIV classification

Brain Leke Betchuoh*, Tshilidzi Marwala and Thando Tettey

In this paper, we introduce a new method to analyse HIV using a combination of autoencoder networks and genetic algorithms. The proposed model is tested on a set of demographic properties of individuals obtained from the South African antenatal survey. When compared to conventional feedforward neural networks, the autoencoder network classifier model proposed yields an accuracy of 92%, compared to an accuracy of 84% obtained from the conventional feedforward neural network models. The area under the ROC curve for the proposed autoencoder network model is 0.86 compared to an area under the curve of 0.8 for the conventional feedforward neural network model. The autoencoder network model for HIV classification, proposed in this paper, thus outperforms the conventional feedforward neural network models and is a much better classifier.

Keywords: Autoencoder networks, genetic algorithms, HIV classification.

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. Risk factor epidemiology examines the individual demographic and social characteristics and attempts to determine factors that place an individual at risk of acquiring a life-threatening disease. In this study, the demographic and social characteristics of the individuals and their behaviour are used to determine the risk of HIV infection; referred to as biomedical individualism. By identifying the individual risk factors that lead to the HIV infection, it is possible to modify social conditions, which give rise to the disease, and thus design effective HIV prevention policies. A model will be created and used to classify the HIV status of individuals based on demographic properties. In this study, the model is created using autoencoder neural networks and genetic algorithms, which have been applied to classification.

An artificial neural network (ANN) is an inter-connected structure of processing elements. The ANN structure used in this study consists of three main components (Figure 1). Neural networks have been successfully used for medical informatics, for decision making, clinical diagnosis, prognosis, and prediction of outcomes and for classification. Marwala used a probabilistic committee of neural networks to classify faults in a population of nominally identical cylindrical shells and obtained an accuracy of 95%, in classifying eight classes of fault cases. Ohno-Machado depicted the limitation on the accuracy of the neural network model due to lack of data balance and increased the accuracy by using sequential neural networks. Lisboa assessed the evidence of healthcare benefits using neural networks. Fernandez and Caballero used ANN to model the activity of cyclic urea HIV-1 protease inhibitors. They showed that ANN were capable of representing the nonlinearity in the HIV model. Lee and Park applied neural networks to classify and predict the symptomatic status of HIV/AIDS patients based on publicly available HIV/AIDS data. A study was also performed to predict the functional health status of HIV/AIDS patients defined as ‘in good health’ or ‘not in good health’, using neural networks. Laumann and Youm used the racial and ethnic group differences to model the prevalence of the disease and succeeded in relating the demographic properties to the transmission of the disease. Poundstone et al. related demographic properties to the spread of HIV. Their work justifies the use of such demographic properties in creating a model to predict the HIV status of individuals, as done in this study. The above models concluded that ANN performed well in HIV classification problems. The methodology presented here aims at using demographic and social factors, to predict the HIV status of an individual, using autoencoder neural networks.

The most common neural network architecture is the multilayer perceptron (MLP). An alternative network is the radial basis function (RBF). The use of MLP over RBF can be attributed to the fact that the RBF usually requires the implementation of the pseudo-inverse of a matrix for training, which is often singular while MLP uses conventional feedforward optimization methods, which are stable. In our study, preliminary design showed that the MLP outperformed the RBF. This can be attributed to the fact that MLP networks, also known as universal approximators, are capable of modelling any complex relationship with one or two hidden layers and are thus most suited for this study. More details on neural networks and
MLP can be found in refs 18–22. In this study neural networks are used with genetic algorithms.

A genetic algorithm (GA) is an optimization method deriving its behaviour from processes of evolution in nature, inspired by Darwin’s theory of natural evolution\(^{23,24}\). This is done by the creation within a machine/computer of a population of individuals. In this study, the population of individuals represents the missing input entries. The individuals then go through the process of evolution. GA uses fitness-proportionate or tournament selection to select the missing entries (individuals) probabilistically that yields the right HIV status for the individuals. Although not guaranteed to provide the globally optimum solution, GA has been shown to be highly efficient at reaching a very near optimum solution in a computationally efficient manner\(^{23,24}\). More details on GA can be found in refs 25 and 26.

In the literature review, there is no method proposed thus far that investigates the use of autoencoder networks for HIV modelling. The aim of this paper will thus be to propose a new method, which is based on autoassociative models\(^{27}\) combined with GA to classify the HIV status of an individual based on demographic properties. The proposed method is tested on the classification of the HIV status of individuals using a data set obtained from the South African antenatal seroprevalence survey. The method is then compared with conventional feedforward neural network models that have already been applied in the HIV modelling problem as presented in the literature review.

**Background**

*Autoassociative networks*

Autoassociative networks are models where the network is trained to recall the inputs\(^{27}\). This network thus predicts the inputs as outputs, whenever an input is presented. These networks have been used in a number of applications\(^{28–31}\). An autoassociative neural network encoder (or simply known as autoencoder) consists of an input and output layer with the same number of inputs and outputs, hence the name autoassociative, combined with a narrow hidden layer\(^{27}\). The networks will be trained using HIV/AIDS demographic data. The hidden layer attempts to reconstruct the inputs to match the outputs, by minimizing the error between the inputs and the outputs when new data is presented. The narrow hidden layer forces the network to reduce any redundancies, but still allows the network to detect non-redundant data. However, it must be noted that for missing data estimation it is absolutely crucial that the network must be as accurate as possible and that this accuracy is not necessarily realized through few hidden nodes as is the case when these networks are used for data compression. It is therefore crucial that some process of identifying the optimal architecture be used. GA is used in this study to find the optimal autoencoder architecture by finding the global optimum solution\(^{23}\). The auto-encoder neural network architecture used in this study is shown in Figure 2.

**Classification as a statistical pattern**

The goal of our classification is to develop an algorithm, which will assign an individual, represented by a vector \(\{x\}\) describing the demographic, social and behavioural characteristics of that individual, to one of the HIV classes, \(C_1\) or \(C_2\) (where \(C_1, C_2\) represents the status of an individual, which may be positive or negative). The data on which the model is based upon contains demographic examples of individuals, as well as the classes to which those individuals belong. The output of the classification system is assigned to the variable \(y\). The classification
model is therefore required to map the inputs $x_1, \ldots, x_d$ to the output $y$. A mathematical function describes this mapping, and since it cannot be explicitly determined, the data is used to determine the parameters. This can be written as follows:

$$[y] = f([x], [w]).$$

(1)

Here $[w]$ is the mapping weights and $[x]$ represents the demographic input parameters and $[y]$ represents the HIV status. In this study, autoencoder neural networks are used to obtain the functional mapping, and supervised learning is used to obtain the parameters. The purpose of the classification model is to design the decision surface to assign new inputs to one of the classes.

**Methodology**

The literature review showed that models for HIV prediction and classification have been developed using conventional feedforward neural networks architectures and have worked well. However, it was found from the literature review that autoencoder networks have not been applied to HIV modelling, for prediction and classification. Our work thus focuses on proposing a methodology for HIV classification from demographic properties using autoencoder neural networks and GA. Our work also focuses on comparing the proposed autoencoder method to a conventional feedforward neural networks model, by creating a feedforward MLP neural network model and comparing the results with the autoencoder network model results.

**HIV classification using autoencoder networks**

The NETLAB toolbox was used to create and train an autoencoder MLP architecture. This toolbox has a 2-layer MLP network, which according to literature review is capable of modelling any complex relationship, such as the HIV model. The network implemented consisted of an input layer, representing different demographic inputs and the HIV status, mapped to an output layer representing the same characteristics as the input layer via the hidden layer. The network was thus trained to recall itself (predict the demographic inputs). This network is shown in Figure 2. One of the input nodes in Figure 2, $x_5$, represented the HIV status of individuals, which was ultimately represented by one of the output nodes, $y_2$, as well. The neural network equation can be written as in eq. (1).

Since the network is trained to recall the demographic inputs, the output vector $[y]$ (predicted demographic properties) obtained will be approximately equal to the input vector $[x]$ (actual demographic properties). An error, however, exists between the input vector $[x]$ and the output vector $[y]$, which can be expressed as the difference between the input and output vector. This error is formulated as

$$e = [x] - [y].$$

(2)

Substituting for $[y]$ from eq. (1) into eq. (2) we get

$$e = [x] - f([x], [w]).$$

(3)

In our work, a minimum and non-negative error is required. This can be obtained by squaring the error function in eq. (3) to obtain

$$e = ([x] - f([x], [w]))^2.$$  

(4)

To predict the HIV status of individuals, the HIV status input, in the input vector $[x]$ was assumed as an unknown input, while the demographic input properties were considered as the known inputs. When the input vector $[x]$ has unknown elements, the input vector set can be categorized into $[x]$ known represented by $[x_k]$ and $[x]$ unknown represented by $[x_u]$. Rewriting (4) in terms of $[x_k]$ and $[x_u]$, we obtain

$$e = \left(\sum_k x_k - f\left(\begin{array}{c} x_k \\ x_u \end{array}, [w] \right)\right)^2.$$  

(5)

Here $[x_u]$ represents the HIV status of the individual, which is unknown. $[x_k]$ represents the demographic input parameters of the individuals in Table 1, $[w]$ represents the weight vector that maps the autoencoder network input vector $[x]$ to the same input vector $[x]$.

An estimated value for the HIV status is then obtained by minimizing eq. (5) using a GA which was chosen because it finds the global optimum solution. GA, however, always finds the maximum value. To cater for this, the negative of eq. (5) was used as the fitness function for the GA. The error function to be minimized is thus
Table 1. Summary of input and output variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Type</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input variables</td>
<td></td>
<td></td>
</tr>
<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>Gravity</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>1–9</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>WTREV</td>
<td>Continuous</td>
<td>0.638–1.2743</td>
</tr>
<tr>
<td>Output variable</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HIV status</td>
<td>Binary</td>
<td>[0, 1]</td>
</tr>
</tbody>
</table>

\[
e = \left( \frac{x_0}{x_k} - f\left( \left[ \frac{x_0}{x_k} \right], \{w\} \right) \right)^2.
\] (6)

This estimated value from the autoencoder network and genetic algorithm was a continuous value representing the HIV status. A threshold was thus required to convert the HIV output node value to a binary value, representative of the HIV class of the individual. Figure 3 shows the implementation of this proposed model in a flowchart.

**HIV classification using neural networks**

In this model, the NETLAB toolbox\textsuperscript{32} was used to create and train an MLP neural network architecture. The network implemented consisted of an input layer, representing different demographic inputs of an individual, mapped to an output layer representing the HIV status of an individual via the hidden layer. The network thus mapped the demographic inputs of individuals to the HIV status, This network is shown in Figure 1. The neural network equation can be written as in eq. (1). In this model, however, the output vector \( \{y\} \) represents the HIV status of the individual. The network is thus trained to find the relationship between the HIV status of the individual and the individual's demographic input properties. An error, however, exists between the individual's predicted HIV status (output vector) \( \{y\} \) and the individual's actual HIV status (target vector) \( \{t\} \) during training, which can be expressed
as the difference between the target and output vector. For the neural network HIV classification, the mean square error function between the target output vector \{t\} and the output vector \{y\} is insufficient as a classification accuracy measure, as it only indicates the total number of correct classifications. A confusion matrix was thus constructed and the accuracy was obtained from the confusion matrix. The accuracy can be formulated as
\[
\text{Accuracy} = \frac{\text{TN} + \text{TP}}{\text{TN} + \text{FN} + \text{TP} + \text{FP}}.
\]

Here 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 accuracy function was then used as the fitness function in the GA to obtain the optimal neural network parameters. GA was used as it finds the maximum value of the fitness function, which was required in this case. GA was also used to obtain the threshold value to convert the continuous network output to a binary value representative of HIV.

Results and discussion

The demographic and medical data, used in this study, came from the South African antenatal seroprevalence survey\textsuperscript{13} of 2001. This is a national survey, and pregnant women attending selected public health care clinics participating for the first time in the survey were eligible. The variables obtained are shown in Table 1. These include: age of mother, age of partner, educational level of mother, gravidity (number of complete or incomplete pregnancies), parity (number of complete pregnancies), province of origin, race of mother, and region of origin. The qualitative variables such as the province of origin, race of mother and region of origin were encoded to integers. For example, the encoding scheme for race is shown in Table 2. The HIV status was encoded using an integer scheme, whereby a 1 represents a positive HIV status meanwhile a 0 represents a negative HIV status. The parameter distributions are also listed in Table 1. A total of 1986 training inputs were provided for the network. The GA used for the autoencoder network model proposed in this study and the neural network model used arithmetic cross-over, non-uniform mutation and normalized geometric selection. The probability of cross-over was chosen to be 0.75 as proposed in Marwala et al.\textsuperscript{24}. The probability of mutation was chosen to be 0.0333 as recommended by Marwala et al.\textsuperscript{24}. GA had a population of 40 and was run for 150 generations.

The first experiment investigated the use of autoencoder networks for HIV classification. An autoencoder network with 9 inputs and 9 outputs was constructed and several number of hidden units were investigated, using Matlab\textsuperscript{5} (ref. 35). A GA was used to obtain the optimum number of hidden units and yielded an optimum number of hidden units of 2, hence the structure 9–2–9. Linear optimization using the mean square error versus hidden units was also investigated. As shown in Figure 4, the

![Figure 4. The prediction error versus the number of hidden nodes.](image1)

![Figure 5. ROC curve for the autoencoder network classifier.](image2)
linear optimization yielded 6 hidden units as the optimal network that gives the best prediction since as the error
does not change significantly from 6 units onwards (the
difference in error is about 8.5% from 6 hidden units to
20 hidden units). It must be noted, however, that it is
generally assumed that the best autoencoder network is
the one that has the lowest possible number of hidden
units. A hidden unit of 2 was thus used as the optimal
autoencoder network number of hidden units. The perform-
ance analysis for the autoencoder network model is
based on classification accuracy and the area under the
ROC curve. The proposed autoencoder network model
obtained an HIV classification accuracy of 92%. The con-
fusion matrix obtained for the above network is shown
in Table 3. The ROC curve for this classification is shown
in Figure 5 and the area under the curve was computed as
0.86, thus giving a very good classifier according to ROC
curves documentation.

The second experiment investigated the use of conven-
tional feedforward neural network MLP architecture to
classify the HIV status of an individual using the demo-
graphic input properties. The MLP was constructed with
9 inputs and 1 output. A GA was then used to obtain the
optimal structure and yielded an optimal number of hidden
units of 77, hence the structure was 9-77-1. The perform-
ance analysis for this network model is also based on
classification accuracy and the area under the ROC curve.
This network gave an accuracy of 84%. The confusion
matrix obtained for the above network is shown in Table 4.
The ROC curve obtained for this classification is shown
in Figure 6 and the area under this ROC curve obtained
was 0.8, which according to ROC curves documentation
is a very good classifier.

The reason why autoencoder networks performed better
than the conventional feedforward neural network can be
attributed to the fact that the autoencoder network focuses
on characterizing the positive classes independently of the
negative classes, whereas the conventional feedforward
neural networks may overlook under-represented classes.
We hypothesize that this may be due to lower effective
dimension of the autoencoder network classifier. The dif-
fERENCE in performance can also be attributed to the fact
that in the autoencoder network, classification is done by
choosing the best fitting model using probability distributions.
The class of the network with the smallest recon-
struction error is selected. Conventional feedforward
neural networks on the other hand just map an input vector
to an output vector using scenario and encodes the classes
directly. This plays a role because, for nonlinear models
such as the HIV model, it is usually difficult to compute the
derivatives for the scenarios since they require that we
integrate all the possible representations that could have
been used for each particular observed input vector. The
distance measure in classification is thus better minimized
in the autoencoder network than in the conventional feed-
forward network model.

| Table 3. Classifier confusion matrix of the autoencoder network classifier |
|---------------------------|---------------------|
| Confusion matrix | Predicted positive | Predicted negative |
|-------------------|---------------------|
| Actual positive   | 899                 | 94                 |
| Actual negative   | 65                  | 928                |

| Table 4. Classifier confusion matrix of conventional feedforward neural network classifier |
|-----------------------------|---------------------|
| Confusion matrix | Predicted positive | Predicted negative |
|-------------------|---------------------|
| Actual positive   | 680                 | 313                |
| Actual negative   | 0                   | 993                |

Figure 6. ROC curve for the conventional feedforward neural network classifier.

Conclusion

A method based on autoassociative neural networks and
genetic algorithms is proposed to classify the HIV status
of an individual from demographic properties. This method
is proposed in order to investigate whether using autoen-
coder networks improves on the accuracy of classifica-
tion, of an individual’s HIV status, from demographic
properties. The proposed method is tested on an HIV data
set obtained from the South African antenatal seropreva-
ience survey of 2001. The method is then compared to a
conventional feedforward neural network model, imple-
mented using the MLP architecture. A classification ac-
curacy of 92% was obtained for the autoencoder network
compared to 84% obtained for the conventional feedfor-
ward neural network model implementation. The area un-
der the ROC curve for the autoencoder network classifier
was computed as 0.86 compared to 0.8 computed for the
conventional feedforward neural network classifier. The
result thus suggest that autoencoder network models are more accurate and better classifiers for the HIV model than conventional feedforward neural network models, since autoencoder networks focus on characterizing the positive classes independently of the negative classes, whereas the conventional feedforward neural networks may overlook under-represented classes.


ACKNOWLEDGEMENTS. We thank Mr Chris Lines, Mr David Starfield, Mr David Yanucci and the National Research Foundation for support.

Received 6 June 2006; revised accepted 30 August 2006

CURRENT SCIENCE, VOL. 91, NO. 11, 10 DECEMBER 2006 1473