A Causality Learning of E-banking Operational Risk using Tree Augmented Naïve Bayes Classifier

Ako Rita Erhovwo, Okpako Ejaita Abugor

Abstract: ***E-banking systems have been shown to increase and modify particularly Operational Risk (OR). It has increased the technical complexity of the banks operational and security issues. The mode of occurrence, magnitude, and consequences often takes on new dimensions. It has become increasingly important to effectively identify potential OR issues underlying the E-banking operations, their causal relationships, the effectiveness of controls implemented, the inherent risk exposure level, and the residual risk. This research work seeks to propose Tree Augmented Naïve Bayes (TAN) Classifier in the modeling of the causal relationships among operational risks factors. To validate the proposed use of TAN classifier, we comparatively analyzed the performance of the TAN classifier with three other soft computing tools; C4.5 Decision Tree, Naïve Bayes (NB) and Artificial Neural Networks (ANN). These soft computing tools were evaluated in terms of the CPU training time complexity, classification measured by prediction accuracy, ranking measured by AUROC, and the Mean and Relative absolute error rate. The dataset was pre-processed and transformed by conducting a factor analysis procedure using SPSS statistical measurement tool, to identify risks that may require urgent actions and to reduce the dimensionality of the dataset into a smaller subset of most significant measurable variables. WEKA was then used as the developmental tool for training and testing the soft computing classifiers. Through causality learning from the collected E-banking Customers' data, we demonstrated that the proposed classifier cannot only discover causalities but also perform better in prediction than other algorithms, such as C4.5, NB, and ANN. The TAN network structure revealed the interdependencies among operational risk factors.***

Index Terms: Causal Relationships, Operational Risk, Soft Computing, Classifiers, E-banking.

I. INTRODUCTION

Recent financial reports have revealed substantial financial losses in the E-banking system as a result of Information Systems (IS) malfunctions such as; SQL injections [1], virus attacks, phishing attacks, and other factors [2]-[4]. These financial losses have been categorized as OR and as a result the area of E-banking OR assessment is becoming extremely attractive for researchers to explore ways to developing risk assessment methodologies and tools, for measuring inherent risk exposures and in computing the economic capital requirements.

The Basel Committee on Banking Supervision in its continued focus on monitoring the implementation of its standards and guidance, and in light of the significant number of recent operational risk-related losses incurred by banks, reviewed its "principles for a sound management of operational risk" guidance issued in June 2011 [5]. The committee based on its review recommended that banks should improve the implementation of each of the operational risk identification and assessment tools, including risk and control self-assessments, key risk indicators, external loss data, business process mapping, comparative analysis, and the monitoring of action plans generated from various operational risk management tools [5].

Despite the critical need for effective risk assessment approaches, there is still no consensus regarding the evaluation tools and techniques for OR assessment. Over the past decades, several tools and techniques have been developed for operational risk assessment and include Key Risk Indicators (KRIs), Monte Carlo simulations, soft computing tools (e.g. Decision trees, Bayesian Network (BN), Fuzzy Inference Systems (FIS), (ANN), and so on. However, a number of factors influence the selection of risk assessment tools and techniques [6]. These include but not exhaustively availability of resources and information, the nature and degree of uncertainty in the data, and complexity of the application [6]. Moreover, the Basel II accord allows some degree of flexibility in computing the economic capital requirements, especially within the Advanced Measurement Approach (AMA) [7]. As a result the development of tools and techniques for conducting operational risk assessment is very dynamic.

The research question explored in this paper is: how to identify the main variable dependencies on E-banking OR factors, that will help risk officers/management to review and make predictions on their E-banking risk profile, and also help to reveal the effects of possible interventions on their planned system? The main objective of this paper consist of using TAN classifier for modeling the causal relationships between risk factors, key risk indicators and other attributes in the context of E-banking systems.

TAN classifier can be constructed into a "multi-level" model [8], which can show several levels of dependency among several risk factors (e.g. frequency of outsider fraud attacks as a result of successful Trojan attacks on a customer computer, which is also enhanced by the weaknesses of the bank IT systems, such as the cryptographic techniques). TAN classifier inputs are similar to Monte Carlo model inputs,

Revised Version Manuscript Received on October 20, 2018.

Dr. Rita Erhovwo Ako, Mathematical Sciences, Edwin Clark University, Kiagbodo, Nigeria, E-mail: ochukorita2@gamil.com

Dr. Ejaita Abugor Okpako, Mathematical Sciences, Edwin Clark University, Kiagbodo, Nigeria, E-mail: okpako.ejaita@gmail.com

in that they can define system variables, causal relationships between variables, add evidence to network and outcome variables are described by deterministic equations [6]. The classifier has the additional benefits of not requiring recording prior to investigation of different scenarios, representing graphically the relationships between parameters and making the assumptions underlying the model explicit [9].

In the application example, the studied process relies on the resulting output of the preliminary data analysis, which was used as input to classifying the most significant risk issues and causal relationships. Reliability of scale was carried out on the pre-processed dataset, in order to test consistency of the scale and ensure that the scale is measuring the same construct. Factor analysis was undertaken to reduce the dimensionality of the dataset. These processes were considered relevant to optimize performance and prediction accuracy of the proposed soft computing algorithm, as every prediction algorithm has its weaknesses and strength [10], [11] such as capability to handle noise, missing data, various attribute coding and so on.

II. LITERATURE REVIEW AND RELATED WORK

Recent advances in the field of soft computing are materializing into a wider usage and are considered outstanding tools for modeling cause-effect relationships at multiple levels (such as ***X*** depends on ***Y*** which depends on ***Z*** and so on) [8]. They are capable of reasoning under conditions of uncertainty and vagueness [8], [12]-[18]. They are also capable of predicting future occurrences and possible intervention, which makes them attractive tools for OR assessment and management.

Significant amount of research have been conducted on the use of soft computing such as Decision Trees, BN, FL, ANNs and Hybrid Intelligent Systems. The applications are quite numerous and diverse. To mention a few in OR assessment and management, [19]-[24] applied BN for modeling OR losses as specified in the Basel II accord. References [8], [25]-[28] demonstrated the usefulness of BN in assessing and managing internally OR at the business unit level. Reference [29] on the other hand proposed the use of Credal Networks, which are a generalization of BNs to imprecise probabilities for measuring and managing operational risk. Commercial software packages are well available for quantification and capital allocation (see agena.co.uk and algorithmics.com).

Another approach is to use FL for managing OR and E-banking security assessment; [30]-[33] at the business, micro, and functional unit levels. References [34]-[39] applied FL to many other field which deals with data mining, risk assessment and analysis.

Reference [40] applied BP neural network predictor model for the prediction of OR losses in commercial banks using historical data. Reference [41] attempted the use of ANN for improving the data analysis process of OR assessment in the financial institutions for Loss Distribution Approach (LDA) computation. References [42], [43] proposed ANN to establish a new risk assessment model of high-tech project investment and landslide hazard and risk analysis respectively. ANN was also proposed by [44] for periodontitis risk assessment.

The prospect of Decision Tree classification was not left out. Reference [45] used CHAID (Chi-Square Automatic Interaction Detector) decision tree algorithm to identify financial profiles of firms and determine operational risk factors. Reference [46] analyzed medium-term risks faced by electrical generation companies in competitive environments using a decision tree approach.

Attempt on the combination of soft computing tools to internally assess OR exposure level inherent in such system or process have also been proposed. Reference [47] proposed fuzzy neural network for fraud detection. Reference [48] proposed the use of soft computing methods in assessing the risk to electrostatic fire and explosion hazards in industries. Reference [49] proposed soft computing approach to conduct risk assessment on collision risk.

Significantly missing in the available literature is the proposal for the application of TAN classifier for OR assessment in the context of E-banking, such that OR issues and causal relationships based on Markov conditional independence assumption are determined in order to further assess the effectiveness of implemented controls, the risk exposure level and inherent risk. As a result open new ground for more research, which was explored in this paper.

A. Naïve Bayes Classifier

Naïve Bayes classifier is the widely known and accepted BN classifier. It assumes that all attributes are conditionally independent given the class. Classification in Naïve Bayes is done by applying Bayes rule to compute the probability of class variable *C* given the instances of $X_1, X_2, ..., X_n$ and then predicting the class with the highest posterior probability [50]. Each attribute node has the class node as its parent, but does not have any parent from attributes nodes. Naïve Bayes classifier learns the conditional probability of each attributes X_i given the class label *C* from training data.

The construction of Naïve Bayes is easy since the values of *P*(*c*) and $P(x_i \mid c)$ can easily be estimated from training data. The computation is feasible because a strong probabilistic independence assumption is held in the sense that all attributes X_i are conditionally independent given the value of the class *C*, that is *X* is independent of *Y* given *C* wherever $P(X \mid Y, C) = P(X \mid C)$ for all possible values of *X*, *Y* and *C*, whenever $P(C) > 0$ [51].

The Bayesian approach to classifying the new instance is to assign the most probable class value $c(V)$, given the attributes values $\langle x_1, x_2, ..., x_n \rangle$ that describe the instance.

$$c(V) = \arg \max_{c \in C} P(c \mid x_1, x_2, ..., x_n) \quad (1)$$

Using Bayes theorem,

$$c(V) = \arg \max_{c \in C} \left(\frac{P(x_1, x_2, ..., x_n \mid c)\, p(c)}{P(x_1, x_2, ..., x_n)} \right) \quad (2)$$

$$= \arg \max_{c \in C} P(x_1, x_2, ..., x_n \mid c)\, p(c) \quad (3)$$

The Naïve Bayes classifier simplifies further by making the assumption that the attribute values are conditionally independent given the class. The resulting Naïve Bayes is defined in equation 4 and structurally represented in fig. 1:

$$c(V) = \arg \max_{c \in C} P(c) \prod_{i=1}^{n} P(x_i \mid c) \quad (4)$$

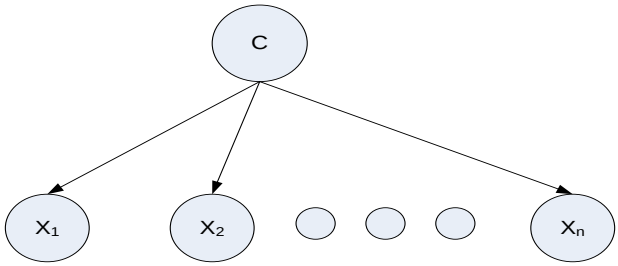


Fig. 1 Structure of a Simple Naïve Bayes Network

Although, the predictive performance of Naïve Bayes is competitive with the state-of-the-art classifiers, it does not represent any variable dependencies given the class variable. The conditional independence assumption is clearly unrealistic, consider for example, a classifier for assessing the risk in E-banking system adoption: it seems counterintuitive to ignore the correlations between age, literacy level, culture and income or for example considering E-banking authorization and authentication process: the correlations between malware attacks, weaknesses in cryptographic techniques of the system and fraudulent attacks. These examples would result in high bias and low variance and therefore harm the performance accuracy of the Naïve Bayes strategy.

B. Tree Augmented Naïve Bayes Classifier

In order to overcome the conditional independence assumption and avoid the intractable complexity of learning BN structure, various improvements on the so called Naïve Bayes have been proposed. Reference [51] proposed a Tree Augmented Naïve Bayes Classifier to overcome the challenges of Naïve Bayes, due to the fact that Naïve Bayes does not represent any variable dependencies given the class variable. TAN is a tree structure learning algorithm that learns maximum spanning tree from attributes, but retains Naïve Bayes model as part of its structure based on conditional mutual distribution (using the Chow-Liu algorithm), which can be defined as:

$$Ip(X;Y \mid Z) = \sum_{x,y,z} P(x,y,z) \log \frac{P(x,y \mid z)}{P(x \mid z)P(y \mid z)} \quad (5)$$

where *x*, *y*, and *z* are the values of variables *X*, *Y*, and *Z*, respectively.

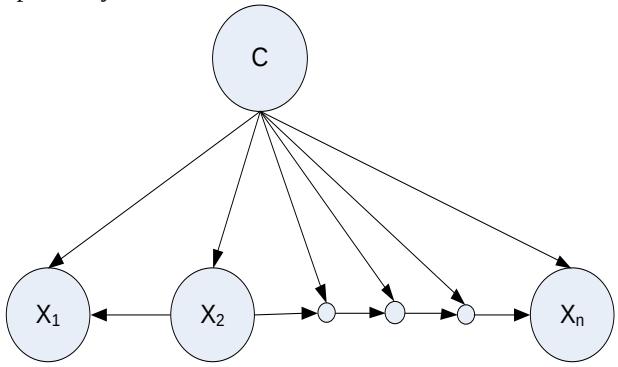


Fig. 2 a Simple TAN Structure

In TAN classification, the class node directly points to all other nodes and allows each attribute apart from the class node, to have at most one other attribute as a parent. The TAN learning process consists of five main steps:

1. Compute $I_{P_D}(A_i; A_j \mid C)$ between each pair of attributes, $i \neq j$.
2. Build a complete undirected graph in which nodes are the attributes $A_1, \ldots, A_n$ and annotate the weight of an edge connecting A_i to A_j by $I_{P_D}(A_i; A_j \mid C)$.
3. Then build a maximum weighted spanning tree.
4. Transform the resulting undirected tree to a directed one by choosing a root variable and setting the direction of all edges to be outward from it.
5. Construct a TAN model by adding a node labelled by C and adding an Arc from C to each A_i.

TAN classifier leads to an acceptable computational complexity and a considerable improvement over the simple Naïve Bayes [49]. It however adds a fixed number of edges irrespective of the distribution of the training data, thereby ignoring the influences from other attributes. In addition structure learning is also unavoidable with TAN classifier.

C. C4.5 Decision Tree

Decision trees are predictive divide-and-conquer techniques used in clustering and classification tasks. When a decision tree is deployed for classification purposes, the tree divides the search space into rectangular regions and a tuple is classified based on the region into which it falls [52], [53]. Reference [54] developed the C4.5 decision tree algorithm as an extension of the basic ID3 algorithm he earlier proposed. Reference [54] developed the C4.5 to address the problem of over fitting the data, trying to find a small decision tree, how to deploy reduced error pruning, rule post-pruning, handling sensibly continuous attributes and missing values, choosing an appropriate selection measure, and handling computational efficiency. C4.5 algorithm uses Shannon's entropy-based measure known as "information gain" as a criterion for selecting the most significant features [55] and defined as:

$$\text{Entropy (S)} = \sum_{i=1}^{c} (-p_i).\log_2(p_i) \quad (6)$$

where p_i is the proportion of the examples that belong to the i^{th} class.

At each decision node, available attributes are computed using the information gain, the attribute with the highest information gain will be chosen as the most significant attribute to split the given set of training data. This process is repeated until the data cannot be split any further. The main advantages of C4.5 algorithm as a classification tool, is that decision trees are self-explanatory and comprehensible, which makes it even more suitable for a novice to understand and use [10], [55]. They are capable of handling both discrete and continuous variables. They also have the capability of dealing with noisy data as well as missing values. The complexity of building a decision tree algorithm is quite straightforward to analyze [56].

However, the estimated complexity of the C4.5 algorithm increases drastically when there are many attributes and the number of numeric values present for each attribute in the training dataset [10]. In addition, C4.5 as a classification tool tends to perform less when there are many complex interactions in the training data, but they however perform well when there are few and highly relevant attributes in the dataset. Further, there is a tendency of oversimplifying the situation in order to be able to represent it as a tree-like graph [11], [53], [55], [56].

D. Artificial Neural Networks

Artificial Neural Networks (ANNs) are reasoning models that is based on the way the human brain processes information. They consists of a graph representing the processing system as well as various algorithms that accesses the graph [11], [13], [35], [17]. Artificial Neural Networks are structured as a directed graph with many nodes and arcs between them. The nodes in the graph are connected by individual numerical weighted links. These weights are the basic means of long-term memory in ANNs. They express the strength or the importance of each neuron input [11], [57]. Fig. 3 represents a simple ANN structure and fig. 4 represent an example of ANN for customers' E-banking risk exposure level.

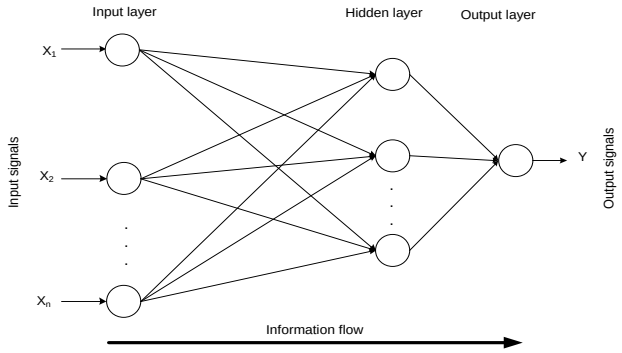


Fig. 3 A simple ANN with Three Layers

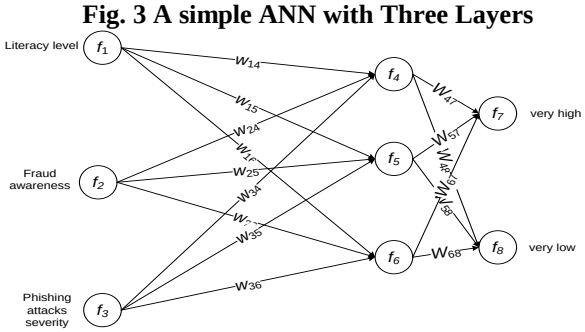


Fig. 4 A Neural Networks for Customers' E-Banking Risk Exposure Level

ANN approaches like decision trees require that a graphical structure be built to represent the model and the structure be applied to the data. In a data mining task, each input corresponds to a single attribute. For example, if the problem is to ascertain the risk exposure level to E-banking systems, some attributes could be customer's age, gender, income level, literacy level, fraud awareness, and phishing attacks severity. The numeric value of an attribute is the input to the network. Further, ANNs can have one input node for each attribute value to be examined unlike decision trees, which has only one input node, [58].

ANN may also be changed after a tuple is processed to improve future performance. The output node is the solution to the problem and it determines what the prediction is. For example, using figure 4, the prediction may be "very high" or "very low" in a binary class problem. The ANN assigns numeric values, such as +1 for "very high" and 0 for "very low". The purpose of the network is to learn and compute the values of the output. The ability to learn the network is determined by its architecture and by the algorithmic method chosen for training. Artificial Neural Networks have the advantages of adapting to unknown situations; they are robust with fault tolerance due to network redundancy and are capable of autonomous learning and generalization [11], [57]. However, major drawback to ANN is that it is considered to be a black box due to low comprehensibility problem inherent in the system. The computational complexity of the structure increases dramatically if the input data increases [11], [57]. ANN applications are also difficult to explain to end users [11], [15], [57], unlike a decision tree which is very easy to understand. Moreover in real time classification or predictions, large computational complexity usually hinders the processing speed both in training and testing. Different training algorithms have been proposed over the years for neural network, one of which is the multilayer perceptron (a feed forward neural network). In this research the multilayer perceptron was adopted because it is a well-established classifier and the most widely used.

III. SYSTEM DESIGN

This section describes the details of the case study and the customer-based survey conducted. Then we provide the data analysis and model development process of our experiment. We conducted a case study on two commercial banks in Nigeria currently providing customers with E-banking services, in order to generalize the results to the Nigeria E-banking system. The survey was conducted over a period of six months to achieve the research objectives. Data from the questionnaire provides a detailed quantitative report on the E-banking customers and their risk experiences. Specifically, the questionnaire was distributed to customers who visited the branch offices of the two understudied banks. Closed-ended questions were developed and formulated from the experimental findings of [59] and other literature. The questionnaire was designed to elicit information on E-banking risk awareness status, frequency of risk experiences, the severity of risks event occurrences, bank control/actions taken for reported risks experiences, and other E-banking risk contributors.

The questionnaire consisted of 120 questions which consisted of four parts: **Part 1** gives a brief profile of the respondent and their E-banking usage/opinion. **Part 2** addresses the users' E-banking risk awareness status and frequency of risks experienced within the last 12 months period. **Part 3** asks respondents to estimate qualitatively the severity of the risks experienced in part 2. In addition, the respondents were asked to identify the control mechanisms or actions that were taken by

The banks after their various risk experiences. Further, respondents were asked to rank some of the identified key risk indicators for E- banking operational risk. Following their risk experiences, respondents were asked to give their perception on their bank's E-banking exposure level. Finally, **Part 4** gives an indication about the perceived reasons for non E-banking adoption by non-adopters.

Due to data protection act and privacy policy, the names of the banks cannot be revealed. Thus, the names of the banks are represented by bank AAA and BBB. The sample size needed was calculated and achieved using the specified acceptable error level of 5%, confidence level of 90%, and estimated response rate of 50%. A total of 300 questionnaires were delivered to bank AAA and 250 to bank BBB. A response rate of 80% was achieved from bank AAA and a response rate of 50% from bank BBB was achieved. An overall total of 365 questionnaires were returned by customers who visited both banks. Of this total, 65 incomplete responses to questions were seen in the questionnaire returned by respondents of both banks. Major reasons for incomplete responses to some questions in this survey, was because customers refused to answer them, and others felt the questions were too many.

We used the dataset to evaluate and assess the soft computing classifiers, in order to choose the best classifier. The dataset was further applied to the TAN classifier to identify the risk issues and causal relationships inherent in the E-banking system under study.

A. Data Analysis

A range of statistical procedures was adopted to explore the research questions and objective. Prior to in-depth analysis, an initial data pre-processing was undertaken to optimize and clean the dataset collected. Reliability of scale was then carried out on the pre-processed dataset, in order to test consistency of the scale and ensure that the scale is measuring the same construct. In addition, factor analysis was undertaken to reduce the dimensionality of the dataset. These processes were considered relevant to optimize performance and prediction accuracy of the proposed soft computing algorithms, as every prediction algorithm has its weaknesses and strength [10], [11] such as capability to handle noise, missing data, various attribute coding and so on. The proposed TAN classifier was tested and compared with three (C4.5 decision tree, NB and ANN) soft computing tools, based on some statistical analysis evaluation metrics (CPU training / testing time, prediction accuracy, both mean absolute error and relative error, and Receiver Operating Characteristics Curve (ROC)). Fig. 5 illustrates the overall E-banking OR data analysis and model development process and fig. 6 present the detailed flowchart of the data analysis process.

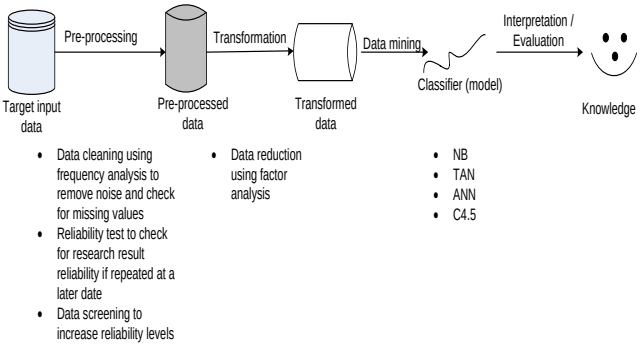


Fig. 5 E-Banking OR Data Analysis Model (Modified From: Dunham 2003)

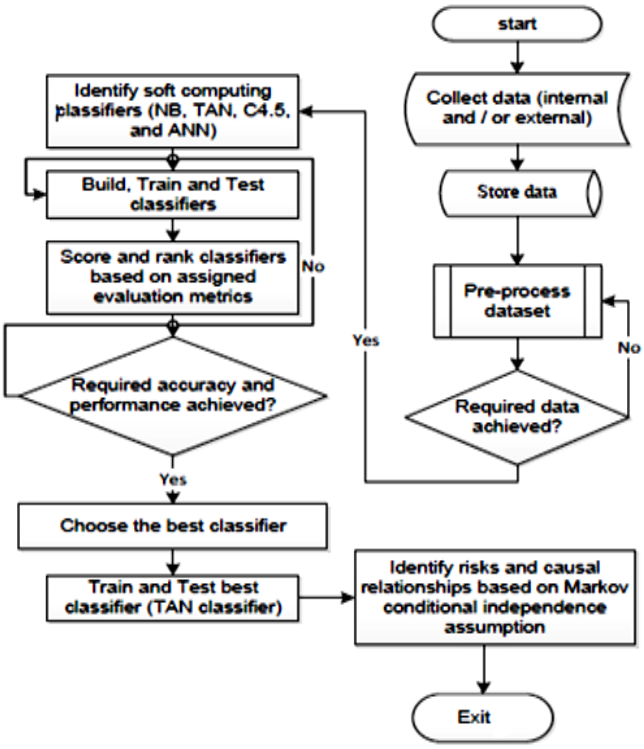


Fig. 6 Flowchart of the E-Banking OR Data Analysis Process

B. Data Pre-processing

The collected data from the first three part of the questionnaire had an initial 89 questions (target input attributes). In order to make the dataset more meaningful, and to provide more accurate results, the dataset was pre-processed and transformed by conducting a factor analysis using SPSS statistical measurement tool, to reduce the dimensionality of the dataset into a smaller subset of measurable variables. Here three aspects of data pre-processing: data cleaning, reliability test, and data screening was carried out. A frequency analysis was conducted for each variable to check for noisy data (such as invalid or incorrect values) and missing values. Missing values occur for several reasons, but in this case study missing values occurred because the survey respondents refused to answer some of the questions. Missing values were excluded by using list wise procedure because data were missing at random;

Meaning for any instance with missing value for any variable is excluded from calculation. In addition, 300 responses after exclusion of missing values is considered a good sample size [60], thus excluding missing values from calculation will not affect the analysis. Fig. 7 presents the data analysis steps.

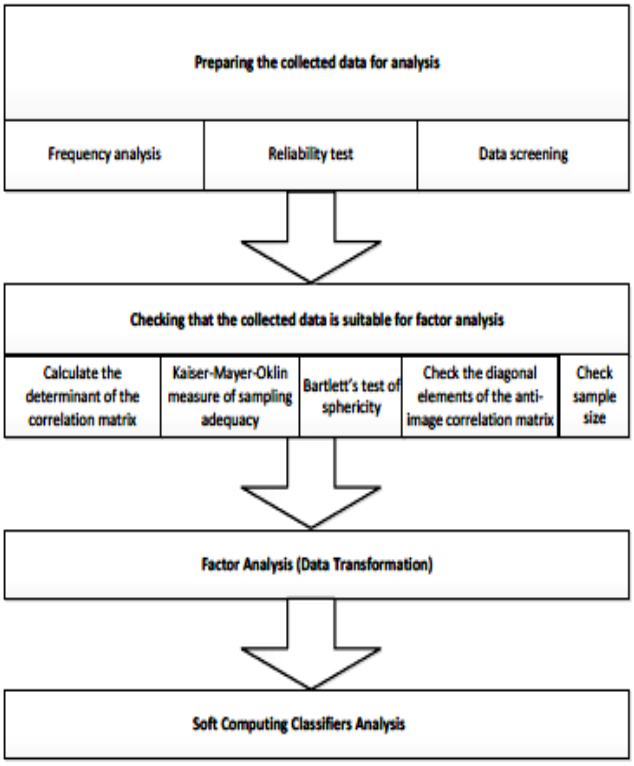


Fig. 7 Details of the Data Analysis Steps

Table 1. Relevant Assumption of the Factor Analysis and How These will Be Tested

Assumption / Issue	How tested
1 Is the sample size adequate for factor analysis to take place?	Obtain the values of the communalities. If all communalities are above 0.6 then samples less than 100 are adequate, if they are in the region of 0.5, samples between 100 and 200 are required.
2 Sampling adequacy	Kaiser-Meyer-Olkin (KMO) value has to be >=0.5 for the sample and for each variable
3 Retaining factors	Eigenvalues > 1
4 Interpretation of factors	Varimax rotation
5 Test for multicollinearity	R-matrix determinant is > 0.00001
6 Importance of given factor – retained if	Correlation coefficient > 0.4
7 Factors need deleting	Consider 3 stage factor analysis, or delete complex variables
8 Matrix is an identity matrix i.e. factor analysis is an appropriate method	Bartlett's test needs to be significant (at the 0.05 level)

C. Soft Computing Performance Evaluation Metrics

In this paper, we carried out a statistical analysis to assess the performance of the four soft computing classifiers for comparison. These statistics are explained here.

- **CPU Training Time**
  When a more complicated algorithm is used the power consumption proportionately increases, because it consumes more CPU cycles in each classification process. The time to build the entire model and classify results is also important for evaluating algorithms performance [55].
- **Prediction Accuracy**
  Prediction accuracy is considered because a correct outcome predicted with a higher degree of probability correctness weighs heavily on an outcome predicted with a lower degree of probability correctness. It is also considered when the prediction is subject to further analysis or perhaps serves as an input to second level learning process (as in this research), then it is important to take into consideration the prediction probabilities [10], [61], [62]. The accuracy of each classifier is evaluated as:

True Positive Rate (TPR)

$$\text{TPR} = \frac{Total\ no.of\ samples\ with\ correctly\ classfied\ outcome}{Total\ no.of\ samples} \quad (7)$$

False Positive Rate (FPR)

$$\text{FPR} = \frac{Total\ no.of\ samples\ with\ incorrectly\ classified\ outcome}{Total\ no.of\ samples} \quad (8)$$

False Negative Rate (FNR)

$$\text{FNR} = \frac{Total\ no.of\ samples\ incorrectly\ classified\ as\ false}{Total\ no.of\ samples} \quad (9)$$

False Positive Rate (FPR)

$$\text{FPR} = \frac{Total\ no.of\ samples\ incorrectly\ classified\ as\ true}{Total\ no.of\ samples} \quad (10)$$

Therefore,

Prediction accuracy

$$\text{Accuracy} = \frac{Total\ no.of\ samples\ correctly\ classified}{Total\ no.of\ samples} \quad (11)$$

- **Mean Absolute Error and Relative Absolute Error**
  An absolute error is the range of possible values in terms of the unit of measurement. The mean absolute error is the weighted average of all the absolute errors found from the cross validations. When considering performance measure an absolute error does not exaggerate the effect of outliers-instances whose prediction error is larger than the others, but instead it treats all sizes of individual errors evenly according to their magnitude and without taking into account their sign [62]. Relative absolute error on the other hand is just the total absolute error, with the same kind of normalization as absolute error [62]. However, sometimes the relative error values are of importance rather than the absolute error values. For example, if a 25% error is equally important whether it is an error of 60 in a prediction of 240 or an error of 42.5 in a prediction of 170, then averages of absolute error will be meaningless, while relative errors will be appropriate. This effect is usually taken into account by using the relative errors in the mean absolute error calculation.

The lower the percentage, the better the performance of the classifier compared to just predicting the mean [10]. Thus we present our statistical analysis using the relative absolute error method.

- **Receiver Operating Characteristics Curve**

A Receiver Operating Characteristics (ROC) curve is a two dimensional visualization graph of the FPR plotted on the horizontal axis against the TPR plotted on the vertical axis. The ROC curve represents the performance of a classifier without taken into consideration the class distribution or the cost of errors. The FRP and the TPR values are expressed as a percentage of the total number of positives and negatives included in the sample. The closer the ROC is to the point (0, 1) which is the upper left of the graph, the better the scoring rule in general [10], [61]. The ROC curve presents the decision-making behaviour of the scoring rule in terms of:

$$\text{ROC curve} = \text{TPR vs. FPR} = \left(\frac{\textit{False Positives}}{\textit{False Positives} + \textit{True Negatives}}, \frac{\textit{True Positives}}{\textit{True Positives} + \textit{False Negatives}} \right) \quad (12)$$

To summarize ROC in a single quantity, Area under the ROC (AUROC) is often used because, the larger the area the better the model (0, 1); 1 being optimal. It is also an appropriate performance measure because the area has a nice interpretation, and it is equivalent to the nonparametric Wilcoxon-Mann Whitney statistic which estimates the probability that the classifier ranks a randomly chosen positive data instance greater than a randomly chosen negative data instance [10], [61]. However this measure is only useful if costs and class distributions are unknown or vague. The AUROC result is reported based on the shaded area (convex hull) of the curves, and operates at the point that lies on the upper boundary of the convex hull.

Based on the factor analysis results, the four classifiers were trained on the dataset. In this research, ANN, C4.5, and two Bayesian network classifiers (TAN and NB) were applied to our E-banking operational risk dataset. WEKA was used as the developmental tool for training and testing the classifiers. Each classifier has a variety of parameter settings available. These parameters are listed and explained in Table A.1 below. The parameter settings that each classifier makes use of is indicated by 'X'.

Once each classifier was run against the dataset, the statistics of their performance are collated against the dataset. This detailed statistical comparison is a comprehensive way of analysing the performance of different classifiers.

IV. EXPERIMENTAL RESULTS

We pre-processed and transformed the dataset by conducting a factor analysis procedure using SPSS 19 statistical measurement tool. Then we trained and tested the four selected classifiers. We observed both the classification performance and ranking performance for each algorithm, measured by the soft computing evaluation metrics discussed above.

A. Frequency Analysis

We conducted a frequency analysis on each variable to check for incorrect data entry, missing values and major occurrences of parameters. From the 365 responses 49.3% of the participants were females, and 50.7% were males. Further, 74.5% of the respondents were E-banking adopters while 25.5% were non-adopters. Of the respondents that contributed, the majority 41.9% use ATM, 9.0% use Mobile phone, 7.9% use Internet, 2.5 % use Telephone, 0.8% use PC banking, while others use a combination of two or more E-banking types as shown in Fig. 8.

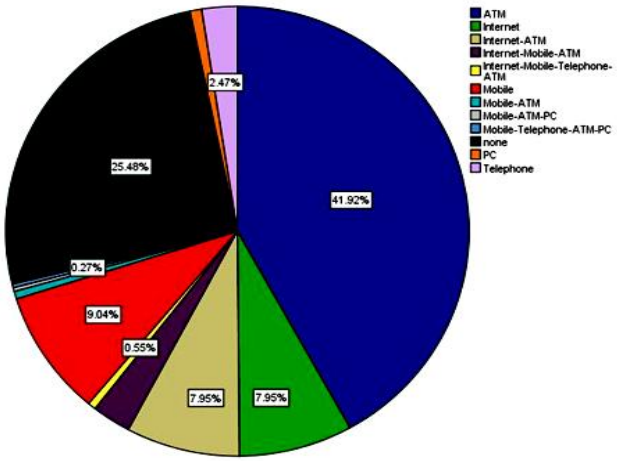


Fig. 8 Distribution of E-Banking System Adopters and Non-Adopters within the Collected Sample

Participants has various qualifications, 47.4 % had Bachelor's degree, 15.1% a Master's degree, 4.1 % holds PhD, 9.0% , 12.9 %, and 7.1% had HND, OND, and High school degree respectively, while the remaining 4.4 % has secondary school leavers degree as illustrated in Fig. 9. In addition, the majority of the respondents in the sample were computer literate.

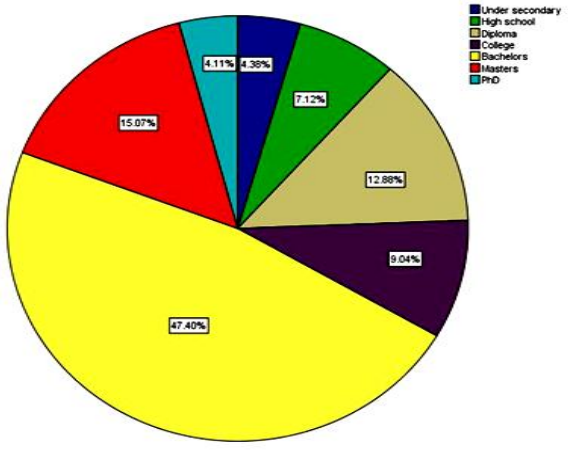


Fig. 9 Respondent Level of Education Attained within the Collected Sample

From the data it can be implied that on the average Nigerian income is between [34,000 – 62,999] naira, while the average age of respondents that contributed were between the age of 25 -34, which can therefore indicate some level of bias towards the age range and income levels. However, the survey reveals that majority of E-banking adopters and frequent users are within this age range,

Therefore helps to validate the survey; as the focus area is on E-banking users risk experiences. The results for variables frequency analysis in each dimension shows that the data is valid and ready to be analysed as missing values, noisy data and data screening were conducted during the frequency distribution analysis.

B. Reliability Test and Data Screening

Next a reliability test was conducted on the dataset, a process of checking the measure of the questionnaire for consistency in what it is measuring. That is the level to which individual items (or set of items) would produce the same results if the investigation was repeated at a later time or with a different sample. We used the Cronbach's coefficient alpha (α) in (13) to carry out testing and measuring the reliability of scale for each inter-item consistency.

Cronbach's α is:

$$\alpha = \frac{N^2 \overline{Cov}}{\sum s^2_{item} + \sum Cov_{item}} \quad (13)$$

Reliabilities less than 0.6 is considered poor, 0.7 is acceptable and greater than 0.8 is considered good. The closer the reliability coefficient is to 1.0 the better the reliability. The generally accepted value for Cronbach's alpha is 0.70. However, our initial reliability test gave some values (0.487, 0.608, 0.817, 0.843, 0.912, and 0.958) below accepted Cronbach's alpha within the six factors (*personal profile/opinion and E-banking perceptions, key risk indicators, actions taken for risks experienced, E-banking risk awareness status, frequency of E-banking risk experienced, and severity of E-banking risk experienced*) respectively. As a result a data screening process was conducted on the initial reliability test result (see [63]) to increase the reliability levels. Twelve attributes with low correlations coefficient [<0.30] between each item and the total score were deleted from the data analysis. This process increased the reliability level of Cronbach's alpha to 0.753, 0.810, 0.817, 0.832, 0.886 and 0.906 for all six factors.

C. Factor Analysis (Data Transformation)

Based on the reliability test results, the dataset was then transformed by conducting a factor analysis on 77 attributes with *orthogonal (verimax)* rotation to check the dimensionality of the construct (dataset). Factor loadings >0.30 are considered significant, >0.40 are more important, and >0.50 are very significant [64], and that there are no fixed cut-off value when interpreting factors; but dependent on the purpose of the study at hand [65]. We used 0.5 as the cut-off value for factor loadings and 1 for *eigenvalue*. This is because eigenvalues represent the amount of variation explained by a factor and an eigenvalue of 1 represents a substantial amount of variation. Moreover, Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy and Bartlett's test of sphericity suggested that all factors with eigenvalues greater than 1 should be retained. Table 2 illustrates the KMO measure of sampling adequacy and Bartlett's test of sphericity.

Table 2 KMO and Bartlett's Test

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.715
Bartlett's Test of Sphericity	Approx. Chi-Square	2412.403
	Df	325
	Sig.	.000

The KMO measure of sampling adequacy statistic was checked to ensure that KMO statistic lies between the value of 0 and 1, where a value close to 1 indicates that patterns of correlation are fairly clustered and as a result the factor analysis gives distinct and reliable factors. In addition to checking the overall KMO statistic, we checked the diagonal elements of the anti-image correlation matrix, to see that the KMO value for individual variables, are greater than 0.50 for all variables. Variables with values below 0.5 were found within the 77 attributes used in the analysis. These variables were then excluded, and a re-run of the factor analysis was conducted on the reduced attributes (26 attributes). These 26 attributes gave values in the range between [0.556 … 0.818] and the KMO value in our reduced dataset is 0.715. This number is considered good and the Bartlett's test of sphericity is less than 0.01, an indication that the correlations between items is highly significant. In addition, factor score (the Anderson-Rubin method) were then used to find an individual score on the subset of measures and to ensure no multicollinearity exists in the data. These shows that factor analysis is an appropriate method for our customer-based dataset and therefore there is confidence in the results obtained.

We carried out an initial analysis to obtain eigenvalues for each factor (component) in the data. Before extraction, SPSS identified 26 factors within the dataset and then extracted all factors with eigenvalues greater than 1; as a result 9 factors were obtained and in combination explained 68.01% of the variance. Table A.2 lists the *eigenvalues* associated with each factor before extraction, after extraction and after rotation. From the scree plot shown in Fig. 10, the point of inflexion on the curve on eleven components is in conformity with the results shown in Table A.2.

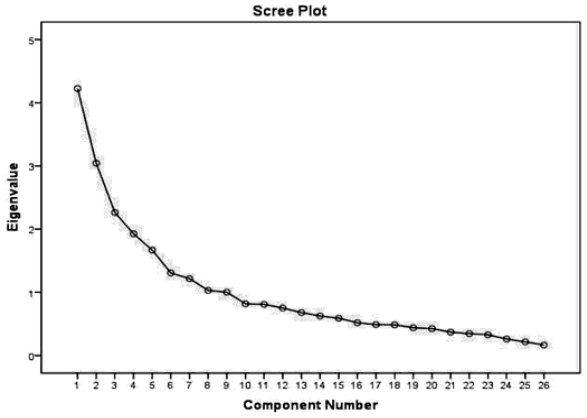


Fig. 10 Scree plot

Table A.3 shows the rotated factor (component) matrix, which is the matrix of the factor loadings for each variable onto each factor. Factor loadings less than 0.5 have not been displayed because these values have been suppressed. As a result the dimensionality of the dataset had been reduced by discarding these factor loadings less than 0.5, leaving 24 variables in total. The analysis shows that there are 24 important attributes as compared with the initial 77 attributes. However, E-banking types adopted attribute was excluded from the analysis, as it is the target output attribute. The transformed E-banking OR attributes selected based on the factor analysis result is shown in Fig. 11.

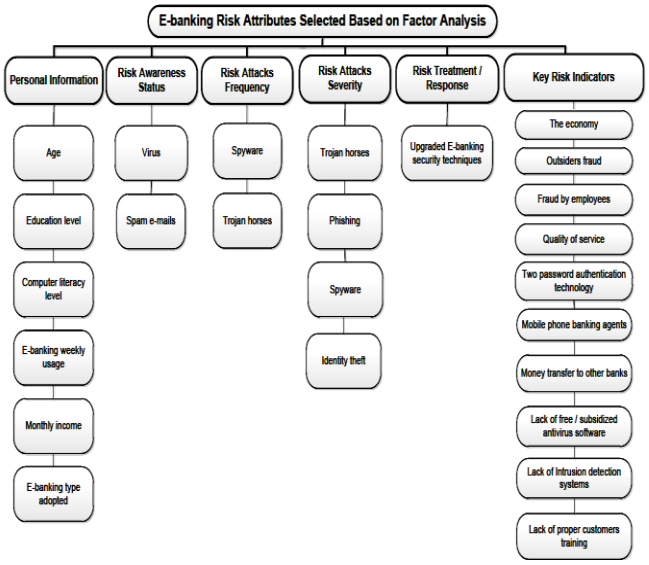


Fig. 11 Transformed E-Banking Risk Attributes based on Factor Analysis

D. Best Performing Soft Computing Classifier

Based on the factor analysis results, the four soft computing classifiers were applied on the attributes shown in Figure 1. In this research, ANN, C4.5, and two Bayesian network classifiers (TAN and NB) were applied to our E-banking operational risk dataset. The performances of the classifiers were evaluated using metrics described in section III. For evaluation purposes a 10 runs of 10-fold stratified cross validation was adopted for both training and testing of the dataset, in order to provide a true error estimates of the classifiers. In addition, cross validation procedure was also used because it is the widely accepted standard way of predicting the error estimates of a learning technique given a single, fixed sample of data. For each of the learning algorithm the same dataset of the same size was used to obtain an accuracy estimate of the dataset, using the stratified 10 fold cross-validation. The error rate is usually in percentage. The performances of these classifiers were collated statistically against the dataset in form of a table. Scores were assigned based on classifiers with the highest AUROC, the highest percentage of successful predictions on the cross-validation, the lowest CPU testing time, and the lowest mean and relative absolute error rate. Next, the classifiers were compared via two-tailed *t-test* with a 95% confidence level. Table 3 to 7 shows the accuracy, AUROC, mean absolute error, relative absolute error, CPU testing time and the results of the paired t-test with significance level 0.05, in which each entry *w / t / l* means that the model in the corresponding row wins in *w*, ties in *t*, and loses in *l* of the dataset, compared to the model in the corresponding column.

Table 3 Comparison of the Results of Two-Tailed T-Test on Accuracy with 95% Confidence Level

	ANN	NB	C4.5 (J48)
TAN	0/1/0	1/0/0	1/0/0
ANN		1/0/0	1/0/0
NB			0/1/0

Table 4 Comparison of The Results of Two-Tailed T-Test on AUROC with 95% Confidence Level

	ANN	NB	C4.5 (J48)
TAN	0/1/0	0/1/0	0/1/0
ANN		0/1/0	0/1/0
NB			0/1/0

Table 5. Comparison of the Results of Two-Tailed T-Test on Mean Absolute Error with 95% Confidence Level

	ANN	NB	C4.5 (J48)
TAN	0/1/0	0/0/1	0/0/1
ANN		0/0/1	0/0/1
NB			1/0/0

Table 6 Comparison of the Results of Two-Tailed T-Test on Relative Absolute Error with 95% Confidence Level

	ANN	NB	C4.5 (J48)
TAN	0/1/0	0/0/1	0/0/1
ANN		0/0/1	0/0/1
NB			1/0/0

Table 7 Comparison of the Results of Two-Tailed T-Test on CPU Testing Time with 95% Confidence Level

	ANN	NB	C4.5 (J48)
TAN	0/1/0	0/1/0	0/1/0
ANN		0/1/0	0/1/0
NB			0/1/0

The algorithm with the overall best performance across the five measurement metrics is assigned the ranking 'first to last'. If two or more algorithms had the same performance across the metrics, they were both given the same ranking.

Table 8 PC, AUROC, Mean absolute error, Relative absolute error, and CPU testing time and standard deviation

	PC	AUROC	Mean absolute error	Relative absolute error	CPU testing time in milliseconds	Classifier Ranking
TAN	72.79(5.95)	0.87(0.18)	0.07(0.01)	38.24(6.81)	0.00(0.01)	1
ANN	72.76(5.45)	0.82(0.28)	0.05(0.01)	38.71(6.79)	0.00(0.01)	3
C4.5	66.23(5.76)	0.80(0.27)	0.05(0.01)	60.47(5.75)	0.00(0.00)	2
NB	65.38(6.41)	0.85(0.23)	0.06(0.01)	49.90(7.19)	0.00(0.00)	4

The best performing algorithm for each parameter is underlined as shown in Table 8 for easy identification. However, when there are two or more algorithms with the same performance, they are both underlined. The experiment finding shows that the performance of TAN, not only in the classification, the AUROC, but also in the relative absolute error and the CPU training time, and it is the overall best performing classifier. Fig. A.1 shows the TAN classifier output and a graphical interface of the CPT for Outsider fraud as a KRI. Based on the TAN classifier performance it was then chosen and used for identifying the causal relationships and risk issues inherent in the E-banking system under study.

Next is the construction of the directed acyclic graph for the TAN structure that encodes assertions of conditional independence. However, computing the full joint probability grows exponentially with the number of variables, thus it will require large space to represent and high computational time complexity, but when the independence assumption is respected in the construction of the BN, the number of conditional probabilities to be evaluated can be reduced substantially as mentioned earlier in section II. Thus, our TAN structure is determined by ordering the variables from the parent node (E-banking system adopted), and determining the variables sets that satisfy the conditions in section II and in the following equation

$$p(x_i \mid x_i, \ldots, x_{i-1}) = p(x_i \mid \pi_i) \text{ for } i = 1, \ldots, n \quad (14)$$

Using the structure ordering (*Esa, Ths, ITs, Sua, Sps, Tp, Phs, MI, IDS, A, Edl, Ct, Mt, Thf, Spf, E-Bwu, Ccll, F-Ss, M-Ba, QoS, Vas, E, IF, Sma, and OF*), we have the conditional independencies as:

$p(Esa) = prior$
$p(Ths) = p(Ths \mid Esa)$
$p(ITs \mid Ths, Esa, Sua, Sps, Tp) = p(ITs \mid Esa, Ths)$
$p(Sua \mid Ths, Esa, ITs, Sps, Tp) = p(Sua \mid Ths)$
$p(Sps \mid Ths, Esa, ITs, Sua, Tp) = p(Sps \mid Ths)$
$p(Tp \mid Ths, Esa, ITs, Sua, Sps) = p(Tp \mid Ths)$

$p(Phs \mid Sps, Ths, Esa, ITs, Sua, Tp, Phs, IDS) = p(Phs \mid Esa, Sps)$

$p(MI \mid Sps, Ths, Esa, ITs, Sua, Tp, Phs, IDS) = p(MI \mid Sps)$
$p(IDS \mid Tp, Ths, Esa, ITs, Sua, Sps, Phs, MI,) = p(IDS \mid Tp)$

$p(A \mid MI, Sps, Ths, Esa, ITs, Sua, Tp, Phs, IDS, Edl, Ct, Mt) = p(A \mid MI)$

$p(Edl \mid MI, Sps, Ths, Esa, ITs, Sua, Tp, Phs, IDS, A, Ct, Mt) = p(Edl \mid MI)$

$p(Ct \mid IDS, Tp, Ths, Esa, ITs, Sua, Sps, Phs, MI, A, Edl, Mt) = p(Ct \mid IDS)$
$p(Mt \mid IDS, Tp, Ths, Esa, ITs, Sua, Sps, Phs, MI, A, Edl, Ct) = p(Mt \mid IDS)$
$p(Thf \mid A, MI, Sps, Ths, Esa, ITs, Sua, Tp, Phs, IDS, Edl, Ct, Mt, Spf, E–Bwu, Ccll, F–Ss, M–Ba) = p(Thf \mid Esa, A)\ p(Spf \mid A, MI, Sps, Ths, Esa, ITs, Sua, Tp, Phs, IDS, Edl, Ct, Mt, Thf, E–Bwu, Ccll, F–Ss, M–Ba) = p(Spf \mid Esa, A)\ p(E–Bwu \mid Edl, MI, Sps, Ths, Esa, ITs, Sua, Tp, Phs, IDS, A, Ct, Mt, Thf, Spf, Ccll, F–Ss, M–Ba) = p(E–Bwu \mid Edl)\ p(Ccll \mid Edl, MI, Sps, Ths, Esa, ITs, Sua, Tp, Phs, IDS, A, Ct, Mt, Thf, Spf, E–Bwu, F–Ss, M–Ba) = p(Ccll \mid Edl)$

$p(F–Ss \mid Ct, IDS, Tp, Ths, Esa, ITs, Sua, Sps, Phs, MI, A, Edl, Mt, Thf, Spf, E–Bwu, Ccll, M–Ba) = p(F–Ss \mid Ct)\ p(M–Ba \mid Mt, IDS, Tp, Ths, Esa, ITs, Sua, Sps, Phs, MI, A, Edl, Ct, Thf, Spf, E–Bwu, Ccll, F–Ss) = p(M–Ba \mid Mt)$

$p(QoS \mid F–Ss, Ct, IDS, Tp, Ths, Esa, ITs, Sua, Sps, Phs, MI, A, Edl, Mt, Thf, Spf, E–Bwu, Ccll, M–Ba, Vas) = p(QoS \mid F–Ss)\ p(Vas \mid F–Ss, Ct, IDS, Tp, Ths, Esa, ITs, Sua, Sps, Phs, MI, A, Edl, Mt, Thf, Spf, E–Bwu, Ccll, M–Ba, QoS) = p(Vas \mid F–Ss)$

$p(E \mid QoS, F–Ss, Ct, IDS, Tp, Ths, Esa, ITs, Sua, Sps, Phs, MI, A, Edl, Mt, Thf, Spf, E–Bwu, Ccll, M–Ba, Vas, IF, Sma) = p(E \mid QoS)\ p(IF \mid QoS, F–Ss, Ct, IDS, Tp, Ths, Esa, ITs, Sua, Sps, Phs, MI, A, Edl, Mt, Thf, Spf, E–Bwu, Ccll, M–Ba, Vas, E, Sma) = p(IF \mid QoS)\ p(Sma \mid QoS, F–Ss, Ct, IDS, Tp, Ths, Esa, ITs, Sua, Sps, Phs, MI, A, Edl, Mt, Thf, Spf, E–Bwu, Ccll, M–Ba, Vas, E, IF) = p(Sma \mid QoS)\ p(OF \mid E, QoS, F–Ss, Ct, IDS, Tp, Ths, Esa, ITs, Sua, Sps, Phs, MI, A, Edl, Mt, Thf, Spf, E–Bwu, Ccll, M–Ba, Vas, IF, Sma) = p(OF \mid E)$

Where

- *Esa* – E-banking system adopted
- *Ths* – Trojan horses severity
- *ITs* – identity theft severity
- *Sua* – security upgrade action
- *Sps* – spyware attacks severity
- *Tp* – two password use only as a key risk indicator
- *Phs* – phishing attacks severity
- *MI* – monthly income
- *IDS* – lack of intrusion detection system as a key risk indicator
- *A* – age
- *Edl* – education level
- *Ct* – lack of proper customers training as a key risk indicator
- *Mt* – money transfer from bank to bank as a key risk indicator
- *Thf* – frequency of Trojan horses occurrence
- *Spf* – frequency of spyware occurrence
- *E-Bwu* – E-banking weekly usage
- *Ccll* – customers computer literacy level
- *F-Ss* – lack of free or subsidized antivirus software as key risk indicator
- *M-Ba* – M-banking agents as a key risk indicator
- *QoS* – quality of service as a key risk indicator
- *Vas* – virus attacks awareness status
- *E* – economy as a key risk indicator
- *IF* – insider fraud as key risk indicator
- *Sma* – spam-email awareness status
- *OF* – outsider fraud as key risk indicator

The TAN graphical network structure for the E-banking operational risk issues is shown in Fig. 12.

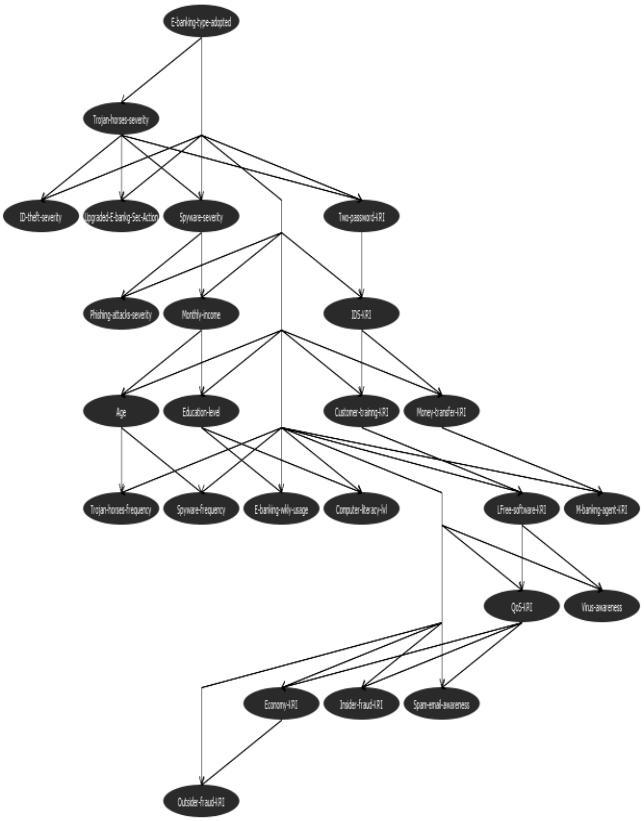


Fig. 12 The E-Banking Operational Risks TAN Structure

Although, using the independence assumption reduces substantially the number of conditional probabilities to be evaluated, choosing the variable ordering may fail to reveal many conditional independencies among the variables, thus to obtain the best and fully connected network structure the variable ordering should be explored n!. However, the exact or approximate inference is difficult when the network structure contains many undirected cycles, making the inference intractable.

To overcome these drawbacks, observing causal relationships among variables from experts or causal relationships that correspond to assertions of conditional dependence can be used. In this research, the interest is on learning and describing the causal relationships among the variables. The *causal Markov* assumption about the causal and probabilistic dependence (see [66]) was used. Given the causal Markov condition deployed in learning the TAN structure, we can infer the causal relationships from conditional independence and conditional dependence relationships for the E-banking OR data. The Inference of the causal relationships presents some interesting findings:

- Twelve risk attributes were seen as the operational risk issues of interest based on their conditional independence: *outsider fraud, insider fraud, Phishing attacks severity, Identity theft severity, Spyware attack frequency, Trojan horses frequency, E-banking weekly usage, computer literacy level, M-banking agents as key risk indicators, virus attack awareness, spam email awareness and upgraded E-banking security* as a control mechanism, which are directly influenced by the type of E-banking system adopted.
- Observing the E-banking users' weekly usage, banks aiming to both acquiring new customers and retaining existing customers may be interested in understanding their customers' level of education as a target market, as weekly usage is directly influenced by their level of education.

For reasons of computational efficiency, simplicity, and validation, the probabilistic inference is analyzed based on six conditional independences. Now we summarize these conditional independences structure as follows:

Given

1. Outsider fraud (*OF*), the independence structure is seen in the network as

$$p\left(\begin{array}{l} OF \mid E, QoS, F-Ss, Ct, IDS, Tp, Ths, Esa, ITs, Sua, Sps, Phs, \\ MI, A, Edl, Mt, Thf, Spf, E-Bwu, Ccll, M-Ba, Vas, IF, Sma \end{array}\right) = p\left(OF \mid Esa, E\right)$$

This means that *E* and *Esa* influences directly *OF*, while *QoS* influences *OF* indirectly through *E*.

2. Insider fraud (*IF*), the independence structure is seen in the network as

$$p\left(\begin{array}{l} IF \mid QoS, F-Ss, Ct, IDS, Tp, Ths, Esa, ITs, Sua, Sps, Phs, MI, \\ A, Edl, Mt, Thf, Spf, E-Bwu, Ccll, M-Ba, Vas, E, Sma \end{array}\right) = p\left(IF \mid Esa, QoS\right)$$

This means that QoS and Esa influences directly *IF*, while *F-Ss* influences *IF* indirectly through *QoS*.

3. Phishing attack severity (*Phs*), the independence structure is seen in the network as

$$p\left(Phs \mid Sps, Ths, Esa, ITs, Sua, Tp, Phs, IDS\right) = p\left(Phs \mid Esa, Sps\right)$$

This means that Sps and Esa influences directly *Phs*, while *Ths* influences *Phs* indirectly through *Sps*.

4. Identity theft severity (*ITs*), the independence structure is seen in the network as

$$p\left(ITs \mid Ths, Esa, Sua, Sps, Tp\right) = p\left(ITs \mid Esa, Ths\right)$$

This means that *Esa* and *Ths* influences directly *ITs*.

5. Spyware attack frequency (*Spf*) the independence structure is seen in the network as

$$p\left(\begin{array}{l} Spf \mid A, MI, Sps, Ths, Esa, ITs, Sua, Tp, Phs, IDS, \\ Edl, Ct, Mt, Thf, E-Bwu, Ccll, F-Ss, M-Ba \end{array}\right) = p\left(Spf \mid Esa, A\right)$$

6. Trojan horses attack frequency (*Thf*) the independence structure is seen in the network as

$$p\left(\begin{array}{l} Thf \mid A, MI, Sps, Ths, Esa, ITs, Sua, Tp, Phs, IDS, \\ Edl, Ct, Mt, Spf, E-Bwu, Ccll, F-Ss, M-Ba \end{array}\right) = p\left(Thf \mid Esa, A\right)$$

This means that both *Spf* and *Thf* are directly influenced by *Customers age* and *Esa*. They are also indirectly influenced through *MI*.

In more detail, Table 9 lists the identified conditional independent risks and risk factors associated with the E-banking system.

Table 9 TAN Conditional Independence Structure for Each Operational Risk Attribute

Identified risk attribute	Conditional independence structure based on TAN classifier	
	Direct influence	Indirect influence
Insider fraud	Quality of service	Free or subsidized anti-virus software
	E-banking type adopted	
Outsider fraud	Economy	Quality of service
	E-banking type adopted	
Phishing attacks severity	Spyware severity	Trojan horses
	E-banking type adopted	
Identity theft severity	Trojan horses severity	
	E-banking type adopted	
Spyware attack frequency	Customers age	Monthly income
	E-banking type adopted	
Trojan horses frequency	Customers age	Monthly income
	E-banking type adopted	

Based on the results reported so far, the indication is that types of E-banking system adopted by the customers, influences their various risk experiences and severity level of attacks. The direct factors and indirect factors affecting E-banking system are therefore identified.

V. CONCLUSION AND FUTURE WORK

This paper focused on identifying the key risk issues and causal relationships of E-banking OR from the customers' perspective. Based on the original data analysis, this research showed that E-banking users widely adopted the ATM system and majority of the respondents were holding bachelor's degree. The research also shows that gender related differences were not significant enough to be taken into account in the overall factor analysis.

A reliability test was conducted on the dataset to check the measure for consistency in what it is measuring. Some attributes however gave values below accepted Cronbach's alpha value, as a result they were deleted from the analysis. A factor analysis approach was used as a means of reducing and checking the dimensionality of the construct by conducting a principal component analysis on the attributes.

Finally, TAN classifier was used to identify some risk issues and causal relationships inherent in the Nigeria E-banking systems. The causal relationship capability of BN is widely responsible for why TAN is a more suitable approach for the E-banking OR assessment.

In future research, we will arrange in a questionnaire format the risk attributes identified in our experimental study to further assess the E-banking risks exposure level from the banks' view point. Specifically, we will evaluate the model using Fuzzy Inference System.

REFERENCES

1. Barnett, R. (2009) The Web Hacking Incidents Database (WHID). Breach Security Inc. Bi-Annual Report.
2. Bonsón E., Escobar, T., & Flores, F. (2008) Operational Risk Measurement in Banking Institutions and Investment Firms: New European Evidences. Financial Markets, Institutions and Instruments, 17(4), pp.287-307.
3. UK Payments Administration, (2011). Fraud Losses Drop on UK Cards, Cheques and Online Banking [online]. UK Payments Administration Press Releases. Available from: http://www.ukpayments.org.uk/media_centre/press_releases/-/page/1324/. [Accessed: 10th August 2011]
4. Financial Fraud Action UK, (2011). Fraud The Facts : The Definitive Overview of Payment Industry Fraud and Measures to Prevent It [online]. London: Financial Fraud Action UK. Available from:
5. http://www.financialfraudaction.org.uk/Publications/#/1/. [Accessed: 10th August 2011]
6. Basel Committee on Banking Supervision (2014) Review of Principles for the Sound Management of Operational Risk. [online]. Switzerland: Bank for International Settlements. Available from: https://www.bis.org/publ/bcbs292.pdf [Accessed: 27th March 2018]
7. British Standards Institution (2010) BS EN 31010. Risk Management – Risk Assessment Techniques. Geneva: International Organization of Standardization.
8. Basel Committee on Banking Supervision (2006). International Convergence of Capital Measurement and Capital Standards: A Revised Framework Comprehensive Version [online]. Switzerland: Bank for International Settlements. Available from: http://www.bis.org/publ/bcbs128.pdf [Accessed: 6th January 2009]
9. Adusei-Poku, K. (2005) Operational Risk Management - Implementing a Bayesian Network for Foreign Exchange and Money Market Settlement. PhD thesis, University of Gottingen.
10. Nash, D. & Murray, H. (2011) Using Monte-Carlo Simulations and Bayesian Networks to Quantify and Demonstrate the Impact of Fertiliser best Management Practices. Environmental Modelling and Software, 26(9), pp.1079-1088.
11. Witten, I. H., & Frank, E. (2005) Data Mining: Practical Machine Learning Tools and Techniques, 2nd edn., San Francisco: Elsevier.
12. Dunham, M. (2003) Data Mining Introductory and Advanced Topics. China: Pearson Education Asia Limited and Tsinghua University Press.
13. Zadeh, L. A. (1992) Fuzzy Logic for the Management of Uncertainty. New York: John Wiley.
14. Jang, J. R., Sun, C., & Mizutani, E. (1997) Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence. Upper Saddle River, NJ: Prentice Hall.
15. Li, X., Ruan, D., Van der Wal, A. J. (1998) Discussion on Soft Computing at FLIN'96. International Journal of Intelligent Systems, 13(2-3), pp.287-300.
16. Negnevitsky, M. (2002) Artificial Intelligence: A guide to Intelligent Systems. 1st edn., Essex: Pearson Education Limited.
17. Mogharreban, N. (2006) Adaptation of a Cluster Discovery Technique to a Decision Support System. Interdisciplinary Journal of Information, Knowledge, and Management, vol. 1, pp.59-68.
18. Venugopal, K. R., Srinivasa, K.G., & Patnaik, L. M. (2009) Soft Computing for Data Mining Applications. New York: Springer.
19. Chakraborty, R. C. (2010) Fuzzy Systems: Soft Computing Course Lecture 35 – 36, notes, slides [online]. Available from:
20. http://www.myreaders.info/07_Fuzzy_Systems.pdf [Accessed: 4th February 2012]
21. Giudici, P. (2004) Integrating Quantitative and Qualitative Operational Risk Data: A Bayesian Approach. In: Cruz, M., ed. Operational Risk Modelling and Analysis. London: Risk Books, pp.131-138.

22. Neil, M., Fenton, N. E., & Tailor, M. (2005) Using Bayesian Networks to Model Expected and Unexpected Operational Losses. Risk Analysis, 25(4), pp.963-972
23. Marquez, D. (2008) Using Bayesian networks to model the operational risk to information technology infrastructure in financial institutions. Computer, 22, pp.131-138.
24. Liu, J. & Liu, R. (2011) Using Bayesian Networks to Model Operational Risk of Bank. International Conference on Computer and Management – CAMAN, Conference Proceedings, Wuhan, China: IEEE, pp.1-4.
25. Shevchenko, P. V. (2011) Modelling Operational Risk Using Bayesian Inference. Media, pp.235-271.
26. Aquaro, V., Bardoscia, M., Bellotti, r., Consiglio, A., De Carlo, F., & Ferri, G. (2012) A Bayesian Networks Approach to Operational Risk. Submitted to Journal of Physica A389(2010), pp.1721-1728.
27. King, J. L. (2001) Operational Risk: Measurement and Modelling. New York: Wiley.
28. Alexander, C. (2003) Operational Risk, Regulation Analysis and Management. London: Pearson Education.
29. Ramamurthy, S., Arora, H., & Ghosh, A. (2005) Operational Risk and Probabilistic Networks - An Application to Corporate Actions Processing [online]. Available from: http://www.infosys.com/industries/financial-services/white-papers/Documents/operational-risk-probabilistic-networks.pdf [Accessed: 8th March 2008]
30. Sanford, A. D., & Moosa, I. A. (2011) A Bayesian Network Structure for Operational Risk Modelling in Structured Finance Operations. Journal of the Operational Research Society, pp.1-14.
31. Antonucci, A., Piatti, A., Zaffalon, M.: Credal networks for operational risk measurement and management. In: Apolloni, B., Howlett, R.J., Jain, L. (eds.) KES 2007, Part II. LNCS (LNAI), vol. 4693, pp. 604–611. Springer, Heidelberg (2007)
32. Shah, S. (2003) Measuring Operational Risks using Fuzzy Logic Modelling [online]. Available from: http://www.prmia.org/Chapter_Pages/Data/WashingtonDC/Shah_Paper_1_26_05.PD F [Accessed: 6th January 2012]
33. Reveiz, A., & Leon, C. (2010) Operational Risk Management using a Fuzzy Logic Inference System. Journal of Financial Transformation, 30(574), pp.141-153
34. Aburrous, M, Hossain, M. A., Dahal, K., & Thabtah, F. (2010) Intelligent Phishing Detection System for E-banking using Fuzzy Data Mining. Journal of Expert Systems with Applications, 37(12), Las Vegas, NV, IEEE, pp.7913-7921
35. Durfee, A., & Tselykh, A. (2011) Evaluating Operational Risk Exposure Using Fuzzy Number Approach to Scenario Analysis. Advances in Intelligent Systems Research - EUSFLAT, 1(1), pp.1045-1051
36. Mahant, N. (2004) Risk Assessment is Fuzzy Business- Fuzzy Logic Provides the Way to Assess Off-site Risk from Industrial Installations. Bechtel Corporation Technical paper: 2004-206.
37. McGill, W. L. & Ayyub, B. M. (2007) Multicriteria Security System Performance Assessment Using Fuzzy Logic. The Journal of Defense Modelling and Simulation: Applications, Methodology, Technology. 4(4), pp.356-376.
38. Sadiq, R., Kleiner, Y., & Rajani, B. (2007) Water Quality Failures in Distribution Networks – Risk Analysis using Fuzzy Logic and Evidential Reasoning. Risk Analysis, 27(5), pp.1381-1394.
39. Pokorádi, L. (2009) Risk Assessment Based Upon Fuzzy Set Theory. 15th "Building Services, Mechanical and Building Industry Days", International Conference, Debrecen, Hungary, 15-16th October 2009, pp.311-318.
40. Aburrous, M, Hossain, M. A., Thabtah, F. Dahal, K., (2008) Intelligent Quality Performance Assessment for E-banking Security using Fuzzy Logic. Fifth International Conference on Information Technology: New Generations - ITNG, Conference Proceedings, 7-9th April 2008, pp.420-425.
41. Zlateva, P., Pashova, L., Stoyanov, K., & Velev, D. (2011) Fuzzy Logic Model for Natural Risk Assessment in SW Bulgaria. 2nd International Conference and Management Technology – IPCSIT, Singapore, IACSIT Press, pp.109-113.
42. Chen, Q., & Wen, Y. (2010) A BP-neural Network Predictor Model for Operational Risk Losses of Commercial Bank. Third International Symposium on Information Processing IEEE Computer Society
43. Balan, C. (2009) The use of Neural Networks in the Operational Risk Data Modelling. Proceedings of the 4th International Conference on Knowledge Management: Projects, Systems and Technologies, November 6-7, pp.225-227.
44. Jiang, H. (2009) The Application of Artificial Neural Networks in Risk Assessment on High-tech Project Investment. International Conference on Business Intelligence and Financial Engineering
45. Pradhan, B., & Lee, S. (2009) Landslide Risk Analysis using Artificial Neural Network Model Focusing on Different Training Sites. International Journal of Physical Sciences, 4(1), pp.1-15.
46. Shankarapillai, R., Mathur, L.K., Nair, M.A., Rai, N., & Mathur, A. (2010) Periodontitis Risk Assessment using Two Artificial Neural
47. Networks-A Pilot Study. International Journal of Dental Clinics, 2(4), pp.36-40.
48. Koyuncugil, A.S., & Ozgulbas, N. (2009) Risk Modelling by CHAID decision Tree Algorithm. International Conference on Computational & Experimental Engineering and Sciences, 11(2), pp.39-46.
49. Mosquera, N., Reneses, J., & Sánchez-Úbeda, E. F. (2008) Medium-term Risk Analysis in Electricity Markets: A Decision-tree Approach. International Journal of Energy Sector Management, 2(3), pp.318 – 339.
50. Lin, J.W., Hwang, M.I., & Becker, J.D. (2003) A Fuzzy Neural Network for Assessing the Risk of Fraudulent Financial Reporting. Managerial Auditing Journal, 18(8), pp.657-665.
51. Berta, I. (2009) Use of Soft Computing Methods in Risk Assessment of Electrostatic Fire and Explosion Hazards in Industries. Journal of Electrostatics, 67(2-3), pp.235-241.
52. Park, S., Kim, B., Choi, B., Kim, E., Lee, H., & Kang, H. (2011) A Soft Computing Approach for Collision Risk Assessments. The 11th International Conference on Control, Automation and Systems (ICCAS), 26-29oct 2011, Gyeonggi-do, Korea.
53. Jiang, L., Zhang, H., & Cai, Z. (2009). A Novel Bayes Model: Hidden Naïve Bayes. IEEE Transactions on Knowledge and Data Engineering, 21(10), 1361-1371.
54. Friedman, N., Geiger, D., & Goldszmidt, M. (1997) Bayesian Network Classifiers. Machine Learning, 29, pp. 131-163.
55. Vaidya, J. S. (2004) Privacy Preserving Data Mining over Vertically Partitioned Data. PhD thesis, Purdue University
56. Ratanamahatana, C. A., & Gunopulos, D. (2002) Scaling up the Naïve Bayesian Classifier: Using Decision Trees for Feature Selection. IEEE International Conference on Data Mining - ICDM 2002, Japan, 9-12th December, IEEE, pp.131-141.
57. Quinlan, J. R. (1993), C4.5: Programs for Machine Learning, Morgan Kaufmann, Los Altos, California
58. Yang, X. (2010) A Wearable Real-Time System for Physical Activity Recognition and Fall Detection. Master of Science Thesis, University of Saskatchewan.
59. Patil, D. D., Wadhai, V. M., & Gokhale, J. A. (2010) Evaluation of Decision Tree Pruning Algorithms for Complexity and Classification Accuracy. International Journal of Computer Applications, 11(2), pp.23-30.
60. Negoita, M., Neagu, D.,& Palade, V. (2005) Computational Intelligence: Engineering of Hybrid Systems. Heidelberg: Springer-Verlag.
61. Turban, E., & Aronson, J.E. (1998) Decision Support Systems and Intelligent Systems. 5th edn., London: Prentice-Hall.
62. Ochuko, R. E., Cullen, J. A. & Neagu, D. (2009) Overview of Factors for Internet Banking Adoption. International Conference on CyberWorlds - CW, Conference Proceedings, Bradford, 7-11th September 2009, Washington, DC: IEEE Computer Society, pp. 163-170.
63. Comrey A. L. & Lee, H. B. (1992) A First Course in Factor Analysis. 2nd ed., New Jersey: Lawrence Erlbaum Associates, Inc.
64. Viaene, S., Derrig, R.A., Dedene, G. (2004) A Case Study of Applying Boosting Naïve Bayes to Claim Fraud Diagnosis. IEEE Transactions on Knowledge and Data Engineering, 16(5), pp. 612-620.
65. Moran, S., He, Y., & Liu, K. (2009) Choosing the Best Bayesian Classifier: An Empirical Study. IAENG International Journal of Computer Science, 36(4), pp. 322-331.
66. Churchill, J. (1999) Marketing Research: Methodological Foundation. 7th edn., Hinsdale, IL: The Dryden Press.
67. Haire, J. F., Tatham, R. L., Anderson, R. E., & Black, W. (1998) Multivariate Data Analysis: With Readings. 5th edn., London: Pearson Education.
68. Field, A. (2009) Discovering Statistics using SPSS. 3rd edn, London: SAGE Publications Ltd.
69. Heckerman, D. (1996) A Tutorial on Learning with Bayesian Networks. Technical Report: 1996- MSR-TR-95-06.

Appendix Table A.1 Parameter Settings for the Soft Computing Tools

Parameters	Parameter Description	TAN	NB	C4.5	ANN
Simple Estimator	- Used for estimating the conditional probability tables of a Bayes network once the structure has been learned. Estimates probabilities directly from data. - alpha used for estimating the probability tables and can be interpreted as the initial count on each value. The default value is 0.5	X			
Markov Blanket Classifier	- when set to true, a Markov Blanket correction is applied to the network structure after a network structure is learned. - it ensures all nodes in the network are part of the Markov blanket of the classifier node	X			
Entropy	- entropy was used as the score type determines the measures used to judge the quality of a network structure - can be BDeu, Minimum Desciption Length (MDL), Akaike Information Criterion (AIC), and Entropy	X			
initAs Naïve Bayes	- When set to true, the initial structure used for learning is a Naïve Bayes network. - When set to false, an empty network is used as the initial network structure.	X	X		
Use AD Tree	- improve search speed when set to true	X			
Use Kernel Estimator	- designed for use with numeric attributes rather than a normal distribution when set to true - it implies that differences in performance will be seen based on the number of numeric values available		X		
Use Supervised Discretization	- used to convert numeric attributes to nominal attributes		X		
Confidence Factor	- The confidence factor used for pruning (smaller values incur more pruning)			X	
Min Num Obj	- The minimum number of instances per leaf			X	
Num Folds	- Determines the amount of data used for reduced-error pruning. One fold is used for pruning, the rest for growing the tree			X	
seed	- The seed is used for randomizing the data when reduced-error pruning is used - Used to initialise the random number generator. Random numbers are used for setting the initial weights of the connections between nodes, and also for shuffling the training data			X	
Subtree Raising	- Whether to consider the subtree raising operation when pruning			X	
Auto Build	- adds and connects up hidden layers in the network				X
Hidden Layers	- defines the hidden layers of the neural network. - It is a list of positive whole numbers. 1 for each hidden layer. To have no hidden layers 0 is specified. This will only be used if auto build is selected. There are also wildcard values 'a' = attribs + classes) / 2, 'i' = attribs, 'o' = classes, 't' = attribs + classes				X
Learning Rate	- The amount the weights are updated				X
momentum	- It is the momentum applied to the weights during updating				X
Nominal To Binary Filter	- This will pre-process the instances with the filter. - It could help to improve the performance if there are nominal attributes in the data.				X
Normalize Attributes	- It normalizes the attributes and help to improve the performance of the network. - It is however not reliant on the class being numeric. - It normalizes the nominal attributes after they have been run through the nominal to binary filter so that the nominal values are between -1 and 1				X
Mormalizenumerical Class	- It normalizes the class if it is numeric. - It could help to improve the performance of the network and normalizes the class to be between -1 and 1. internally, - the output is then scaled back to the original range				X
Training Time	- The number of epochs to train through. If the validation set is non-zero then it can terminate the network early				X

Validation Set Size	- The percentage size of the validation set. - The training will continue until it is observed that the error on the validation set has been consistently getting worse, or if the training time is reached				X
Validation Threshold	- Used to terminate validation testing. - The value here dictates how many times in a row the validation set error can get worse before training is terminated				X

APPENDIX: Table A.2 Total Variance Explained

Total Variance Explained									
Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.226	16.252	16.252	4.226	16.252	16.252	2.848	10.955	10.955
2	3.044	11.706	27.958	3.044	11.706	27.958	2.603	10.012	20.967
3	2.262	8.702	36.660	2.262	8.702	36.660	2.517	9.682	30.649
4	1.926	7.408	44.068	1.926	7.408	44.068	2.084	8.016	38.665
5	1.669	6.417	50.486	1.669	6.417	50.486	2.052	7.891	46.555
6	1.307	5.026	55.512	1.307	5.026	55.512	1.595	6.135	52.691
7	1.216	4.679	60.191	1.216	4.679	60.191	1.437	5.528	58.219
8	1.030	3.961	64.152	1.030	3.961	64.152	1.437	5.527	63.746
9	1.002	3.855	68.007	1.002	3.855	68.007	1.108	4.261	68.007
10	.818	3.147	71.154						
11	.809	3.111	74.265						
12	.752	2.891	77.155						
13	.679	2.610	79.765						
14	.625	2.405	82.170						
15	.590	2.271	84.441						
16	.518	1.991	86.432						
17	.490	1.883	88.315						
18	.485	1.867	90.182						
19	.440	1.691	91.873						
20	.425	1.636	93.509						
21	.371	1.425	94.935						
22	.347	1.333	96.268						
23	.327	1.259	97.527						
24	.261	1.006	98.532						
25	.216	.831	99.363						

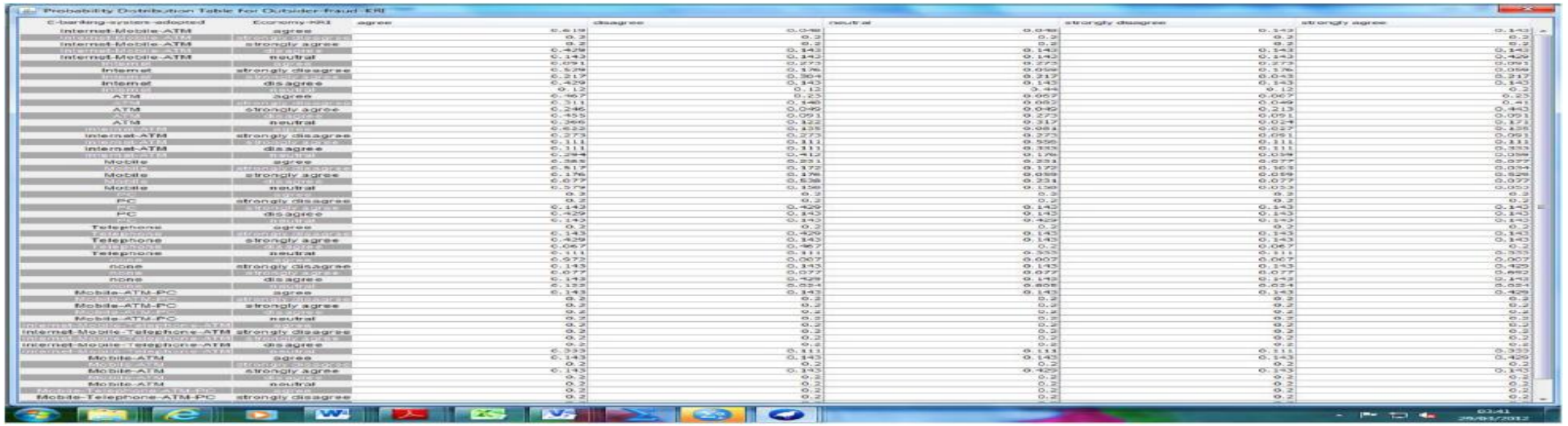
26	.166	.637	100.000						
Extraction Method: Principal Component Analysis.									

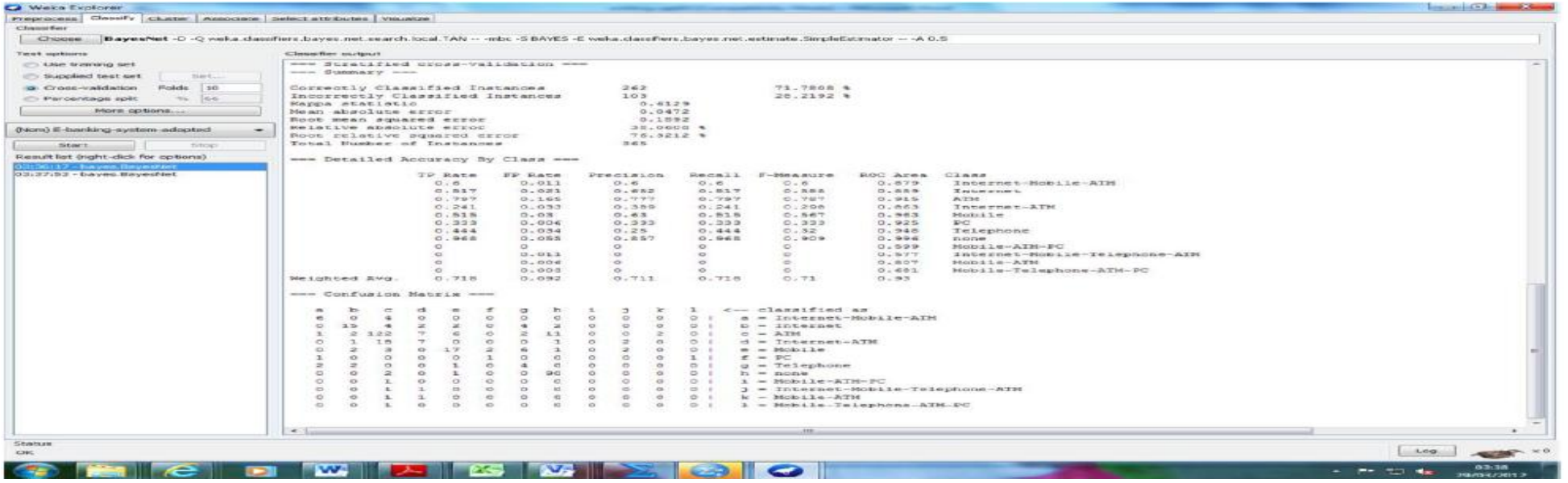
APPENDIX: Table A.3 Rotated Component Matrixa factor loadings for each attribute

Rotated Component Matrix[a]									
	Component								
	1	2	3	4	5	6	7	8	9
Monthly income	.813								
Weekly E-banking usage	.748								
Customers age range	.709								
Computer literacy level	.664								
Level of education attained	.635								
Number of Trojan horses attacks experienced		.833							
Number of spyware attacks experienced		.769							
Virus attack awareness		.617							
Trojan horses attack severity		.557							
Transferring of money to other banks			.735						
Insider fraud as key risk indicator			.726						
Quality of service as key risk indicator			.706						
Phishing attack severity				.830					
Spyware attack severity				.677					
Identity theft severity				.529		.524			
Lack of customers proper training as key risk indicator					.813				
Lack of Free or subsidized antivirus software as key risk indicator					.801				
Lack of intrusion detection systems as key risk indicators					.663				
Economy growth as key risk indicator						.862			
Regular staff training as key risk indicator									
Upgraded the E-banking security techniques							.809		
Outsider fraud as key risk indicator							.553		
Spam E-mails awareness								.791	
Two password authentication only								.636	
Mobile phone banking agents									.699
Number of server failure experienced									
Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization.									
a. Rotation converged in 10 iterations.\									

APPENDIX B: Fig. A.1 Graphical Interface of a CPT and TAN Classifier Output

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication


Probability Distribution Table for Outsider Fraud-ERI


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
BayesNet -D -Q weka.classifiers.bayes.net.search.local.TAN -- -mbc -S BAYES -E weka.classifiers.bayes.net.estimate.SimpleEstimator -- -A 0.5
Test options
Use training set
Supplied test set
Cross-validation
Folds
Percentage split
More options...
(Nom) E-banking-system-adopted
Start
Stop
Result list (right-click for options)
03:36:17 - bayes.BayesNet
03:37:52 - bayes.BayesNet
Classifier output
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 262 71.7808 %
Incorrectly Classified Instances 103 28.2192 %
Kappa statistic 0.6129
Mean absolute error 0.0472
Root mean squared error 0.1592
Relative absolute error 38.0668 %
Root relative squared error 76.9212 %
Total Number of Instances 365
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure ROC Area Class
0.6 0.011 0.6 0.6 0.6 0.879 Internet-Mobile-AIM
0.517 0.021 0.652 0.517 0.577 0.888 Internet
0.797 0.165 0.777 0.797 0.787 0.915 AIM
0.241 0.033 0.359 0.241 0.288 0.863 Internet-AIM
0.515 0.03 0.63 0.515 0.567 0.983 Mobile
0.333 0.006 0.333 0.333 0.333 0.925 PC
0.444 0.034 0.25 0.444 0.32 0.948 Telephone
0.948 0.055 0.857 0.948 0.909 0.996 none
0 0 0 0 0 0.599 Mobile-AIM-PC
0 0.011 0 0 0 0.577 Internet-Mobile-Telephone-AIM
0 0.004 0 0 0 0.807 Mobile-AIM
0 0.003 0 0 0 0.681 Mobile-Telephone-AIM-PC
Weighted Avg. 0.718 0.092 0.711 0.718 0.71 0.93
=== Confusion Matrix ===
a b c d e f g h i j k l <-- classified as
6 0 4 0 0 0 0 0 0 0 0 0 | a = Internet-Mobile-AIM
0 15 4 2 0 0 0 0 0 0 0 0 | b = Internet
1 2 122 7 6 0 2 11 0 0 2 0 | c = AIM
0 1 15 7 0 0 0 1 0 3 0 0 | d = Internet-AIM
0 2 3 0 17 2 6 1 0 2 0 0 | e = Mobile
1 0 0 0 0 7 1 0 0 0 0 1 | f = PC
2 2 0 0 1 0 4 0 0 0 0 0 | g = Telephone
0 0 2 0 1 0 0 90 0 0 0 0 | h = none
0 0 1 1 0 0 0 0 0 0 0 0 | i = Mobile-AIM-PC
0 0 1 1 0 0 0 0 0 0 0 0 | j = Internet-Mobile-Telephone-AIM
0 0 1 1 0 0 0 0 0 0 0 0 | k = Mobile-AIM
0 0 1 0 0 0 0 0 0 0 0 0 | l = Mobile-Telephone-AIM-PC
Status
OK
Log