AUTOMATIC DATA COLLECTION DESIGN FOR NEURAL NETWORKS DETECTION OF OCCUPATIONAL FRAUDS

Bakpo, F.S.,
Department of Computer Science,
University of Nigeria, Nsukka.
E-mail: fbakpo@yahoo.com

ABSTRACT

Automated data collection is necessary to alleviate problems inherent in data collection for investigation of management frauds. Once we have gathered a realistic data, several methods then exist for proper analysis and detection of anomalous transactions. However, in Nigeria, collecting fraudulent data is relatively difficult and the human labour involved is expensive and risky. This paper examines some formal procedures for data collection and proposes designing an automatic data collection system for detection of occupational frauds using artificial neural networks.

Keywords*: Occupational frauds, automatic data collection, artificial neural networks.*

Introduction

Nigeria has consistently been widely rated as one of the most corrupt nations in the World[1,2] and this has been identified as a key constraint to the development of the national economy. High- level management fraud is continuously increasing with each national government coming to power, resulting in the loss of billions of naira every year. An indication of the extent to which Nigeria is perceived as being corrupt is given in the Corruption Perception Index published by Transparency International on 22 September 1998.

This Index for Nigeria is 1.76 on a scale of 1 to 10 in which 1 represents the highest level of perceived corruption[1,2]. While the vast oil industry should in itself boost the Nigerian economy up, the corrupt government controls the fields, and the government has been more interested in personal profit than in improving the countries economy. Over the past twenty years, Nigeria has generated approximately $300 billion from oil revenues. Fraud, waste, and mismanagement have sucked away $200 billion, or two-thirds of the oil revenue. The estimated figure of the amount of money stolen out of the remaining revenue is reckoned to be about $50 billion. Corruption does not only permeate into the government and oil fields of Nigeria. It attacks the entire nation. For example, Ajaokuta, a steel mill in Nigeria, has been under construction for the past seventeen years and throughout that period of time has consumed more than seven billion dollars. It has produced no steel. Another example is Alscon, an aluminium plant in Nigeria, has consumed three billion dollars over the past five years. The project was to produce 190,000 tons of aluminium, but, like its predecessor, Ajaokuta, has not produced any aluminium to date. Management fraud may be defined as "deliberate fraud committed by management that injuries investors and creditors through materially misleading financial statements or intentionally or reckless conduct, whether by act or omission, that leads to materially misleading financial statements[3"]. Major forms of fraudulent behaviour include occupational fraud, money laundering, advance fee frauds etc. This study is concerned with occupational fraud, which is the most prevalent form of fraudulent financial practices (FFPs) in Nigeria.

2. Occupational Frauds Defined

Also known as fund embezzlements, occupational fraud may be defined as "the use of one's occupation for personal enrichment through the deliberate misuse or misapplication of the employing organisation's resources or assets[3,7] ". This definition also covers a wide range of misconduct by employees, managers, and executives. Regardless, all occupational fraud schemes have four key elements in common. The activity is:

(i) Clandestine (i.e. kept secret);

(ii) Violates the perpetrator's fiduciary duties to the victim organisation;

(iii) Committed for the purpose of direct or indirect financial benefit to the Perpetrator;

(iv) Costs the employing organisation assets, revenue, or reserves.

2.1 Types of Occupational Frauds

Occupational fraud can be classified into the following three types:

(i) Asset misappropriations involving the theft or misuse of an organisations asset.

(ii) Corruption- When fraudsters wrongfully use their influence in a business transaction in order to procure some benefit for themselves or another person, contrary to their duty to their employer or the rights of others.

(iii) Fraudulent statements involving the falsification of an organisation's financial statements. Figure 1 shows this classification scheme.

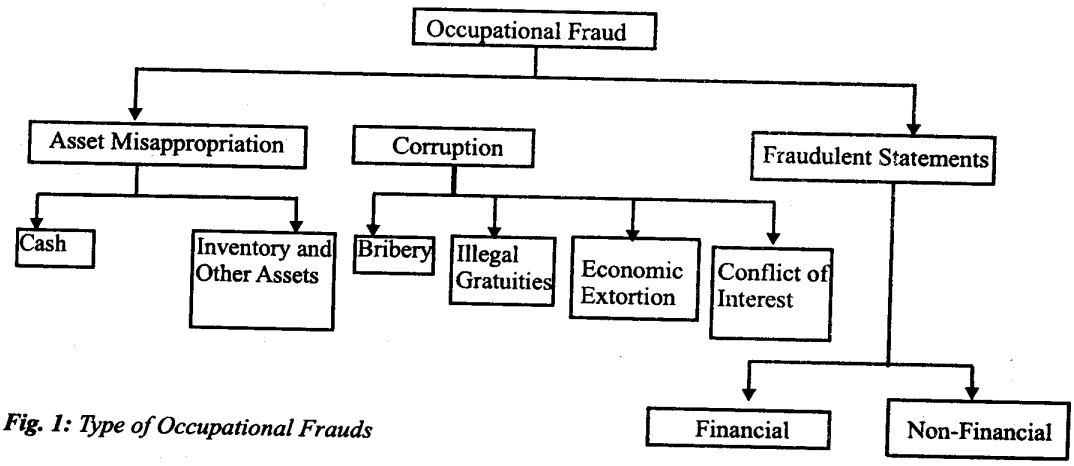

Fig. 1: Type of Occupational Frauds

2.2 Fraud Prevention and Detection

We wish to distinguish between these two concepts: Fraud Prevention and Fraud Detection. Fraud prevention describes measures taken to stop fraud from occurring in the first place. Such measures usually include elaborate design, fluorescent fibres, multi-tone drawings, watermarks, holographs on banknotes, PINs for bankcards, Internet security systems for credit card transactions, SIM cards for mobile phones, and passwords on computer systems and telephone bank accounts. Of course, none of these methods are perfect, and in general, a compromise has to be struck between expense and inconvenience (for examples, to a customer), on the one hand and effectiveness on the other. In contrast, fraud detection involves identifying fraud as quickly as possible once it has been perpetrated. Fraud detection comes into play once fraud prevention has failed. In practice, of course, fraud detection must be used continuously, as one will typically be unaware that fraud prevention has failed. Fraud detection is a continuously evolving discipline[8,9]. Whenever it becomes known that one detection method is in place, criminals usually adapt their strategies and try others. New criminals are also

constantly entering the field. The development of fraud detection methods is also made more difficult by the fact that the exchange of ideas in fraud detection is severely limited. It does not make sense to describe fraud detection techniques in great detail in the public domain, as this gives criminals the information that they require in order to evade detection. Data sets are not made available and results are often censored, making them difficult to assess[8].

3. A Survey of Formal Data Collection Methods.

This section attempts to examine the formal data collection procedures[4]. The strengths and limitations of each method are summarized as follows:

1. Questionnaires or Structured Interviews

Strengths

(i) Consistent for everyone (ii) Relatively easy to code (iii) Good when you need data for large numbers of people or need quantitative data (iv) Provides information on "scope" (i.e., how many people do a behaviour or are affected by an issue).

Limitations

(i) Little flexibility for people to raise their own issues (ii) Little opportunity for people to respond in their own words (iii) Little opportunity to go into depth on any issue (iv) Usually limited to quantitative data

2. Unstructured or Semi-Structured Interview

Strengths

(i) Permits people to raise issues important to them and speak in their own words

(ii) Rapport encourages sharing

(iii) Provides information on "deept" (why do people feel or act as they do and what does it mean to them) (iv) Open-ended responses can be analysed qualitatively or quantitatively

Limitation

(i) Time-consuming (ii) Fewer people can be interviewed than surveyed

(iii) More difficult to code (iv) Interviewers need in-deept training to interview well

3. Focus Groups

Strengths

(i) Interaction encourages people to elaborate on their responses (ii) Ideas can be shared (iii) More people can be interviewed in groups than one-to-one (iv) Program staff can get a lot of good suggestions in a short period of time.

Limitations

(i) Some people do not speak well in groups and others may dominate discussions (ii) Will only work if moderator has good facilitation skills (iii) Not good if you need data from individuals or if topic is sensitive (iv) Scheduling can be problematic for busy people or those with family responsibilities

4. Journals

Strengths

(i) Provide an ongoing record of people's felt experiences with the program (ii) Focus participants and others on their experiences as they happen (iii) Can provide extensive detailed data to be used in analyses or as ideas for interviews or focus groups

Limitations

(i) Time-consuming to keep (ii) Some people do not communicate well in written English (iii) Some people may feel uncomfortable writing about themselves or feel they are giving up their privacy.

5. Observation

Strengths

(i) Provides data other than verbal self-report (ii) Provides an opportunity for in-depth study of behaviour, nonverbal communication and the physical and social environment (iii) Allows immersion in the whole context of the program.

Limitations

(i) Time-consuming and often expensive (ii) May seem intrusive to program staff and participants (iii) Observers need extensive training (iv) Conflict over role may arise (e.g. to what extent can an observer be a participant?).

6. Population Statistics

strengths

(i) Provide information about change on a broad scale (ii) Useful if program is geared to a large population

(iii) Usually are easily available and regularly collected (though government spending cuts have led to less availability of some status or higher charges for accessing the data)

Limitations

(i) Limited to quantitative data (ii) Are influenced by many non-program factors (iii) Short-term changes may be chance fluctuations rather than trends (iv) Only useful if large samples from large populations.

4. Automatic Data Collection System

4.1 Investigative Domain

In Nigeria, Occupational fraud is usually committed at any of the following levels: Federal government, state government parastatals, local government areas and private firms. Figure 2 depicts an organisational structure of a university system.

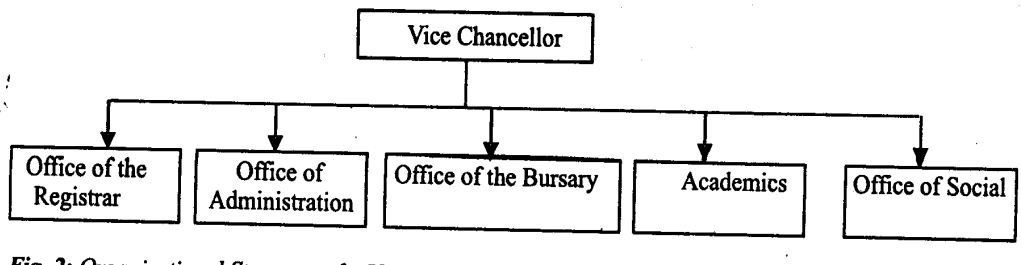

Fig. 2: Organisational Structure of a University System

In the above organisation, apart from the annual and supplementary budgetary allocations (call these external income), other revenues may be generated internally (call these internal income). The executive manager (vice-chancellor) is solely responsible for disbursement of funds to other subordinate units and expenses on the needs of the entire organization e.g. salary payments, project implementation etc. For this reason, if fraudulent financial practices is suspected and need to be investigated, relevant data must be collected from the executive managers domain before or at the end of his tenure in office. Collecting data in a domain suspected as fraudulent using any of the formal methods discussed in section 3, is relatively problematic and the human labour involved is expensive and at risk. These limitations suggest the need for the design of an automatic data collection system.

4.2 Proposed Architecture for Automatic Data Collection and Detection Design.

The major functionalities of the proposed automatic data collection and fraud detection design are as follows: (i) to facilitate real-time data collection and (ii) to react to a suspicious cooperation or transaction that may lead to a fraud. Consequently, the design of the architecture is based on the following conditions:

(i) The system provides a non-stop transaction environment within an identified bank with which an

executive manager transacts.

(ii) Each executive manager's transaction such as withdrawal, saving, agreement records etc is treated as a signature and saved in a large database.

(iii) Deviation from the usual pattern of an entity may imply the existence of a fraud, e.g. a sharp increase in an amount withdrawn as salary, urgent withdrawal of huge sums including at late hours etc may result from accounting fraud.

(iv) The similarity between an entity's current activity and a known fraud scenario indicates the same fraud may occur again.

(v) Analysis of an entity's behaviours in a relatively long period may reveal his real intentions that are covered by good activities (decisive intention). Figure 3 shows this architecture.

The proposal architecture consists of two subsystems: automatic data collection and fraud detection design using Neutral Networks.

4.2.1 The Automatic data collection subsystem consists of the executive manager's transaction workspace and the transaction table. As he carries out each transaction with his bank, transaction profile is added or saved onto a table within the database. A transaction will automatically enter the system from the real-time data feed via a form or a CORBA (Common Object Request Broker Architecture) interface linked to the organization's existing system. Transaction data is further used to build a transaction. table (similar to log files[10]), which provide specific information regarding manager's transactions.

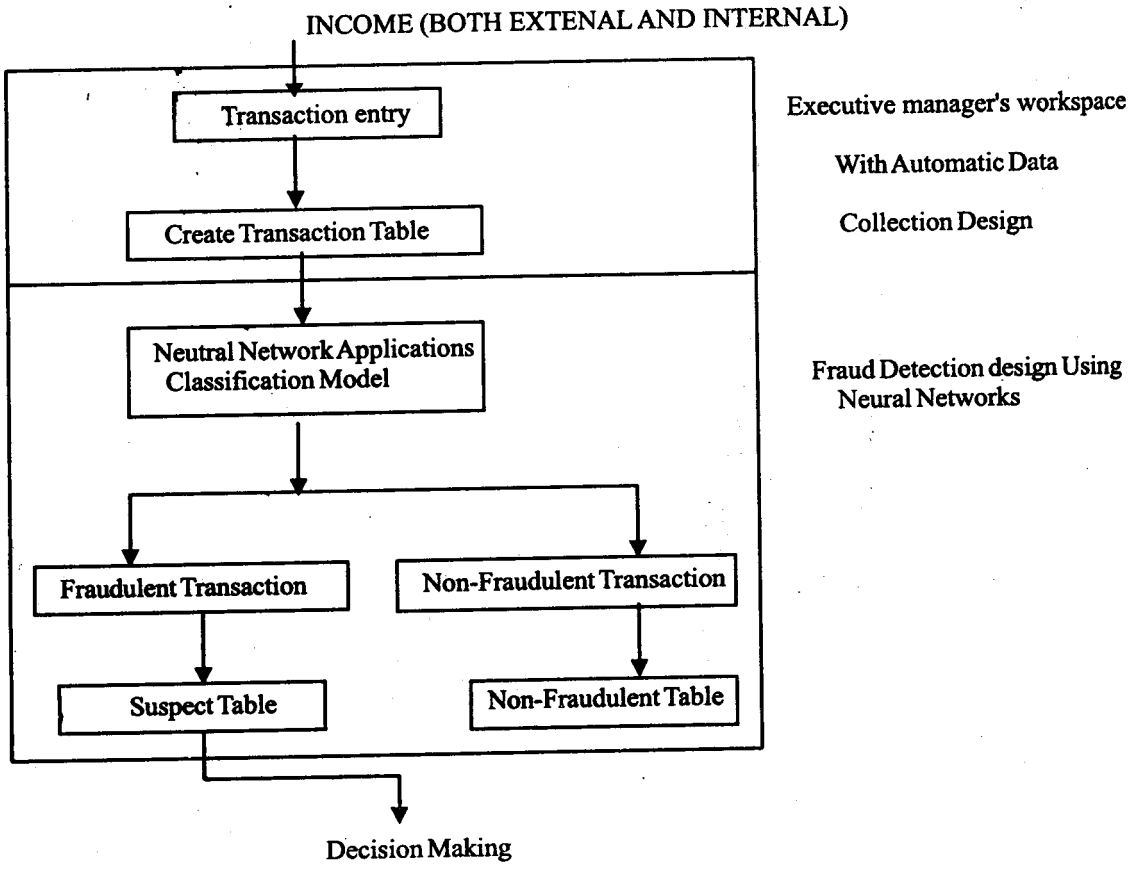

Fig. 3: Proposal Architecture for Automatic Data Collection and fraud detection Design.

Each transaction would normally create one line in the table with parameters given according to the transaction signature. The format of a common transaction table includes: domain Name, host name, manager's authentication, date/time, type of transaction, purpose of transaction, quantity of goods, Amount involved, completion code.

The **domain name** indicates the name of the organisation making a transaction. The **host name** refers to the server/or the bank with which the executive manager is transacting. This will not change if only one bank is patronized. The **manager's authentication** is a useful way to follow specific managers in an attempt to assess if the intended individual is transacting. **Date** and **time** is a very useful part of this signature. It indicates the exact time/date of occurrences of transactions. **Type/purpose** of transaction refers to the intended type and purpose of transaction e.g. Withdrawal for payment of April 2004 salary of workers etc. **Quantity of goods** this entry refers to purchased goods or equipment. It is useful if the manager intends to purchase necessary equipment, otherwise left blank. **Amount involved** shows the amount to be withdrawn, saved or involved in the specified transaction. Completion code-is a submit button, which indicates completion of a transaction. The structure of the transaction table is shown in table 2.

4.2.2 Fraud Detection Subsystem using Neutral Network

Before now, no adequate measures were taken to detect management fraud and consequently punish perpetrators. Uncovering management fraud is a difficult task, using normal audit procedures, which perhaps contributed to the continuous perpetration of this crime in our society. First, there is a shortage of knowledge concerning the characteristics of financial fraud. Second, given its infrequency, most auditors lack the experience necessary to detect it. Finally, bank managers may also be involved and may be deliberately trying to deceive the auditors. Given managers who understand the limitations of an audit and the large volume of data involved, standard auditing procedures may be insufficient. These limitations suggest the need for intelligent procedures for its detection. The present study will apply the techniques of Artificial Neutral Networks (ANNs) to develop a decision making aid to be used in detecting instances of management frauds.

ANN is information processing paradigm that is inspired by the way biological nervous system such as the brain processes information[5,6,9] it is compose of a large number of highly interconnected processing elements (neurons) working in unison to solve specific problems. A neural network learns relationship between sets of input and output patterns, and can then intelligently responds to new inputs using the experience gained during training. Research in this area started in 1943 by the neuro-physiologist Warren McCulloch and the logician Walter Pits[6].

Table 2: Format of the transaction table

Domain	Host	Authentication	Date	Time	Transaction	Purpose	Qty	Amount
ESUT	Union Bank Plc	Prof. OBI C.C	1/01/04	16:18:00	Withdrawal	Salary	-	8million
ESUT	„	„	2/01/04	14:18:00	Withdrawal	Leave Bonus	-	4million
ESUT	„	„	5/01/04	15:00:00	Withdrawal	Entertainment	-	2 million
ESUT	„	„	7/01/04	10:30:00	Withdrawal	Sal. Arrears	„	„

For a thorough study of ANNs, the reader is referred to [5,6]. ANN can be configured for a specific application through a learning process. Generally, we can categorize the learning situations in ANNs into: **supervised** and **unsupervised**.

In supervised methods, also known as back propagation training, samples of both fraudulent and non-fraudulent records are used to construct models, which allow one to assign new observations into one of the two classes. This requires one to be confident about the true classes of the original data used to build the models. It also requires that one have examples of both classes. Furthermore, it can only be used to detect frauds of a known type, which have previously occurred. In contrast unsupervised methods, also known severally as Kohonen's self-organizing maps, the competitive learning, the winner- take-it-all learning, are used when there are no prior sets of legitimate and fraudulent observations. Unsupervised classification learning is based on clustering of input data. Clustering is understood to be the grouping of similar objects and separation of dissimilar ones. Figure 4 shows a simple clustering network.

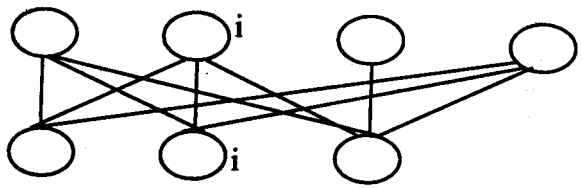

Fig. 4: A simple clustering ANN

The network is fully connected with weight W_{ij} connecting neuron i in the output layer to neuron j in the input layer. Both the input and weights vectors are normalized to unit length initially. The activation of neuron i in the output layer is given by

$$a_i = \sum_j w_{ij} \times_j = w_i^T \times$$

The algorithm typically uses a set of representatives wj, which are moved around in space R^D to match (represent) regions that include relatively large amount of samples. Every time a new sample is introduced the representatives compete with each other and the winner is updated (moved). Other representatives can be updated at a slower rate, left alone or be punished (moved away from the sample). There are a number of algorithms known to belong to the unsupervised clustering algorithms. These include:

i) Modified Basic Sequential Algorithmic Scheme (MBSAS),
ii) Two-threshold Sequential Algorithmic Scheme (TTSAS),
iii) General Agglomerative scheme (GAS),
iv) General Divisive Scheme (GDS),
v) Generalized Competitive Learning Scheme (GCLS).

A generalize competitive learning scheme may be written as follows:

1. t=o; //Time =o
2. m=minit; //Number of clusters
3. While NOT convergence AND (t<tmax)
   a. t = t+l
   b. Present a new sample x and calculate winner wj
   a. if(x NOT similar with wj)AND (m<mmax)
      i. m =m+l; //New cluster
      ii. Wm =x
   d. else // update parameters
      i. if winner wj (t) =Wj (t-l) + nh (x, wj (t-l)
      ii. if not winner Wj (t) =Wj (t-l) + nh (x, wj (t-l))

4. Clusters are ready and represented by wjs. Assign each sample to the cluster whose representative is the closest. Parameters n and n' are called learning rate parameters. They control the rate of change of the winners and losers. The function h is some function usually dependent on distance. Calculating the total change in the vectors wj and comparing it to a selected threshold value can consider

convergence, for example.

4.3 Implementation

The brain behind this new system design is a user interface or form, which is used to collect real- time transaction data. As each transaction enters the system it is added to the database as shown in figure 3 and table 2. A transaction will automatically enter the system from the real-time data via a CORBA (common object request broker architecture) interface linked to the organization's existing systems or via a client operator interface as shown in figure 5.

Transaction data obtained from the database is used to automatically generate a neutral network-clustering model. When a transaction enters the system its history is calculated. This is the historical data associated with the account, such as balances (over a 30 days), frequencies (use over 30 days) and cash velocity (rate of cash flow over a period of time). The model generated is then used to classify all incoming transactions collected in real- time, to determine the cluster centre it should belong to and the distance from this centre. When the model classifies a transaction as normal (that is, where the transaction data is within a certain distance of the cluster centre) the transaction is passed without action and sent to the non- fraudulent transaction table. Where a transaction falls outside of a specified distance from the centroid, that transaction is flagged and sent to the fraudulent transaction table for further investigation and decision-making.

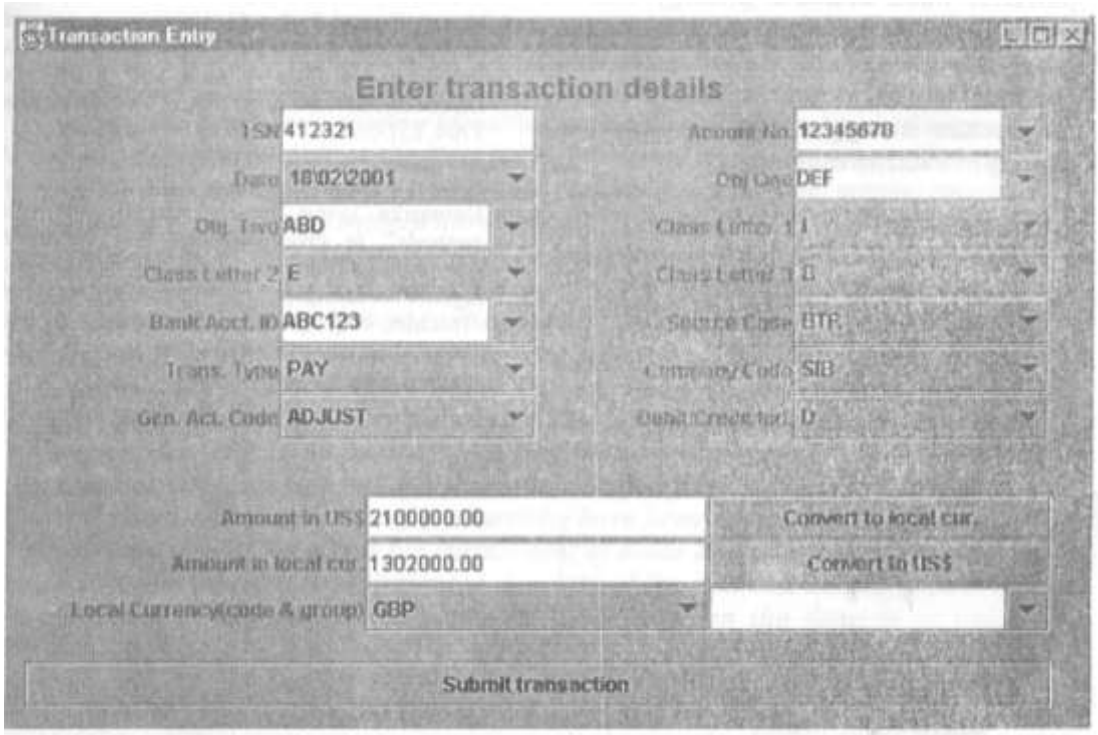

Fig. 5: Client operator interface for Automatic data collection

4.4 System Requirements

The system requirement is as in a computerized or electronic banking system with which the executive manager of any organization transacts. The framework (software) can be embedded into such custom applications where this whole process would be run secretly in the background. The complete software has been developed as part of an on-going PhD thesis, which I am currently carrying out in the Computer Engineering Department, Enugu State University of Science and Technology.

5. Conclusion

The contribution made by this work is threefold. First, the formal data collection schemes are discussed to expose their inherent shortcoming in environments where fraudulent financial practices are being suspected. Secondly, a concept of artificial neural networks is presented as a powerful fraud detection strategy that is well suited for detection of occupational frauds in Nigeria. Thirdly, an automatic data collection and neural network-based architecture is suggested to aid decision- making in fraud investigations.

REFERENCES

1. Daily Times. October 8, 2003 "Nigeria ranked second most corrupt nation"*http:/www.dailytimesofnigeria.com /DailTimes/2003/October/8/Nigeria ranked.asp/*

2. Daily Trust. October 8, 2003 "Nigeria still Second most corrupt country-TI"*http:/www.mtruslonline.com/dailytrust/ cbn08102003.htm*

3. Fanning, Cogger, K and Srivastava, R (1995) "Detection of Management Fraud: A neural Network approach". *International Journal of Intelligent Systems in Accounting, Finance & Management, Vol.* 4, *No* 2, *pp* 113 126.

4. Horne, T (1995) "Summary of Data Gathering Methods" http://www.web.net /tamhorne /data.htm

5. Loi, L.L. (1998) *Intelligent System Application in Power Engineering: Evolutionary Programming and Neural Networks*, John Wiley & Sons, UK, 264p.

6. Haykin, S (1999) *Neural Networks: A comprehensive Foundation.* Second Edition. Prentice Hall, 842p.

7. National Commission of Fraudulent Financial Reporting (NCFFR) 1987. *Reports of the NCFFR: The Role of the SEC. Oct. NCFFR*, New York.

8. Murad, U., and Pinkas, G (1999) "Unsupervised profiling for identifying supervised frauds". In Principles of Data Mining and Knowledge Discovery, Lecture notes in Artificial Intelligence 1704, 251-261.

9. Taniguchi, M., Haft, M., Hollmen, J. and Tresp, V. (1998). "Fraud Detection in Communication Networks using Neural and Probabilistic methods". *In Proceedings of the* 1998 *IEEE International Conference in Acoustic. Speech and Signal Processing (ICASSP98). Vol 2.1241-1244.*

10. Trochim, W.M.K. Cirillon, D. (2003) "Automatic Data Collection with log files" on Internet: http:// trochim.human.cornel.edu.