

Personnel Audit Using a Forensic Mining Technique

Adesesan B. Adeyemo[1] and Oluwafemi Oriola[2]

**[1] Computer Science, University of Ibadan
Ibadan, Oyo State, Nigeria**

**[2] Computer Science, University of Ibadan
Ibadan, Oyo State, Nigeria**

Abstract
This paper applies forensic data mining to determine the true status of employees and thereafter provide useful evidences for proper administration of administrative rules in a Typical Nigerian Teaching Service. The conventional technique of personnel audit was studied and a new technique for personnel audit was modeled using Artificial Neural Networks and Decision Tree algorithms. Atwo-layer classifier architecture was modeled. The outcome of the experiment proved that Radial Basis Function Artificial Neural Network is better than Feed-forward Multilayer Perceptron in modeling of appointment and promotion audit in layer 1 while Logitboost Multiclass Alternating Decision Tree in Layer 2 is best in modeling suspicious appointment audit and abnormal promotion audit among the tested Decision Trees. The evidential rules derived from the decision trees for determining the suspicious appointment and abnormal promotion were also presented.
Keywords: *Data Mining, Forensic, Audit, Appointment, Promotion*

1. Introduction

By general definition, *audit* is the evaluation of a person, organization, system, process, enterprise, project or product. There are two types of audit namely: Financial Audit and Compliance Audit. In organizations, personnel are employed and assigned to various departments or units based on certain criteria which include skill, qualification, experience, and social/gender status. It is the desire and goal of every employer to secure and maintain a set of personnel that are legitimate within the consent/service condition of the institution and as such, verification and validation of employees is done severally through various means to determine the employment status of its staff. This is referred to as Personnel Audit. Personnel Audit is an example of Financial Audit. According to the Employment Equity, Planning and Policy Development Division of Personnel Policy Branch, Treasury Board in Canada, personnel audit is the systematic, independent review and appraisal of all departmental (personnel) operations, to determine the efficiency, effectiveness and economy of the departmental (personnel) management practices and controls. Personnel Audit entails verification and validation of the compliance of appointment, promotion, payment, background and job performance records. In this work, the focus will be on the audit of appointment and promotion records because of their significance and the challenges posed in Nigerian environment.

Appointment Audit refers to a process that involves verification and validation of employees to ascertain their compliance with principles guiding recruitment and assignment to jobs. In this aspect, the employees can either be legitimate or illegitimate (ghost or ineligible). Legitimate employees refer to group of employees whose appointments comply with rules and regulations that govern its implementation within an organization. Ghost employees refer to group of employees that are not in existence and fraudulently enjoying staff benefits while Ineligible employees refer to group of employees that are not competent or fit for appointment.

Promotion Audit is a process that involves verification and validation of employees to ascertain their compliance with principles guiding advancement and progress of employees through cadres or positions. In this aspect, the employee can be categorized as having a normal or abnormal promotion, being qualified for promotion or unqualified for a promotion. Abnormal promotion is a promotion that does not follow the due process and regulations. *Employees qualified for promotion* refers to employees whose promotion are indiscriminately impeded or delayed, while *employees unqualified for promotion* refer to employees that occupy positions that they are not yet entitled to.

Audit, in some cases can be Forensic in its approach depending on the level of duty or complexity. Forensic Science has been defined as the application of science and technology to the potential for evidence, at all stages of the investigative process, so that it can be located, recovered, analyzed and interpreted for the purpose of impacting on crime and criminality in a way that supports the effective administration of justice and inspires public confidence (Mennell, 2009). The primary objectives of forensic science include: Bringing offenders to justice; exonerating the innocent; detecting crime; ensuring efficient and effective investigations and gaining a better understanding of criminality, for example by understanding criminal behavior, links and associates via information provided through forensic data such as fingerprints, footwear marks, drug composition and tool marks. Forensic data mining which takes its roots from Forensic Science can be described as the application of data mining techniques and other scientific tools to investigative process for good and sound evidence (Chatterji, 2001). It often requires thorough IT knowledge of data matching and data mining techniques. This project intends to determine the authenticity of employees on the personnel and pay roll list of a Nigerian government agency using data mining techniques supported by forensic principles.

1.1 Forensic Data Mining

Data Mining refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases (Zaiane, 1999). It is a key step of knowledge discovery in databases (KDD). In other words, data mining involves the systematic analysis of large data sets using automated methods. By probing data in this manner, it is possible to prove or disprove existing hypotheses or ideas regarding data or information while discovering new or previously unknown information. It is noted for its Pattern Recognition ability that ensures that information is obtained from vague data (Baker, 2009). In particular, unique or valuable relationships between and within the data can be identified and used proactively to categorize or anticipate additional data. Through the use of exploratory graphics in combination with advanced statistics, machine learning tools, and artificial intelligence, critical "nuggets" of information can be mined from large repositories of data. While data mining has been applied to many areas of human endeavour such as business, education, manufacturing, and government, the application of data mining in personnel audit has not well been explored. In a report by Phua et al. (2004) on fraud detection based researches, it was shown that very few of these researches focused on employees' fraud.

Forensic Data Mining (or Forensic Mining) which originated from Forensic Science can be described as the application of data mining techniques and other scientific tools to investigative process for good and sound evidence. In the same vein, the investigative process of auditing called forensic auditing could be defined as the application of auditing skills to situations that have legal consequences (Chatterji, 2001). Forensic audit methodologies can be used to obtain a more detailed understanding of the entity and its activities to identify areas of risk both in determining the direction of the audit and in expressing an opinion. In forensic auditing, the following actions are important: working relations with the investigating and prosecuting agencies, authorisation and control of the audit investigation, documentation of relevant information and safeguarding all prime records pertaining to the case, rules of evidence governing admissibility or authentication of records, confidentiality, evaluation of the evidence to assess whether the case is sustainable, legal advice where appropriate and reporting the findings in a manner that meets legal requirements. As such, the knowledge of entity's business and legal environment, awareness of computer assisted audit procedures and innovative approach and sceptic of routine audit practices are required for forensic audit.

The audit of personnel of personnel who are mostly teachers, laboratory attendants, administrative workers and guards in State's Teaching Service is carried out regularly to ensure that qualified employees and accurate record of personnel are on the state government payroll list. This is handled by internal auditors who are employees of the state civil service commission and external auditors who are consultants who work within the government auditing standard. However, the technique being used has not yielded reliable results. In our case study which is State Teaching Service, the conventional physical appearance screening of personnel and manual methods of personnel's records checking which are being used are faced with the problems of erroneous results, lack of support for effective administration of justice and inability in generating useful patterns that could be documented for later personnel audits in the organization.

This poor result of audit of the State Teaching Service personnel has caused government a lot of resources in terms of training, payment of salary and other entitlements of illegitimate employees. Apart from this, a lot of wastage has been incurred, in terms of monetary waste and human waste during physical appearance screening of personnel due to poor method of processing and large volume of

ISSN (Online): 1694-0814
www.IJCSI.org

data being processed. Therefore the techniques that are employed in carrying out the audit has posed many challenges like waste on the part of government and stakeholders in form of time, effort and money; and loss of lives and negative consequences on health of employees during audit exercise.

In this research work, soft computing data mining methods (Artificial Neural Networks and Decision Trees) that are guided by forensic principles will be used to develop a model for personnel audit purpose using case data from one of the States Teaching Service Personnel Databases of in Nigeria. This will help to expose behavioural patterns of crime associated with appointment and promotion among employees in the Teaching Service, guide against human judgement that is distorted by an array of cognitive, perceptual and motivational biases applied during physical appearance screening of employees, reduce the anomalies in employees record, detect appointment and promotion fraud and give accurate staffing report in sufficient details to further allow accurate resizing and restructuring.

1.2 Decision Trees and Artificial Neural Networks

A Decision Tree (DT) is a logical model represented as a binary or multiclass tree that shows how the value of a target variable can be predicted by using the values of a set of predictor variables. In the tree structures, leaves represent classifications and branches represent conjunctions of features that lead to those classifications. In decision analysis, a decision tree can be used visually and explicitly to represent decisions and decision making. The concept of information gain is used to decide the splitting value at an internal node. The splitting value that would provide the most information gain is chosen. Formally, information gain is defined by entropy.

In other to improve the accuracy and generalization of classification and regression trees, various techniques were introduced like boosting and pruning. Boosting is a technique for improving the accuracy of a predictive function by applying the function repeatedly in a series and combining the output of each function with weighting so that the total error of the prediction is minimized or growing a number of independent trees in parallel and combine them after all the trees have been developed. Pruning is carried out on the tree to optimize the size of trees and thus reduce overfitting which is a problem in large, single-tree models where the model begins to fit noise in the data. When such a model is applied to data that was not used to build the model, the model will not be able to generalize. Many decision tree algorithms exist and these include: Alternating Decision Tree, Logitboost Alternating Decision Tree (LAD), C4.5 and Classification and Regression Tree (CART).

An Artificial Neural Network (ANN) is a mathematical model or computational model that tries to simulate the structure and/or functional aspects of biological neural networks. It consists of an interconnected group of artificial neurons and processes information using a connectionist approach to computation. In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase. Modern neural networks are non-linear statistical data modeling tools. They are usually used to model complex relationships between inputs and outputs or to find patterns in data. When creating a functional model of the biological neuron, there are three basic components of importance. First, the synapses of the neuron are modeled as weights. The strength of the connection between an input and a neuron is noted by the value of the weight. Negative weight values reflect inhibitory connections, while positive values designate excitatory connections. The next two components model the actual activity within the neuron cell. An adder sums up all the inputs modified by their respective weights. This activity is referred to as linear combination. Finally, an activation function controls the amplitude of the output of the neuron. An acceptable range of output is usually between 0 and 1, or -1 and 1. This process is illustrated diagrammatically in figure 1. From this model the interval activity of the neuron can be shown to be equal to an output function represented by

$$v_k = \sum_{j=1}^{p} w_{kj} x_j$$

Eq. (1)

where the output of the neuron, y_k, is the outcome of some activation function on the value of v_k. Examples of ANN include Multilayer Perceptron (MLP) Neural Networks, Radial Basis Function (RBF) Neural Networks, Generalized FeedForward (GFFN) Neural Networks, Probabilistic and Generalized Regression Neural Networks.

2. Methodology

2.1 Classifier System Design

Classifier Systems (Chan et al. (1999); Schetinin (2001) and Phua et al. (2004)) that combine artificial Neural Networks and Decision Trees were proposed.

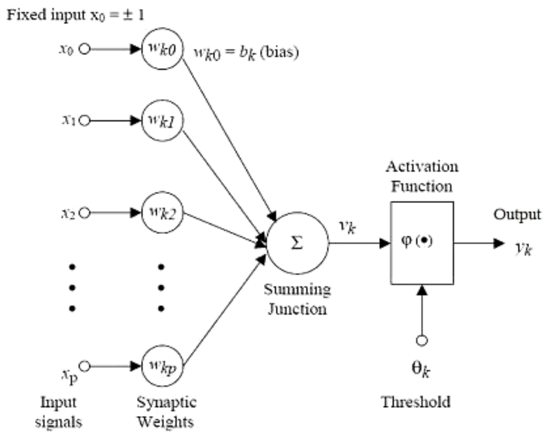

Fig. 1: Artificial Neural Network Process

Table 1: Confusion Matrix

Actual	Prediction		
	A	B	
	TN	FP	a
	FP	TP	b

In Mahesh et al. (2009), Decision Tree models recorded impressive prediction accuracy. On the other hand, the effectiveness of feed forward neural networks classifications has been tested empirically in (Guoqiang, 2000). Neural networks have been successfully applied to a variety of real-world classification tasks in industry, business and science (Widrow et al., 1994). In (Lampinen et al. (1998); Petche et al. (1998); Barlett et al. (1992) and Hoskins et al. (1990)), neural networks were applied to inspection of processes and detection and diagnosis of faults with good outcomes. Therefore in this compliance audit that has to do with inspection and detection, we experimented with a two layer classifier system for the proposed personnel audit system. Layer 1 consists of an Artificial Neural Network that models the appointment and promotion datasets, while Layer 2 is made up of a Decision Tree classifier that models suspicious appointment and abnormal promotion datasets derived from the output of the datasets of layer 1. This two-layer model reduces the size of the trees generated by the decision tree algorithm as the very effective and accurate neural networks would have been able to model the classification of the original dataset so that the decision tree classification of desired suspicious appointment and abnormal promotion dataset could yield a reliable result. Figure 2 presents the schematic diagram of the two layer Classifier System model. Different ANN and Decision Tree algorithms are tested on the dataset in order to determine that which best models the data at each layer.

The MLP and RBF Artificial Neural Networks were used for the modeling of the dataset in layer 1 while that the CART, Logitboost Alternating Decision and C4.5 decision tree algorithms were used for modeling the dataset layer 2. The comparison of the performance of the various algorithms was carried using standard metrics of accuracy, precision, recall and f-measure for classification. These are calculated using the predictive classification table called a Confusion Matrix (Table 1). Also Eq. 2-10 present the metrics used in the experiment and their definitions.

From Table 1:
TN (True Negative) = Number of correct predictions that an instance is *invalid*
FP (False Positive) = Number of incorrect predictions that an instance is *valid*
FN (False Negative) = Number of incorrect predictions that an instance is *invalid*
TP (True Positive) = Number of correct predictions that an instance is *valid*
Accuracy = The proportion of the total number of predictions that were correct.
Accuracy (%) =
(TN + TP) / (TN + FN + FP + TP) (2)
Precision = The proportion of the predicted *valid instances* that were correct:
Precision (%) = TP / (FP + TP) (3)
Recall = The proportion of the *valid instances* pages that were correctly identified
Recall(%)= TP/(FN+TP) (4)
F-Measure = This is derived from precision and recall values:
F-Measure (%) = (2 x Recall x Precision) /(Recall+ Precision) (5)
Sensitivity or true positive rate (TPR) equivalent to hit rate, recall
$TPR = TP/P = TP/(TP+FN)$ (6)
Specificity (SPC) Or True Negative Rate
$SPC = TN/N = TN/(FP+TN) = 1 - FPR$ (7)

The Kappa Statistics (κ): Is used to measure the concordance level between categorical data during prediction. Cohen's kappa measures the agreement between two raters that each classifies *N* items into *C* mutually exclusive categories.

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}, \qquad (8)$$

where Pr(*a*) is the relative observed agreement among raters, and Pr(*e*) is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly saying each category. If there is no agreement among the raters (other than what would be expected by chance), then κ ≤ 0.

The F-Measure: Is used because despite the Precision and Recall values being valid metrics in their own right, one of them can be optimized at the expense of the other. The F-Measure only produces a high result when Precision and Recall are both balanced, thus this is very significant.

The Receiver Operating Characteristic (ROC) curve: This shows the sensitivity (FN classifications) and specificity (FP classifications) of a test. The ROC curve is a comparison of two characteristics: TPR (true positive rate) and FPR (false positive rate). The TPR measures the number of *valid instances* that were correctly identified.

$$TPR=TP/(TP+FN) \quad (9)$$

The FPR measures the number of incorrect classifications of *valid instances* out of all *invalid* test instances.

$$FPR=FP/(FP+TN) \quad (10)$$

2.2 Data Modeling and Pre-processing

The data used for this research was collected from employee's data records through their credentials submitted for audit at the State's Auditor General's Office tagged as *present data* with suffix "PR". Another set of data of the employees tagged as *past data* with suffix "P" containing original and duplicate information about employees was collected from the State Teaching Service Commission (employer). The data was cleaned, normalized and organized in a form suitable for data mining. WEKA version 3.6.2, an open source data mining software developed at the Waikato University was used for the data mining processing. Table 2 presents the attributes of appointment dataset, Table 3 presents the attributes of promotion dataset and Table 4 presents the categorization of data for the data mining process.

The datasets were divided into two which includes the training and testing datasets. 66% of each of the datasets was devoted to training while the remaining 34% was used for testing of randomly selected new data. First MLP Neural Networks and RBF Neural Networks with varied parameters were used. The neurons in the hidden layers and the number of layers themselves were varied between 1 and 4 with a momentum value of 0.2, learning rate of 0.3, using the Gaussian activation function. These algorithms were used to model both appointment and promotion dataset. The Logitboost Multiclass Alternating Decision Tree (LAD), Classification and Regression Tree (CART) and C4.5 decision tree algorithms were used to model the suspicious appointment and abnormal promotion dataset. The experimental procedure used is:

1. Pre-processing: Load the formatted and normalized appointment dataset in the employees' database (file) into the classifier application.
2. Set the number of hidden layers of MLP Artificial Neural Networks to zero, with momentum of 0.2 and Learning rate of 0.3.
3. Perform training and testing of dataset using percentage split option. If the result is satisfactory, stop modeling.
4. Repeat Experiment for one to four hidden layers of MLP Artificial Neural Networks if the result of step 3 is not satisfactory.
5. Repeat Experiment for one to four iterations of RBF Artificial Neural Networks.
6. Select the algorithm that best models the dataset based on the performance measures.
7. Load suspicious employees' dataset drawn from appointment dataset into the application.
8. Perform training and testing of dataset using percentage split using LAD, CART and C4.5.
9. Select the algorithm that best model the dataset based on the performance measures.
10. Repeat step 1 to 6 for promotion dataset in employees' database.
11. Repeat step 7 to 9 for abnormal promotion dataset drawn from promotion dataset.

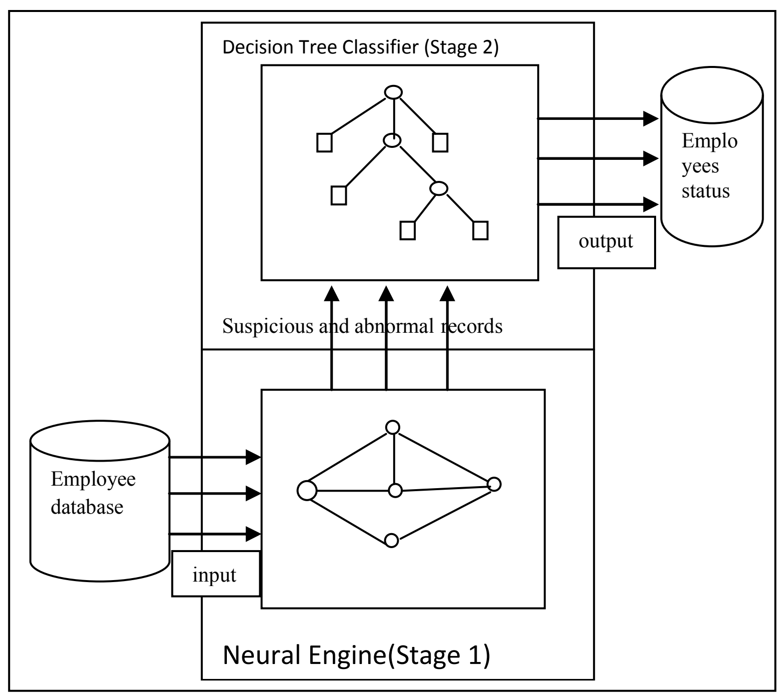

Fig. 2: Classifier System Model

Table 2: Attributes of Appointment Dataset

Attribute	Type	Description
FormNo	Numerical	Number on Verification Form
SurnameP	Categorical	Past – Surname
SurnamePR	Categorical	Present Surname
FirstNameP	Categorical	Past - First Name
FirstNamePR	Categorical	Present - First Name
PFileNoP	Categorical	Past - Personal File Number
PFileNoPR	Categorical	Present- Personal File Number
SexP	Categorical	Past – Sex
SexPR	Categorical	Present –Sex
HomeTownP	Categorical	Past- Hometown
HomeTownPR	Categorical	Present-Home Town
QualificationswithDateP	Categorical	Past - Qualification and Date of Qualification
HomeTownP	Categorical	Past- Hometown
HomeTownPR	Categorical	Present-Home Town
QualificationswithDateP	Categorical	Past - Qualification and Date of Qualification
QualificationswithDatePR	Categorical	Present- Qualification and Date of Qualification
DateofBirthP	Categorical	Past- Date of Birth
DateofBirthPR	Categorical	Present- Date of Birth
DateofFirstAppP	Categorical	Past- Date on First Appointment
DateofFirstAppP	Categorical	Present- Date of First Appointment
GLon1stAppP	Numerical	Past- Grade level when first appointed
GLon1stAppPR	Numerical	Present- Grade level when first appointed
PresentGradeLevelP	Numerical	Past- Current Grade level
PresentGradeLevelPR	Numerical	Present- Current Grade level
StepP	Numerical	Past- Current Step
StepPR	Numerical	Present- Current Step
DatePostedPresentSchlP	Categorical	Past- Date Posted to Present School
DatePostedPresentSchlPR	Categorical	Present- Date Posted to Present School
NamePresentZEAP	Categorical	Past – Name of Present Zonal Education Authority
NamePresentZEAPR	Categorical	Present – Name of Present Zonal Education Authority
PresentRankP	Categorical	Past- Current Rank
PresentRankPR	Categorical	Present – Current Rank
PresentSchlNameP	Categorical	Past – Current School Name
PresentSchlNamePR	Categorical	Present- Current School Name
Status	Categorical	Appointment Status(Class)

Table 3: Attributes of Promotion Dataset

Attribute	Type	Description
SurnameP	Categorical	Past –Surname
FirstNameP	Categorical	Past – First Name
PFileNoP	Categorical	Past- Personal File Number
SexP	Categorical	Past- Sex
QualificationwithdateP	Categorical	Past- Qualification and Date of Qualification
DateofFirstAppP	Categorical	Past – Date of First Appointment
GLOn1stApptP	Numerical	Past – Grade Level on First appointment
DateofPreviousPromotionP	Categorical	Past- Date Promoted before the last promotion
PreviousGradelevelP	Numerical	Past- Grade level in the previous promotion
StepPreviousP	Numerical	Past – Step in the previous promotion
DateLastPromotionP	Categorical	Past- Date on the current promotion letter
PresentGradelevelP	Numerical	Past- Grade level on the current promotion letter
StepPresentP	Numerical	Past – Current Step
DatePostedPresentSchP	Categorical	Past – Date Posted to Present School
NamePresentZEAP	Categorical	Past – Name of Present Zonal Education Authority
PresentRankP	Categorical	Past – Current Rank
PresentSchlNameP	Categorical	Past – Current School Name
Status	Categorical	Promotion Status(class)

Table 4: Categorization of data for the data mining process

	Number of Exemplars (Instances)			
Dataset	Algorithm	Training	Testing	Total
Appointment	*ANN*	132	70	202
Suspicious Appointment	*Decision Tree*	43	22	65
Promotion	*ANN*	139	72	211
Abnormal Promotion	*Decision Tree*	72	37	109

3 RESULTS AND DISCUSSION

The result of the Artificial Neural Networks Models of the Appointment dataset is presented in Table 5. The results show that RBF neural network with two iterations was able to model the data better than the MLP neural networks. Also, the results of experiment on suspicious dataset using decision trees presented in Table 6 shows that the Logitboost Multiclass Alternating Decision Trees (LAD) was able to model the Suspicious Appointment Dataset better than the other algorithms used while the result of the promotion dataset and abnormal promotion dataset is presented in Table 7 and Table 8 respectively. The results of Table 7 show that RBF neural network with two iterations was able to model the data better than the MLP neural networks. Also, the results of experiment on abnormal promotion dataset using decision trees presented in Table 8 shows that the Logitboost Multiclass Alternating Decision Trees (LAD) was able to model the Suspicious Appointment Dataset better than the other algorithms used. The LAD tree models satisfied the support and confidence criteria that is greater than 0.5 (>0.5). In Table 9, the results of the weighted average measures of the selected best algorithms is presented with RBF recording TP rate, F-Measure and ROC that are above 0.8 while LAD tree recorded TP rate, F-Measure and ROC that are above 0.7 for all categories of dataset as presented in Table 4.

Table 5: Results of Artificial Neural Networks Models of Appointment Dataset after Testing

Performance Measure	**Instances**	**Number of Layer (Multilayer Perceptron using Momentum= 0.2 and Learning Rate = 0.3)**					**Number of Iteration (Radial Basis Function using Gaussian Radian Function)**			
		0	1	2	3	4	1	2	3	4
Time to model (Secs)		215.33	141.24	180.88	195.41	228.38	0.19	0.09	0.09	0.09
% classification	Correct	68.5714	68.5714	68.5714	68.5714	68.5714	91.4286	91.4286	91.4286	91.4286
	Incorrect	31.4286	31.4286	31.4286	31.4286	31.4286	8.5714	8.5714	8.5714	8.5714
Kappa Statistic		0	0	0	0	0	0.8016	0.8016	0.8016	0.8016
MAE		0.3143	0.4178	0.4178	0.4178	0.4178	0.0869	0.0869	0.0869	0.0869
RMSE		0.5606	0.4656	0.4656	0.4656	0.4656	0.2915	0.2915	0.2915	0.2915
RAE(%)		73.8725	98.194	98.194	98.194	98.194	20.4829	20.4829	20.4829	20.4829
RRSE(%)		120.6984	100.2429	100.2429	100.2429	100.2429	62.7557	62.7557	62.7557	62.7557
ROC	Legitimate	0.5	0.873	0.815	0.873	0.815	0.969	0.969	0.969	0.969
	Suspect	0.5	0.873	0.815	0.873	0.815	0.961	0.961	0.961	0.961
Precision	Legitimate	0.686	0.686	0.686	0.686	0.686	0.938	0.938	0.938	0.938
	Suspect	0	0	0	0	0	0.864	0.864	0.864	0.864
Recall	Legitimate	1	1	1	1	1	0.938	0.938	0.938	0.938
	Suspect	0	0	0	0	0	0.864	0.864	0.864	0.864
F-Measure	Legitimate	0.814	0.814	0.814	0.814	0.814	0.938	0.938	0.938	0.938
	Suspect	0	0	0	0	0	0.864	0.864	0.864	0.864

Table 6: Results of Decision Trees Models of Suspicious Appointment Dataset after Testing

Performance Measure	**Instances**	**LAD**	**CART**	**C4.5**
Time to Model(seconds)		1.56	1.06	0.08
Tree Size		28	1	57
No of Leaves		12	1	55
% Classification	Correct	86.3636	68.1818	77.2727
	Incorrect	13.6364	31.8182	22.7273
Kappa Statistics		0.667	0	0.4737
MAE		0.0956	0.2995	0.1632
RMSE		0.3013	0.4049	0.1632
RAE (%)		30.9425	96.9405	52.8372
RRSE (%)		74.7561	100.455	94.8297
ROC	Ghost	0.914	0.5	0.776
	Ineligible	1	0.5	0.792
	CannotSay	0.632	0.5	0.588
Precision	Ghost	0.838	0.682	0.824
	Ineligible	1	0	0.9
	CannotSay	0	0	0
Recall	Ghost	1	1	0.933
	Ineligible	1	0	0.75
	CannotSay	0	0	0
F-Measure	Ghost	0.909	0.811	0.875
	Ineligible	1	0	0.667
	CannotSay	0	0	0

Table 7: Results of Artificial Neural Networks Models of Promotion Dataset after Testing

ISSN (Online): 1694-0814
www.IJCSI.org

Performance Measure	Instances	Multilayer Perceptron (Number of Layer) Momentum= 0.2 Learning Rate = 0.3					Radial Basis Function (Number of Iterations) Gaussian radian Function			
		0	1	2	3	4	1	2	3	4
Time to model (secs)		56.51	57.16	69.91	77.22	91.32	0.3	0.07	0.9	0.2
% class-ification	Correct	54.1667	45.8333	45.8333	44.8333	45.8333	84.7222	86.1111	86.1111	86.1111
	Incorrect	45.8333	54.1667	54.1667	54.1667	54.1667	15.2778	13.8889	13.8889	13.8889
Kappa Statistic		0	0	0	0	0	0.6872	0.7176	0.7176	0.7176
MAE		0.4583	0.5077	0.5077	0.5125	0.5163	0.1424	0.2466	0.1532	0.1587
RMSE		0.677	0.5161	0.5161	0.534	0.5518	0.3667	0.3567	0.3598	0.3573
RAE(%)		90.9677	100.7716	100.7716	101.7242	102.4698	28.2694	48.9455	30.4047	31.4935
RRSE(%)		133.8132	102.0047	102.0047	105.5435	109.065	72.4881	70.5021	71.1147	70.6211
ROC	Normal	0.737	0.922	0.934	0.929	0.93	0.944	0.945	0.945	0.945
	Abnormal	0.738	0.922	0.934	0.929	0.93	0.944	0.945	0.945	0.945
Precision	Normal	0.542	0	0	0	0	0.804	0.837	0.837	0.923
	Abnormal	0	0.458	0.458	0.458	0.458	0.923	0.897	0.897	0.788
Recall	Normal	1	0	0	0	0	0.946	0.923	0.923	0.837
	Abnormal	0	1	1	1	1	0.727	0.788	0.788	0.897
F-Measure	Normal	0.703	0	0	0	0	0.871	0.878	0.878	0.878
	Abnormal	0	0.629	0.288	0.626	0.629	0.814	0.839	0.839	0.839

Table 8: Results of Decision Trees Models of Abnormal Promotion Dataset after Testing

Performance Measure	Instances	LAD	CART	C4.5
Time to Model(seconds)		2.15	2.07	0.01
Tree Size		31	1	221
No of Leaves		15	1	217
% Classification	Correct	70.2703	40.5405	51.3514
	Incorrect	29.7297	59.4595	48.6486
Kappa Statistics		0.5547	0	0.3218
MAE		0.2321	0.4394	0.3207
RMSE		0.4213	0.4732	0.4352
RAE (%)		52.7867	99.9545	72.9362
RRSE (%)		89.1088	100.0724	92.046
ROC	Unqualified	0.855	0.5	0.821
	Referred	0.624	0.5	0.623
	Qualified	0.707	0.5	0.707
Precision	Unqualified	0.867	0.405	0.821
	Referred	0.5	0	0.623
	Qualified	0.75	0	0.707
Recall	Unqualified	0.867	1	0.733
	Referred	0.5	0	0
	Qualified	0.75	0	1
F-Measure	Unqualified	0.867	0.577	0.786
	Referred	0.609	0	0
	Qualified	0.571	0	0.5

Table 9: Weighted Average Performance Measure of Best Algorithms

RECORD	ALGORITHM	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
APPOINTMENT	RBF	0.914	0.113	0.914	0.914	0.914	0.966
SUSPICIOUS APPOINTMENT	LAD	0.864	0.292	0.75	0.864	0.802	0.891
PROMOTION	RBF	0.8661	0.15	0.864	0.861	0.86	0.945
ABNORMAL PROMOTION	LAD	0.703	0.122	0.745	0.703	0.705	0.735

Table 10: Sample of the Staffing Report for Appointment Audit

PFileNoP	FormNo	Rule	Status
TR 36876	3038	6	CannotSay
TR 29032	2356	7	CannotSay
TR23861	1880	11	CannotSay

NULL	1619	3	Ghost
NULL	1986	3	Ghost
NULL	1610	3	Ghost
TR/2007/2116	2333	2,12	Ineligible
TR.2000/151	2328	2,12	Ineligible
TR 2000/1263	3248	1	Ineligible
TR 37096	4842	Comply	Legitimate
TR.2000/2049	4912	Comply	Legitimate
TR 2000/925	4918	Comply	Legitimate

Table 11: Sample of the Staffing Report for Promotion Audit

PFileNoP	Rule	Type
NTS/PF/3516	6	Qualified
TR/29906	3	Qualified
TR/2001/01/192B	6	Qualified
NTS/PF/3428	4	Referred
NTS/PF/3453	4	Referred
TR/14924/9	13,15	Referred
NTS/PF/3079	1	Unqualified
NTS/PF/3452	1	Unqualified
31011	10	Unqualified
TR/2001/1466	Comply	Normal
TS/PF/4783	Comply	Normal
TR/37228/12	Comply	Normal

Table 12: Summary of Audit Report

Audit	Type	Status	Number of Employees
Appointment	Legitimate	Comply	137
	Suspect	Ghost	47
		Ineligible	12
		Cannotsay	6
Promotion	Normal	Comply	102
	Abnormal	Unqualified	47
		Referred	32
		Qualified	30

In the Staffing Report of Table 10 and Table 11; and Summary of Results in Table 12, employees records with status “Cannotsay” and “Referred” means such records should be subjected to further probing as the data released and used for the audit cannot provide sufficient claims for true status of the employees. From the results obtained, the level of compliance of appointment records is above average (high significance) because the percentage compliance is more than 50% while the level of compliance of promotion records is below average (low significance) as the percentage compliance is less than 50%.

3.1 General Administrative Rules

Rules generated from the LAD decision tree for suspicious appointment dataset are presented in Table 13 while rules generated from the LAD decision tree for abnormal promotion are presented in Table 14.

4 Conclusion

The two-layer Classifier System model proposed and tested on case study data from a state Teaching Service Commission database was successfully used for Personnel Audit processing. The system which is based on a soft computing data mining process combines both ANN and Decision Tree algorithms in such a way that the ANN layer is first used to determine the compliance of records while the Decision Tree layer is used to determine derivation of behaviour patterns and rules from such records. The ANN layer uses a Radial Basis Function Neural Network while the Logitboost Multiclass Alternating Decision Tree algorithm was used in the Decision Tree layer. The two algorithms recorded an accuracy that was above 70% with an average F-Measure value of over 70%. This model is being further refined for the development of a Personnel Audit Expert System which will include identity verification using fingerprint and facial recognition based mining.

Table 13: Administrative Rules Generated for Suspicious Appointments

S/No	Rule
1	IF (hometownpr = null) THEN Status = ineligible
2	IF(qualificationwithdatep=nce and hometownpr!=null) THEN status=ineligible
3	IF(presentgradelevelp<11 and qualificationwithdatep!=nce and hometownpr!=null) THEN status= ghost
4	IF(presentgradelevelp>=11 and qualificationwithdatep!=nce and hometownpr!=null) THEN status= Cannotsay
5	IF(hometownp= idanre and qualificationwithdate!= nce and hometownpr= null) THEN Status= cannotsay
6	IF(surname=ajelabi and presentgradelevelpr<15.5 and hometown!=idanre and qualificationwithdate!=nce and hometownpr!=null) THEN status = cannotsay
7	IF(firstnamep=Dickson and surname!=ajelabi and presentgradelevelpr<15.5 and qualificationwithdatep!=nce and hometownp != null) THEN status= cannotsay
8	IF(surname=ajayi and firstnamep!= Dickson and surname!= ajelabi and presentgradelevelpr<15.5 and hometownp!= idanre and qualificationwithdatep !=nce and hometownpr!=null) THEN status=cannotsay
9	IF(surname!=ajayi and firstnamep!= Dickson and surname!= ajelabi and presentgradelevelpr<15.5 and hometownp!= idanre and qualificationwithdatep !=nce and hometownpr!=null) THEN status= ghost
10	IF(presentgradelevelpr>=15.5 and hometown!= idanreand qualificationwithdatep!=nce and hometownpr!=null) THEN status=cannotsay
11	IF(hometownp=idanre)THEN status=cannotsay
12	IF(hometownp!=idanre)THEN status= ineligible

Table 14: Administrative Rules Generated for Abnormal Promotion

S/No	Rule
1	IF(presentrankp=typist and presentgradelevelp<13.5) THEN

	status=unqualified
2	IF(qualificationwithdatep=nce91 and presentrank!=typist and prsentgradelevelp<13.5)THEN status=unqualified
3	IF(qualificationwithdate!=nce91 and present!=typist and presentgradelevelp<13.5) THEN status= qualified.
4	IF (presentgradelevelp<4.5 and previousgradelevel!=null and presentgradelevelp<13.5) THEN status=referred
5	IF(presentgradelevelp>=4.5 and previousgradelevel!=null and presentgradelevelp<13.5)THEN status=qualified
6	IF(presentgradelevelp<11 and previiousgradelevel!=null and presentgradelevelp<13.5)THEN status=unqualified
7	IF (presentgradelevelp>=11 and previousgradelevel!=null and presentgradelevelp<13.5) THEN status=referred
8	IF (presentgradelevelp>=11 and previousgradelevel!=null and presentgradelevelp<13.5) THEN status=referred
9	IF (presentgradelevelp>=11 and previousgradelevel!=null and presentgradelevelp<13.5) THEN staus=referred
10	IF(dateoffirstapptp!=28/10/1985 and presentgradelevelp>13.5) THEN status=unqualified
11	IF(stepprevious>=2.5 and presentgradelevelp>=13.5)THEN status=referred
12	IF(dateofpreviouspromotion=01/01/2003)THEN status=qualified
13	IF(dateofpreviouspromotion!=01/01/2003)THEN status= referred
14	IF(datepostedpresentschoolp=19/09/2008)THEN status=qualified
15	IF(datepostedpresentschoolp!=19/09/2008)THEN status=referred

References

[1] M.J. Abdolmohammedi, S.L. Elliot, Audit: West's Encyclopaedia of American Law, Full Article, Retrieved March3, 2010 from www.answers.com/topic/audit, 2010.

[2] B. Baker, "Forensic Audit and Automated Oversight", Office of Auditor General based on logistic model tree. JBiSE. Vol.2, No.6, 2009, pp. 405-411.

[3] E. B. Barlett and R. E. Uhrig "Nuclear power plant status diagnostics using artificial neural networks," *Nucl. Technol.*, Vol. 97, 1992, pp. 272–281.

[4] P. Chan, W. Fan, A. Prodromidis and S. Stolfo "Distributed Data Mining in Credit Card Fraud Detection", IEEE Intelligent Systems, Vol. 14: 1999, pp.67-74.

[5] A.N. Chatterji, Forensic Auditing. Conference of Common Wealth Auditor General, SAI-India. Retrieved April 3, 2010 from http://www.asosai .org/ asosai_old/ journal2001 /style.css, 2001.

[6] T. Cutting, How to Survive an Audit, PM Hut. Retrieved December 13, 2009. from http://www.pmhut.com/how-to-survive-an-audit, 2008

[7] P. Z. Guoqiang, "Neural Networks for Classification", in A survey. IEEE transactions on Systems, Man, and Cybernetics Part C: Applications and Reviews, 2000, Vol. 30, No. 4.

[8] J.C. Hoskins, K. M. Kaliyur and D.M. Himmelblau D. M., "Incipient fault detection and diagnosis using artificial neural networks," in *Proc. Int. Joint Conf. Neural Networks*, 1990, pp. 81–86.

[8] J. Lampinen, S. Smolander and M. Korhonen, "Wood Surface Inspection System based on Generic Visual Features," in *Industrial Applications of Neural Networks*,1998, pp.35–42.

[9] T. Petsche , A. Marcantonio, C.Darken, S.J. Hanson, G.M. Huhn, and I. Santoso, "An autoassociator for on-line motor monitoring," in *Industrial Applications of Neural Networks*, 1998, pp. 91–97.

[10] V. Mahesh, A. Kandaswamy, C. Vimal and B. Sathish "ECG arrhythmia classification based on logistic model tree," JBiSE, 2009, Vol.2, No.6, pp. 405-411.

[11] C. Phua, D. Alahakoon and V. Lee, "Monitoring Report in Fraud Detection: Classification of Skewed Data," SIGKDD Exporation , 2004,Vol. 6, No 1, pp. 50-56.

[12] C. Phua , V. Lee, K. Smith and R. Gayler, A Comprehensive Survey of Data Mining-based Fraud Detection Research, Retrieved March 17, 2010fromhttp://www.clifton.phua.googlepages.co m /**fraud-detection-survey**.pdf, 2004.

[13] V.Schetinin, "A Neural Network DecisionTreefor Learning Concepts from EEG Data," in NIMIA - SC2001 - 2001 NATO Advanced Study Institute on Neural Networks for Instrumentation, Measurement, and Related Industrial Applications: Study Cases Crema, Italy, 2001, pp. 9-20.

[14] B. Widrow, D. E. Rumelhard, and M.A. Lehr, "Neural networks: Applications in industry, business and science," in *Commun. ACM*, 1994, Vol. 37, pp. 93–105.

[15] O. R. Zaiane, Principle of Knowledge Discovery in Databases, University of Alberta. Department of Computer Science. CMPUT690. 1999.

First Author Adesesan B. Adeyemo obtained M.Sc and Ph. D. in Computer Science from Federal University of Technology, Akure, Nigeria in 1998 and 2008 respectively. He is a lecturer in Department of Computer Science of University of Ibadan, Nigeria. His research interest includes Data Mining and Computer Netwoks. He is a member of CPN and NCS in Nigeria

Second Author Oluwafemi Oriola obtained M.Sc. in Computer Science from University of Ibadan, Ibadan, Nigeria in 2010. He works with SITA, Akure, Nigeria. His research interest includes Data Mining and Image Processing.