

**KWAME NKRUMAH UNIVERSITY OF SCIENCE AND TECHNOLOGY,
KUMASI**

**COLLEGE OF SCIENCE
DEPARTMENT OF COMPUTER SCIENCE**

**FRAUD DETECTION PROCESS IN DATA MINING USING ADAPTIVE
AND RULE LEARNING ALGORITHM TECHNIQUES**

**BY
KORNYO OLIVER**



**A THESIS SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE,
KWAME NKRUMAH UNIVERSITY OF SCIENCE AND TECHNOLOGY, IN
PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF
MASTER OF INFORMATION TECHNOLOGY.**

MARCH, 2014.

DECLARATION

I hereby declare that this submission is my own work towards the award of Msc. (Information Technology) degree and that, to the best of my knowledge, it contains no material previously published by another person or material which had been accepted for the award of any other degree of the University, except where due acknowledgement had been made in the text.

KORNYO OLIVER (PG6557711)

(Student Name and ID)



Signature

13/03/14

Date

Certified by:

Dr. Michael Asante

(Supervisor)



Signature

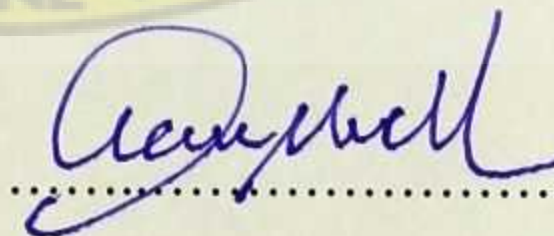
28/03/14

Date

Certified by:

Dr Michael Asante

(Head, Department of Computer Science, KNUST)



Signature

28/03/14

Date

LIBRARY
KWAME NINSIN UNIVERSITY
SCIENCE AND TECHNOLOGY
KUMASI-GHANA

DEDICATION

This work is dedicated to my lovely and caring Wife and daughter (Selina and Esther) for their continuous cooperation and support in my life.

KNUST



ACKNOWLEDGEMENT

My sincere thanks go to the Almighty God for His protection and guidance through the successful completion of the course.

A hand shake that exceeds the elbow is no more a hand shake but a moment of thanks given to my supervisor and the same my Head of Department Dr. Michael Asante for spending a lot of his time reading this thesis, criticizing where necessary and offering very good suggestions and pieces of advice.

My deepest and profound gratitude to Dr Kwasi Preko for his guidance, encouragement and suggestions in this project.

Also I wish to thank all the staff of the Electricity company of Ghana Limited, especially Mr. Jones Talor(ASH- West (DBA)), Mr Nelson Adziman (Accra west(DBA)), Mr. Gilbert Abosti (Volta (DBA)), Miss kathren Nico-anna(Tema(DBA)), Miss Ewurasi(Eastern(DBA)) for their support in the data collection processes.

Further, I wish to express my appreciation to Mrs. Celestina Lordina Rockson. In addition, all my brothers and sisters for being there for me.

In addition, I wish to express my gratitude to my entire church member at SRBC- Kumasi and SRBC-Lebanon for their spiritual encouragement.

Finally, my gratitude go to all the MIT final year students of the Computer Science Department and to Evelyn Kornyo who have in diverse ways contributed to this work but their names have not been mentioned, I deeply appreciate you all.

God bless you all.

ABSTRACT

The main objective of this study was to apply adaptive and rule learning techniques in data mining for fraud detection in Electricity Company of Ghana (ECG). This work Combines the rule learning techniques in supervised and unsupervised fraud detection in data set and considers the behaviour of data set patenting to identify a suspected fraud for further investigation The rule-learning (supervised and unsupervised) and adaptive techniques were use to uncover indicators of fraudulent behaviour from a large database of customer transaction dataset. Then the indicators were used to create a set of monitors, which profile legitimate customer behaviour and indicate anomalies. Finally, the outputs of the monitors are used as features in a system that learns to combine evidence to generate high-confidence alarms. The system has been applied to the problem of detecting fraud based on a database of records. The findings indicated that there were always some amount of fraud in every pool of data set, thus processed data already in their system or data thus yet to be updated. It also showed that, these algorithms, can be used to test for existence of fraud in any data set with three unique attributes like account number, receipt number and transaction identification(id).

Furthermore, it was found that as the populations of data set increases, the timing of the algorithm to detect fraud in a particular data for investigation also increases.

TABLE OF CONTENT

DECLARATION	I
DEDICATION	II
ACKNOWLEDGEMENT	III
ABSTRACT.....	IV
CHAPTER 1.....	1
INTRODUCTION	1
1.2 PROBLEM STATEMENT	4
1.3 AIMS AND OBJECTIVES OF THE RESEARCH	5
1.4 RESEARCH QUESTIONS.....	5
1.5 JUSTIFICATION OF OBJECTIVES.....	6
1.6 METHODOLOGY.....	6
CHAPTER 2.....	9
LITERATURE REVIEW.....	9
2.2 MOST EFFECTIVE SKILLS AT DETECTING TOP FRAUD RISKS.....	22
CHAPTER 3.....	29
MATERIALS AND METHODS	29
3.1 DESCRIPTION OF PROJECT SITE	29
3.1.1 ADAPTIVE ALGORITHMS.....	29
3.1.2 SUPERVISED LEARNING METHOD (DIRECTED DATA MINING).....	30
3.1.3 UNSUPERVISED LEARNING.....	31
3.1.4 Process for Algorithms Techniques of Adaptive and Rule Learning Used For Fraud Detection	32
3.1 5 Adaptive And Rule Learning Algorithm Generated For Deploying The Application With Acl (Audit Command Language).	35
3.2 METHODS FOR FRAUD DETECTION ANALYSIS (MFDA).....	37
3.3. DATA COLLECTION	40
Table 3.10: Payment collection for Ashanti East (source: Ashanti East Database)	46
3.5 MODELING OBJECTIVES	47
3.6 DATA ANALYSIS	47
CHAPTER 4.....	50
RESULTS AND DISCUSSION.....	50
4.1 INTRODUCTION.....	50
4.1 DESCRIPTIVE ANALYZED DATA	50
4.2 SUMMARY OF FINDINGS	54
CHAPTER 5.....	57
CONCLUSIONS AND RECOMMANDATIONS	57
5.1 CONCLUSION	57

5.2 RECOMMENDATIONS	58
REFERENCES	59
APPENDICES.....	62
APPENDIX A	62
APPENDIX B	65

LIST OF FIGURES

<i>Figure 3.0 Verification of Feilds for analysis</i>	38
<i>Figure 3.1: filtering and data transformations applied</i>	39
<i>Figure 3.2: payment transaction processing system (PTPS) Flowchart</i>	48
<i>Figure 3.3 Model Diagram for fraud detection system (MDFDS)</i>	49
<i>Figure 4.1 Fraud detection process on Accra West</i>	50
<i>Figure 4.2 Gap checks on receipt number in data set</i>	51
<i>Figure 4.3. Sequence check on receipts numbers and Trans id</i>	52
<i>Figure 4.4 Receipts Duplication and Transaction ID check</i>	53
<i>Figure 4.5 Receipts Duplication and Transaction ID check in ASH EAST</i>	54

LIST OF TABLES

<i>Table 3.0: Payments history of a customer (source : ECG Payment Platform)</i>	30
<i>Table 3.1: Suspected customers account information is shown (source: ECG Payment Platform)</i>	31
<i>Table 3.2. PARAMETER DIFINITION MODEL (PDM)</i>	33
<i>Table 3.3 Stages for Model</i>	34

<u>Table 3.4 Payment collection for Accra East (source: Accra east database).....</u>	<u>40</u>
<u>Table3.5: Payment collection for Accra West (Source: Accra west database)</u>	<u>41</u>
<u>Table 3.63: Payment collection for Tema (Source: Tema database).....</u>	<u>42</u>
<u>Table 3.7: Payment collection for Volta (source: Volta database).....</u>	<u>43</u>
<u>Table 3.8: Payment collection for Eastern (source: Eastern database).....</u>	<u>44</u>
<u>Table 3.9: Payment collection for Ashanti West (source: Table 3.10: Payment collection</u> <u>for Ashanti East (source: Ashanti East Database)</u>	<u>46</u>
<u>Ashanti West database).....</u>	<u>45</u>
<u>Table 4.1 SUMMARY OF RESULTS FROM ANALYSIS</u>	<u>56</u>



CHAPTER 1

INTRODUCTION

Data mining refers to the process of extracting or mining knowledge from large an amount of data or the application of specific algorithms for extracting patterns from data. It has the ability to learn from past success and failures and to predict what will happen in the future. Data mining is therefore useful in any field where there are large quantities of data from which to extract meaningful patterns and rules. Data mining has been used by financial organizations, statisticians, data analysts, information systems managers and in the telecommunication industry.

Data mining in financial transaction is an important application because transactions are routinely generated and stores a tremendous amount of high quality data. The quantity of data is always so large that manual analysis of the data is practically impossible. The advent of data mining technology promised solutions to these problems and that is the main reason why the financial organization was an earlier adopter of data mining technology for fraud detection in their datasets.

A financial organization e.g. The Electricity Company of Ghana (ECG) is a service-oriented business, hence data mining can be viewed as an extension of the use of expert systems, which was majorly designed to address the complexity associated with maintaining a huge dataset and the need to maximize network reliability while minimizing labour cost. Also, within huge amount data usually lies a hidden knowledge of strategic importance which the natural ability cannot analyze unless with powerful tools. Hence, a deep understanding of the knowledge hidden in financial organization data is vital to the industry's competitive position and

organizational decision-making. This research discusses how data mining can be used to discover and extract useful patterns from large databases to find observable patterns.

There is a growing increase in the quantity of data in the world today, especially in the financial organization. These data include customer detail data, payment data and customer behaviors. The need to handle such large volumes of data has paved way to the development of computational techniques of extracting knowledge from large amount of data.

Financial industries, customer detail data is useful for fraud detection applications. All financial industries maintain data about the payment and crediting customers that traverse their network in the form of customer detailed records. These detailed records are kept on-line for many months; hence, billions of detailed records were usually available for data mining.

Fraud in financial organization of services delivery is a major problem as it impacts from 1% to 3% of the revenues. This is in most cases a customer specific behaviour that companies need to detect in order to minimize it. Detecting specific types of behaviour as soon as it happens is critical, and to that purpose companies deploy sophisticated detection systems. The biggest challenge to fraud detection systems is to accurately predict in near real time that a customer is a fraudster or that a service is being used in a fraudulent way. As this may happen to any customer at any, time, it is mandatory to monitor the entire customer base – sometimes several million customers making several transactions per day.

The term fraud here refers to the abuse of the profit in an organizational system without necessarily leading to direct legal consequences. In a competitive environment, fraud can become a business critical problem if it is very prevalent and especially where prevention procedures do not become fail-safe. Fraud detection, being part of the overall fraud control,

automates and helps reduce the manual parts of a screening/checking process. This area has become one of the most established industry/government data mining applications.

It is impossible to be certain about the legitimacy of and intention behind an application or transaction. Given the reality, the best cost effective option is to tease out possible evidences of fraud from the available data using mathematical algorithms such as artificial immune systems, artificial intelligence, auditing, database, distributed and parallel computing, econometrics, expert systems, fuzzy logic, genetic algorithms, machine learning, neural networks, pattern recognition, statistics, visualization and others. There are a lot of

Specialized fraud detection solutions and software, which protect businesses such as credit card, e-commerce, insurance, retail, telecommunications industries. There are often two main criticisms of data mining-based fraud detection research: the dearth of publicly available real data to perform experiments on; and the lack of published well-researched methods and techniques. Data mining for telecommunications companies involve the use of simple, traditional and advanced mathematical techniques used to analyze large populations of data and deliver insights, forecasts, explanations and predictions of how systems, customers, network and marketplace are likely to react to different situations. The hypercompetitive nature of the telecom industry has created a need to understand customers, keep them, and model effective ways to market new products. This creates a great demand for innovation like data mining technique to help understand the new business trends involved, catch fraudulent activities, identify customer payments patterns history , make better use of resources, and improve the quality of services.

1.2 Problem Statement

With the current huge customer population of ECG and demand for quality service, it has become necessary for ECG to move from the manual technology of service delivery to automated electronic service. This could be done by introducing the following through the pooling of data set:

- a. Automatic Meter reading system (AMR)
- b. Online service supply application
- c. Personal Computer (PC)-based Payment System
- d. Online Checking and payment of all bills through the customer data set.

There is also a growing fraud phenomenon common at most, customer service centres of ECG for either payment or purchase of credit; this has to do with long queues of customers waiting to receive one service or another.

Research has revealed that, commercial losses form about 15 % as results of various fraud cases through the network system leading to customer's inability to purchase credit or pay their bills promptly leading to illegal electricity connections in their homes. This research work will seek to detect fraud in various transactions in the network systems involving payments by customers and crediting of customers which are major problems contributing to about 6% to 10% loss of revenue to the company. This is in most cases a customer specific behaviour that companies need to detect in order to minimize it. Detecting specific types of behaviour as soon as it happens is critical, and to that purpose companies deploy sophisticated detection systems. The biggest challenge to fraud detection systems is accurately predict in near real time that a customer is a fraudster or that a service is being used in a fraudulent way. As this may happen to any customer

at any, time, it is mandatory to monitor the entire customer base, which sometimes comprises several million customers making various transactions per day.

1.3 Aims and Objectives of the Research

The aim of the research is to detect fraud processes in the payments and crediting of customer in account history in the database using fraud detection techniques, algorithms and possible model to analyse and prevent duplication of update of customer crediting in the database.

The specific objectives are;

1. To design an algorithm and model for fraud detection analysis in crediting customer payments.
2. To develop an expert system for checking of double crediting of customers.
3. To design an accurate fraud detection technique for payment systems
4. To optimize fraud detection process in a financial organization.

1.4 Research Questions

It is hoped that through this research reliable answers could be found to the following questions:

1. What is the various payment platforms used in the company?
2. What are the details of customer payments?
3. How many times can a customer do payment in a day?

How does one differentiate one payment from the other by the same customer?

4. Is there something like duplication of payment by a customer? If yes? How does one check such duplication?

5. Has a customer ever been credited double before? If yes? How was it detected?
6. How does one check for any fraud of payment in payment data?

1.5 Justification of Objectives

To advance as researchers and humans in general, it behoves on us all to make investigations that advance the development of humans and society in general. The study is intended to add to the existing knowledge of fraud detection techniques especially in financial institutions and organizations that are handicapped in this area. Results from the study will particularly provide workable information for managers of ECG and would allow them a better opportunity to reduce fraud in the company, boost revenue and help gain a customer trust for service delivery and discover patterns of customer payment history in data for the purpose of predictions and description.

1.6 Methodology

Data mining techniques have different features that make them suitable for analyzing large quantities of data, and these techniques are the result of a long process of research and product development. Data mining tools have the capability of analyzing massive databases and to deliver answers to questions at a very fast speed. OLAP (online analytical processing) is one of the analytical tools that focuses on providing multidimensional data analysis which are based on verification where the system is limited to verifying user's hypotheses.

They are mainly used for simplifying and supporting interactive data analysis. They also check the statistical significant of the predicted patterns and give relevant reports on them. Although the boundaries between prediction and description are not sharp, the distinction is very vital in

understanding the overall discovery goal, which is achieved through the following data mining techniques.

Selection will be made from multiple databases in all ECG operational regions for more than two (two) years about customer's monthly payment history. These data may include transaction databases, personnel databases, and accounting databases. Care will be taken though, as the data used to train data mining models must match the data for which models will be deployed in an operational setting. Thus, the tasks that need to be performed in data mining fraud detection will have to be repeated when data mining models are deployed. Finally, data mining tools (Techniques) would be used to perform data transformations involving aggregation, arithmetic operations, and other aspects of data preparation. Data mining strategies fall into two broad categories: supervised learning and unsupervised learning. Supervised learning methods are deployed when there exist a target variable with known values about which predictions will be made using the values of other variables as input.

Unsupervised learning methods tend to be deployed on data for which there exists no target variable with known values, but for which input variables does exist. Although unsupervised learning methods are used most often in cases where a target variable does not exist, the existence of a target variable does not preclude the use of unsupervised learn.

An adaptive algorithm is an algorithm that changes its behavior based on information available at the time it is run. This might be information about computational resources available, or the history of data recently received and whose behavior changes upon the presortedness of its input. An example of an adaptive algorithm in radar systems is the constant false alarm rate (CFAR) detector. In machine learning and optimization, many algorithms are adaptive or have adaptive variants, which usually mean that the algorithm parameters are automatically adjusted

according to statistics about the optimization thus far (e.g. the rate of convergence). In data compression, adaptive coding such as Prediction by partial matching can take a stream of data as input, and adapt their compression technique based on the symbols that they have already encountered.

KNUST



CHAPTER 2

LITERATURE REVIEW

A number of researches have been carried out on the Fraud Detection Techniques processes in financial organizations and telecommunications that contain data mining techniques. This chapter discusses a review of literature on these previous researches.

Zaiane et al (2010) present an overview of fraud detection in security transactions as well as a comprehensive literature review of data mining methods are used to address the issue. The best practices are based on data mining methods for detecting known fraudulent patterns and discovering new predatory strategies are identified. Furthermore, the challenges faced in the development and implementation of data mining systems for detecting manipulation in security transactions are highlighted and we provide recommendations for future research provided accordingly.

Data mining methods support fraud detection in security transactions and experimental results in literature are encouraging. However, there are many challenges in designing and developing data mining methods for detecting transaction manipulation in security transactions including heterogeneous datasets, privacy, performance and legal consequences.

The significant growth of the capital transaction due to increasing interests in investing in security transaction requires regulatory organizations to expand their efforts to ensure a fair and orderly transaction for the participants. Some of the challenges in designing and developing data mining methods include massive datasets, different sources and forms of data and using appropriate performance measures to evaluate the method. Kluwer et al (2008), talk about using adaptive method for fraud detection by checking for suspicious changes in user behaviour. It describes the automatic design of user profiling method for the purpose of fraud detection, using

a series of data mining techniques. Specifically, a rule – learning program is developed to uncover indicators of fraudulent behaviour from a large database of customer transactions; indicators are then used to create a set of monitors, which profile legitimate customer behavior and anomalies. These monitors are then used as features in a system that learns to combine evidence to generate high confidence alarm for fraud detection. The system has been applied to the problem of detecting cellular cloning fraud base on a database of transactions. The experiment indicates that this automatic approach performs better than hand – crafted methods for detecting frauds. They recommend that this approach can be adapted to the changing conditions of fraud detection environments. The learning Rule involves searching through the transaction data for indicators of fraud using the conjunctive rule discovered by a standard rule-learning program, thus, being able to use the rule to define normal transactions from abnormal transactions through defined perimeters of algorithm for the transaction data set.

Fraud behaviour changes frequently as bandits adapt to detection techniques and fraud detection system should be adaptive as well Provost and Aronis et al. (1996 - 7). However, in order to build usage monitors one must know which aspects of customers' behaviour to profile. Again, their frame work is not specific to cloning fraud, but maybe applied to fraud problems in any domain. Prime candidates are tall fraud, computer intrusions and credits - cards fraud. Finally, it can act to prevent fraud; bandits in favour of easier prey will avoid a system that quickly adapts to new patterns.

Gayle et al (2010) conducted a comprehensive survey on Data Mining Based Fraud Detection Research, Categories compared and summarized from almost all published technical and review articles in automated fraud detection. They define the professional fraudster, formalize the main types and sub – types of known frauds and present the nature of data evidence collected from

affected industries. Compared to all related reviews on fraud detection, the survey covers much more technical articles, which propose alternative data and solution from related domains. They report about Elkan et al, (2001) that say data mining is about finding insights, which are statistically reliable, unknown previously and attainable from data; this data must be available, relevant, adequate and clean. Data mining problem must be well- defined; it cannot be solved by query and reporting tools, and guided by data mining process model Havrac et al, (2004). They describe fraud detection, being part of the overall fraud control, and automate which help reduce the manual parts of a screening or checking process in data set for data mining. It is impossible to be certain about the legitimacy of and intention behind an application or transaction. Given the reliability, the best cost effective option is to tease out possible evidence of fraud from the available data using mathematical algorithm. There are plenty of specialized fraud detection solutions and software which protect business, such as credit card, e-commerce, insurance, retail, telecommunication industries which are driven by artificial immune systems, artificial intelligence, expert systems, fuzzy logic genetic algorithms, machine learning, neural network, pattern recognition, statistics, visualization and others.

There are four subgroups of insurance fraud detection: home insurance Bentley et al, (2000); Von Altrock et al, (1997), crop insurance Little *et al*, (2002), automobile insurance (Phua *et al*, (2004); Viaene *et al*, (2004); Brockett *et al*, (2002); Stefano and Gisella, (2001); Belhadji *et al*, (2000); Artis *et al*, (1999), and medical insurance Yamanishi *et al*, (2004); Major and Riedinger, (2002); Williams, (1999); He *et al*, (1999); Cox, (1995). Credit fraud detection refers to screening credit applications.

Wheeler et al (2000) logged credit card transactions Fan, (2004); Chen *et al*, (2004); Chiu and Tsai, (2004); Foster and Stine, (2004); Kim and Kim,(2002); Maes *et al*, (2002); Syeda *et al*,

(2002); Bolton and Hand, (2001); Bentley *et al*, (2000); Brause *et al*, (1999); Chan *et al*, (1999); Aleskerov *et al*, (1997); Dorronsoro *et al*, (1997); Kokkinaki, (1997); Ghosh and Reilly, (1994). Similar to credit fraud detection, telecommunications subscription data (Cortes *et al*, (2003); Cahill *et al*, (2002); Moreau and Vandewalle, (1997) Rosset *et al*, (1999), and/or wire-line and wire-less phone calls Kim *et al*, (2003); Burge and Shawe-Taylor, (2001); Fawcett and Provost, (1997); Hollmen and Tresp, (1998); Moreau *et al*, (1999); Murad and Pinkas, (1999); Taniguchi *et al*, (1998); Cox, (1997); Ezawa and Norton, (1996) are monitored. Credit transactional fraud detection has received the most attention from researchers although it has been loosely used here to include bankruptcy prediction Foster and Stine, et al (2004) and bad debts prediction (Ezawa and Norton, 1996). Employee/retail Kim *et al*, (2003), national crop insurance Little *et al*, (2002), and credit application Wheeler and Aitken *et al*, (2000) each has only one academic publication.

The main purpose of these detection systems is to identify general trends of suspicious/fraudulent applications and transactions. In the case of application fraud, these fraudsters apply for insurance entitlements using falsified information, and apply for credit and telecommunications products/services using non-existent identity information or someone else's identity information. In the case of transactional fraud, these fraudsters take over or add to the usage of an existing legitimate credit or telecommunications account.

There are other fraud detection domains. E-businesses and ecommerce on the Internet present a challenging data mining task because it blurs the boundaries between fraud detection systems and network intrusion detection systems.

Related literature focus on video-on-demand websites Barse *et al*, (2003) and IP-based telecommunication services McGibney and Hearne, *et al*, (2003). Online sellers Bhargava *et al*, (2003) and online buyers (Sherman, 2002) can be monitored by automated systems.

Fraud detection in government organizations such as tax Bonchi *et al*, (1999) and customs Shao *et al*, (2002) have also been reported.

Most fraud departments place monetary value on predictions to maximize cost savings/profit and according to their policies. They can either define explicit cost Phua *et al*, (2004); Chan *et al*, (1999); Fawcett and Provost, (1997) or benefit models Fan *et al*, (2004); Wang *et al*, (2003).

Cahill *et al* (2002) suggest giving a score for an instance (phone call) by determining the similarity of it to known fraud examples (fraud styles) divided by the dissimilarity of it to known legal examples (legitimate transactional account). Most of the fraud detection studies using supervised algorithms since 2001 have abandoned measurements such as true positive rate (correctly detected fraud divided by actual fraud) and accuracy at a chosen threshold (number of instances predicted correctly, divided by the total number of instances). In fraud detection, misclassification costs (false positive and false negative error costs) are unequal, uncertain, can differ from example to example, and can change over time. In fraud detection, a false negative error is usually more costly than a false positive error. Regrettably, some recent studies on credit card transactional fraud Chen *et al*, (2004) and telecommunications superimposed fraud

Kim *et al*, (2003) still aim to only accuracy. Some use Receiver Operating Characteristic (ROC) analysis (true positive rate versus false positive rate). Apart from (Viaene *et al* 2004), no other fraud detection study on supervised algorithms has sought to Area under the Receiver Operating Curve (AUC) and minimize cross entropy (CXE). AUC measures how many times the instances

have to be swapped with their neighbours when sorting data by predicted scores; and CXE measures how close predicted scores are to target scores.

In addition, Viaene *et al* (2004) and Stine *et al* (2004) seek to minimize Brier score (mean squared error of predictions). Niculescu-Mizil *et al*, (2004) argue that the most effective way to assess supervised algorithms is to use one metric from threshold, ordering, and probability metrics; and they justify using the average of mean squared error, accuracy, and AUC. Provost *et al* (1999) recommend Activity Monitoring Operating Characteristic (AMOC) (average score versus false alarm rate) suited for timely credit transactional and telecommunications superimposition fraud detection for semi-supervised approaches such as anomaly detection.

Lee *et al*, (2001) propose entropy, conditional entropy, relative conditional entropy, information gain, and information cost. For unsupervised algorithms, Yamanishi *et al* (2004) used the Hellinger and logarithmic scores to find statistical outliers for insurance. Shawe-Taylor *et al* 2001 employed Hellinger score to determine the difference between short-term and long-term profiles for the transactional account (Hand *et al* 2001 recommends the *t*-statistic as a score to compute the standardized distance of the target account with centred ; and also to detect large spending changes within accounts Reilly *et al* , (1994). Other important considerations include how fast the frauds can be detected (detection time/time to alarm), how many styles/types of fraud detected, whether the detection was done in online/real time (event-driven) or batch mode (time-driven)

There are problem-domain specific criteria in insurance fraud detection. To evaluate automated insurance fraud detection, some domain expert comparisons and involvement have been described. Von Altrock *et al*, (1995) claimed that their algorithm performed marginally better than the experienced auditors.

Brockett *et al* (2002) and Gisella *et al* (2001) summed up their performance as being consistent with the human experts and their regression scores. Belhadji *et al* (2000) stated that both automated and manual methods are complementary. Williams *et al* 1999 support the role of the fraud specialist to explore and evolve rules. NetMap *et al*, (2004) report visual analysis of insurance claims by which the user helped discover the fraudster.

Sherman *et al*, (2002) used predictive supervised algorithms to examine all previous labeled transactions to mathematically determine how a standard fraudulent transaction looks like by assigning a risk score. Neural networks are popular and support vector machines (SVMs) have been applied. Reilly *et al*, (1994) used a three-layer, feed-forward Radial Basis Function (RBF) neural network with only two training passes needed to produce a fraud score in every two hours for new credit card transactions. Barse *et al* (2003) used a multi-layer neural network with exponential trace memory to handle temporal dependencies in synthetic Video-on-Demand log data.

Syeda *et al* (2002) propose fuzzy neural networks on parallel machines to speed up rule production for customer-specific credit card fraud detection. Kim *et al* (2003) proposes SVM ensembles with either bagging and boosting with aggregation methods for telecommunications subscription fraud.

The neural network and Bayesian network comparison study Maes *et al*, (2002) uses the STAGE algorithm for Bayesian networks and back propagation algorithm for neural networks in financial transaction fraud detection. Comparative results show that Bayesian networks were more accurate and much faster to train, but Bayesian networks are slower when applied to new instances.

Norton et al (1996) developed Bayesian network models in four stages with two parameters. They argue that regression, nearest neighbour, and neural networks are too slow and decision attributes Label Fraud Legal Evaluation instances Training and testing examples trees have difficulties with certain discrete variables. The model with most variables and with some dependencies performed best for their transaction uncollectible debt data.

Viaene *et al* (2004) apply the weight of evidence formulation of Ada Boosted naive Bayes (boosted fully independent Bayesian network) scoring. This allows the computing of the relative importance (weight) for individual components of suspicion and displaying the aggregation of evidence pro and contra fraud as a balance of evidence which is governed by a simple additively principle. Compared to unboosted and boosted naive Bayes, the framework showed slightly better accuracy and AUC but clearly improved on the cross entropy and Brier scores. It is also readily accessible and naturally interpretable decision support and allows for flexible human expert interaction and tuning on a financial insurance dataset.

Fan et al, (2004) used decision trees, rule induction, case-based reasoning, and systematic data selection to mine concept-drifting, possibly insufficient, data streams. The paper proposed a framework to select the optimal model from four different models (based on old data only, new data only, new data with selected old data, and old and new data chunks). The selected old data is the example which both optimal models at the consecutive time steps predict correctly. The cross validated decision tree ensemble is consistently better than all other decision tree classifiers and weighted averaging ensembles under all concept-drifting data sizes, especially when the new data size of the transactions are small. With the same transaction data Fan et al (2004), Wang *et al* (2003) demonstrate a pruned classifier C4.5 ensemble which is derived by weighting each base classifier according to its expected benefits and then averaging their outputs. The

authors show that the ensemble will most likely perform better than a single classifier, which uses exponential weighted average to emphasise more influence on recent data.

Rosset *et al* (1999) present a two-stage rules-based fraud detection system which first involves generating rules using a modified C4.5 algorithm. Next, it involves sorting rules based on accuracy of customer level rules, and selecting rules based on coverage of fraud of customer rules and difference between behavioural level rules. It was applied to a telecommunications subscription fraud. Bonchi *et al* (1999) used boosted C5.0 algorithm on tax declarations of companies. Shao *et al* (2002) applied a variant of C4.5 for customs fraud detection. Case-based reasoning (CBR) was used by Wheeler *et al*, (2000) to analyse the hardest cases, which have been misclassified by existing methods and techniques. Retrieval was performed by threshold nearest neighbour matching. Diagnosis utilised multiple selection criteria (probabilistic curve, best match, negative selection, density selection, and default) and resolution strategies (sequential resolution-default, best guess, and combined confidence) which analysed the retrieved cases. The authors claimed that CBR had 20% higher true positive and true negative rates than common algorithms on credit applications. Statistical modelling such as regression has been extensively utilised.

Foster *et al*, (2004) use least squares regression and stepwise selection of predictors to show that standard statistical methods are competitive. Their version of fully automatic stepwise regression has three useful modifications; this survey has explored almost all published fraud detection studies. It defines the adversary, the types and subtypes of fraud, the technical nature of data, performance metrics, and the methods and techniques. After identifying the limitations in methods and techniques of fraud detection, this paper shows that this field can benefit from other related fields. Specifically, unsupervised approaches from counter-terrorism work, actual

monitoring systems and text mining from law enforcement, and semi supervised and game-theoretic approaches from intrusion and spam detection communities can contribute to future fraud detection research.

Provost et al (1999) show that there are no guarantees when they successfully applied their fraud detection method to news story monitoring but unsuccessfully to intrusion detection. Future work will be in the form of financial application fraud detection. Fraud is a costly problem for organizations.

The Association of Certified Fraud Examiners' (ACFE) (2008) survey results reported that U.S. organizations lost an estimated 7% of their annual revenues to fraud (ACFE 2008). This percentage increased from the estimated 5% for 2006 and 6% for 2004 (ACFE 2006). With layoffs and cuts in travel budgets for internal auditors, there is concern that as economic stresses increase due to the poor economy there will be more instances of fraud and corruption (Sullivan et al, (2009). As organizations work to reduce the incidence of fraud, their anti-fraud programs continue to rely heavily on the internal audit activity. Over time as internal auditors review systems in the organization, they develop an overall knowledge of the organization's processes, risks, control systems and personnel (IIA 2009c). These factors contribute to their effectiveness at detecting fraud. The ACFE's 2008 survey provides empirical evidence to this effect as the survey found that over 19% of the respondents' fraud cases were initially detected by internal audits versus about 9% that were discovered by external audits. The survey respondents noted that their organizations' internal audit departments played the most important role in uncovering or limiting asset misappropriations and corruption schemes. The internal auditor's role was greater than management review of internal controls, surprise audits, fraud hotlines, rewards for

whistleblowers, mandatory job rotation and vacations, and audits of internal controls for financial reporting. Internal audit department's role in detecting or limiting financial statement Priscilla Burnaby and Martha et al, (2009) however fraud schemes are was ranked second behind rewards for whistleblowers (ACFE 2008). The survey did not address IT as a separate fraud area. This study first reviews internal auditors' responsibilities for detecting fraud in their Organizations. Then the information gathered about the demographics of the respondents and their organizations are presented. The next section examines how internal auditors ranked the Likelihood and impact of fraud in their organizations in four areas: financial statement reporting, asset misappropriation, corruption and information technology (IT) and summarizes the audit procedures that respondents indicated they use to detect fraud in each of these areas. Another section lists the types of software used by internal auditors to search for fraud. Finally, the study presents the key skills that respondents suggested are needed by internal auditors for fraud detection Baker et al, (2007).

As an important first step in developing an anti-fraud program, a comprehensive fraud risk assessment should be completed. In conjunction with input from different perspectives, including management, general counsel and the audit committee, internal auditors should work together to develop the anti-fraud program. Assessing fraud risk is a natural role for internal auditors, and one that is in keeping with The IIA's fraud-related Practice Standards (IIA 2009c) with their understanding of the business and its processes, their knowledge of the organization's performance and related pressures, and their ability to assess internal controls. Zikmund et al, (2008),

Internal auditors' skills and abilities to determine the top fraud risks in terms of likelihood and impact are vital to the company's anti-fraud program. This fraud assessment is a first step toward

designing procedures to prevent and detect fraud. In this study, internal auditors were asked to complete an assessment of the fraud risks for their organization, in terms of likelihood of occurrence and in severity of impact. When there were enough respondents in an industry, analyses were performed to evaluate the subjects' level of consensus of likelihood and risk of each issue between those industries.

Research suggests internal auditors believe that business intelligence tools are effective.

For example, Bierstaker et al. (2006) found that internal auditors and accountants tended to rate these techniques as effective in combating fraud, but the study's subjects also noted that these techniques were not used very frequently, except in the largest organizations.

Cook & Clements et al. (2009) shared their concern about the lack of use of the best tools that are available and their hope that internal auditors would develop the skills necessary to continue the fight against fraud by using the best tools available.

In The IIA's Global Technology Audit Guide, *Fraud Prevention and Detection in an Automated World et al (2011)*, the use of technology as an audit tool allows the internal auditor to go from using IT as a detective control to a continuous monitoring technique. Data analysis technology allows auditors to examine data for indications of fraud (IIA 2009b). The subjects in this study were asked specifically about their use of such business intelligence tools. These questions provided data to form a clearer picture of the use of information technology by internal auditors to detect fraud. One of the challenges for auditors is to look beyond manual internal controls and find Loopholes in information systems where fraud could occur (IIA, 2009b). For all industries, the impact for security over employees' access to the systems or data was the highest, followed by security of systems and data in terms of inappropriate external parties and physical security of hardware. The means of the likelihood of fraud occurring for security over employees' access to

the systems or data in security of systems and data in terms of inappropriate external parties and for physical security of hardware, indicating a low probability of occurrence .

The results for financial impact in the Banking and Manufacturing industries followed the same order as all industries. Both the Accounting and Technology respondents selected the security of systems and data as the issue that would have the most impact for their organizations and rated security over employees' access to the systems or data second (Accounting and Technology) and physical security of hardware third (Accounting and Technology). For likelihood of issue occurrence, only security of systems and data in terms of inappropriate external parties had a significant difference between the four industries with the Accounting industry respondents ranking the likelihood of occurrence the highest.

For the Banking industry, the mean responses for security over employees' access to the systems had the second highest likelihood. In the Manufacturing industry, the mean likelihood of concerns about the security over employees' access to the systems or data was the third highest of the three issues. The means of likelihood responses in the Technology industry were the same for security over employees' access to the systems or data and physical security of hardware and lower for security of systems and data in terms of inappropriate external parties. Three out of four respondents who indicated other IT risk areas were in the Banking industry. Their additional areas of concern were payment card industry (PCI) compliance, Internet fraud utilizing electronic banking product delivery channels, and disaster recovery and identity theft.

When asked about audit procedures used to find potential fraud IT areas, respondents Reported internal control reviews as the most effective procedures that the internal audit activity performed to detect fraud, followed by risk assessment and audit steps.

Among the procedures related to internal controls, reviews of access controls, separation of duties and physical security were reported most frequently. Penetration and vulnerability testing were the most frequently reported audit tests. The risk assessment audit procedures were dominated by assessments of IT risk areas in general and IT security including firewalls and anti-virus software. Some respondents reported that this area in their organizations was covered by external auditors and not by the internal audit activity.

Use of Business Intelligence Tools to Detect Fraud Wang & Yang et al, (2009) To determine the extent to which the respondents' organizations used business intelligence tools to find and address their identified fraud risks, they were asked to select all these types of tools used in their internal audit work to detect fraud. Of the 48 respondents, 17 stated that one or more of these tools are used. It was shown that all 17 respondents used data mining. This result is consistent with an increase in the use of data mining to detect fraud, but lamented an overall underutilization. Seven respondents also used relational reporting and six used online analytical processing. Several respondents selected the "other" option; one respondent complained that all the tools listed were deficient in that the processes only caught the low hanging fraud; another noted that they used MS Access, and a third stated that they monitored email looking for transmission of credit card numbers.

2.2 Most Effective Skills at Detecting Top Fraud Risks

Reding et al, (2007). Internal auditing requires a large range of skills. These skills include inherent personal qualities and acquired knowledge and skills. The IIA has issued an *Internal Auditor Competency Framework* (Framework) (IIA 2008) that outlines the minimum level of knowledge and skills needed to effectively operate and maintain an internal audit function. The Framework is organized into four general skill buckets: interpersonal skills, tools and techniques,

internal audit standards, theory, and methodology, and knowledge areas. The participants of this study were asked to list from most important to least important the top three skills for an internal auditor to have for fraud investigation. The four skill buckets from The IIA Competency Framework and the relevant skills within each bucket that were selected by the respondents organize the following discussion. As much as space would allow, the respondents' actual words or phrases were included. The skill areas most often selected from the Framework were in the tools and techniques bucket in the risk and control assessment techniques category. Skepticism was the most frequent response for all three levels of the most important to third most important skills followed by risk assessment and recognizing fraud opportunities.

The second highest scoring skills selected by respondents from the Framework were from internal audit standards, theory, and methodology bucket in the categories of proficiency and due care. The participants ranked as most important the ability to adapt, awareness, the CFE designation, curiosity, experience, being perceptive, suspicious and understanding. The second most important skills included awareness and knowledge. The responses in the third-most important category were common sense, intelligence, creative thinking, curiosity, detail Orientation, dogged determination, being naturally suspicious and keeping an open mind that fraud may happen and the internal auditor must be willing to look for it.

Another set of skills from the tools and techniques bucket in data collection and analysis tools and techniques were the third most selected type of skills. Respondents listed analytical skills in the category as the top skill to discover fraud at all three levels. Other skills listed included data analysis, fraud investigation skills and observation. Individual respondents reported imaginative testing, documenting evidence and obtaining relevant data as the third most important skill.

Also in the tools and techniques bucket, the fourth area of most effective skills to find fraud was problem solving tools and techniques. Skills listed by respondents were thinking outside the box, digging deeper, understanding and elevating concerns, recognizing an answer that needs further investigations or information proof, understanding patterns, evaluating the results from the gathered data and problem solving skills.

This paper provides insights gained from a survey of internal auditors into how they perform audits to fulfill their professional responsibilities and the skills that are most important to find fraud. Across all industries, respondents listed fraud in the IT area as having the greatest potential for large losses with a high likelihood of occurrence. The respondents were most concerned with data and systems security due to inappropriate access by employees and by external parties. Media attention to recent large losses at TJX and other companies due to IT issues are a very real concern for loss of reputation for organizations. The IIA has placed great emphasis on IT fraud as reflected in its recent GTAG, "Fraud Prevention and Detection in an Automated World" (IIA, 2009b). Respondents are aware of the potential impact of losses. The relatively high likelihood ratings may also indicate that auditors are not particularly confident that the organization can prevent or detect an IT fraud in time to mitigate any losses. Respondents indicated the impact of fraud across all industries was highest in cash and marketable securities, timing of revenue recognition and existence of revenue. The fraud issue with the highest impact rating in the entire survey was the timing of revenue recognition in the Technology and Manufacturing industries.

VFCPA et al, (2010) None of the corruption issues were perceived to have a great impact or likelihood, and there were no industry differences in this area. Violation of the *Foreign Corrupt*

Practices Act (VF CPA) was rated with a very low likelihood, a result that is perhaps due to the relatively low (34%) percentage of respondents from organizations of international reach.

The audit procedures listed by respondents as the most effective to find fraud ranged from review of separation of duties and tests of controls to analysis of risks in the area under audit. For financial statement asset reporting and asset misappropriation, audit procedures included reconciliations and cut-off tests. Audit procedures for corruption were cantered around whistleblower policies and observing employees. For audit steps to find IT fraud, respondents listed reviewing controls over physical security and penetration and vulnerability testing. When respondents were asked about the business intelligence tools used to detect fraud, a surprisingly few relied on such tools. Only 35% of the respondents used data mining. It would be interesting to pursue this further by trying to determine whether cost, lack of skills or other reasons lie behind the low utilization of IT tools to detect fraud.

Finally, internal auditors were asked to list the top three skills needed to be the most effective at detecting their organizations' top fraud risks. The responses were loosely mapped into the four buckets of The IIA's Framework for internal auditors' skills: knowledge, tools and techniques, internal audit standards, theory, and methodology, and interpersonal skills. The respondents tended to value "soft" skills and behaviours as opposed to factual knowledge. Skills in "tools and techniques" were listed the most often. Although some of the items within this category are related to factual knowledge, many of the items are more closely related to behaviours or generic skills such as skepticism, putting you in the fraudster's shoes and ability to look at something logically. Many responses fell into the area of "internal audit standards, theory, and methodology." These items frequently tend to be aligned with softer skills and behaviours including curiosity, independence, objectivity, and creative thinking. Finally, some of

the respondents listed skills that were mapped within the category of interpersonal skills, such as interviewing and listening.

Although this study had a small sample of internal auditors so that generalizations to all internal auditors should be limited, it is a beginning on gaining an understanding of which areas of potential frauds are incorporated into internal audits, internal auditors' perceptions of the impact and likelihood of these frauds and the types of skills needed by internal auditors to find fraud. It appears that the respondents were planning their audits with the risk of many types of fraud in mind. Further research should replicate this study with a larger population using individuals from large and small internal audit activities to determine if those with more resources and larger organizations use different audit techniques to test for fraud Nuno Homem and Joao Paulo Carvalho et al, (2009)

Fraud in telecommunications services or financial transactions is a major problem it impacts from 1% to 3% of the revenues. This is in most cases a customer specific behaviour that companies need to detect in order to minimize it. Detecting specific types of behaviour as soon as it happens is critical, and to that purpose companies deploy sophisticated detection systems. The biggest challenge to fraud detection systems is to accurately predict in near real time that a customer is a fraudster or that its service is being used in fraudulent way. As this may happen to any customer at any time it is mandatory to monitor the entire customer base – sometimes several million customers making several transactions per day.

Optimizing the detection process is therefore critical. To model and analyze the problem Finite State Automata and Markov Chains were used. To solve the optimization problem Dynamic Programming and Stochastic, Hill Climber algorithms were chosen.

This work shows how a fraud detection system can be modelled and analyzed in order to

optimize the quality of its output. The approach used can easily be extended to systems that are more complex. Clearly, considering not only the tests dependency but also the test quality, results in better quality of fraud detection.

The initial FSA model presents some interesting characteristics that can be seen in real life fraud systems; the multiple testing stages and cascaded decisions. Although the analytical complexity of such a detection system may prevent a closed expression to be extracted, it is always possible to numerically compute the values – even if the tests are not independent as was assumed in this work. The overall reward model is not complex (although it might be more complex than this). As long as the staged test model is retained, it would be easy to introduce a transfer matrix between stages. This could further increase the modelling power. From the results, it is also obvious that the fraud analysis history of a given customer is also relevant to optimize the tests to be performed. This clearly shows that the very common practice of applying the same test battery for customers regardless of their previous test history is not advisable. Future work should consider correlated test results, test execution time (as this may also be a constraint) and variations to the overall cost and reward function. The inclusion of execution time constraints is also relevant as this may affect the decision at each step. This is motivated by the fact that all customers should be checked within a detection cycle and some tests may require more time or processing power than is available.

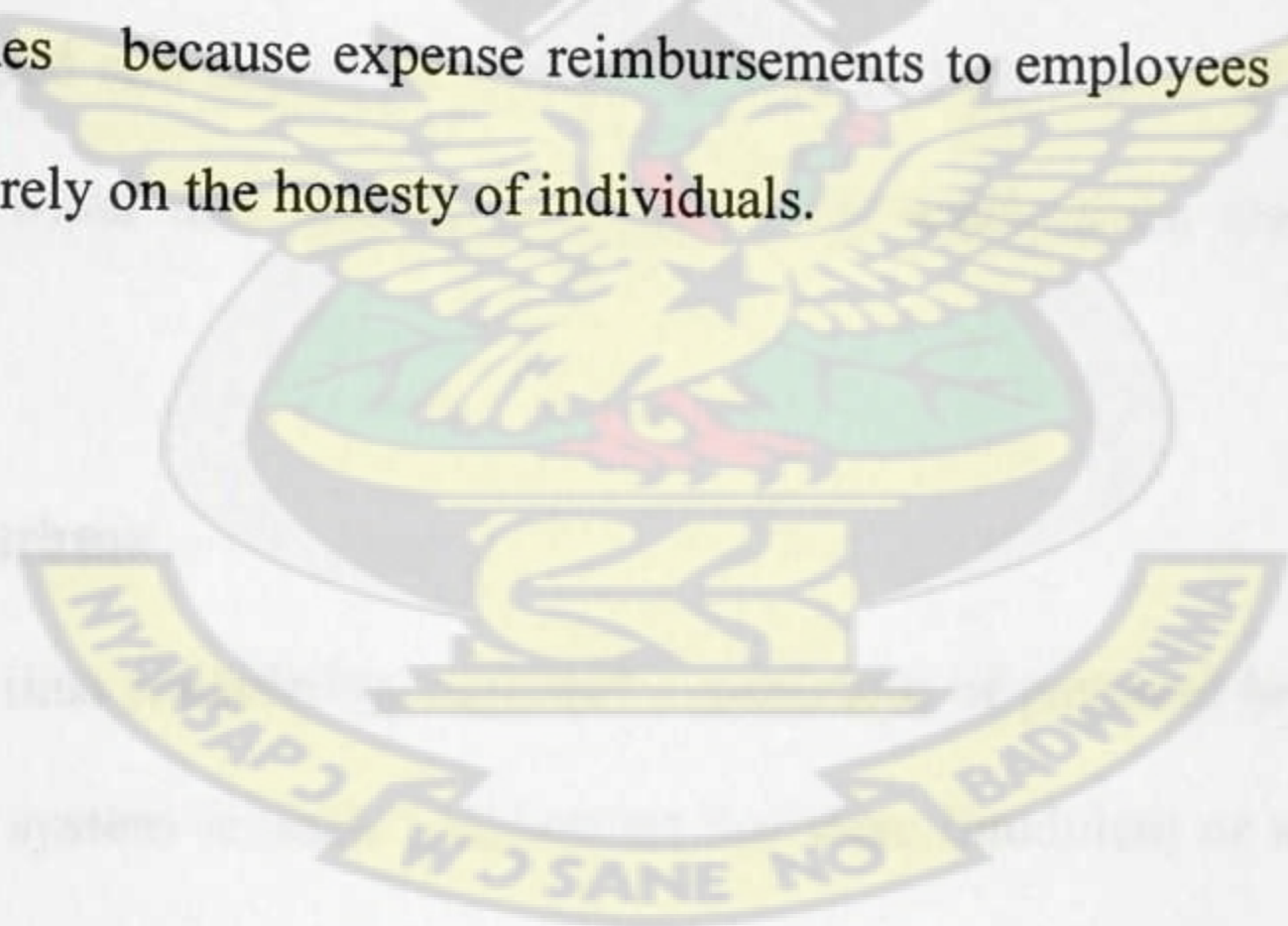
Changes to the overall fraud detection process should also handle the possibility of a customer becoming a fraudster not at the beginning but later on (maybe due to theft of the mobile phone or financial difficulties of the customer itself).

From the above literatures review it has been noted that fraud detection techniques has the probability of precision or accuracy in data mining in the various organizations being it

financial, telecommunications, credit card transactions, payments crediting systems and etc has its problem of detecting a fraudster from the beginning of transaction or at the end of transaction in a pull of data.

This is not unexpected since inappropriate revenue recognition has been a factor in many infamous financial statement frauds over the past decade. It appears that despite years of attention, prevention measures have not been sufficient to reduce the likelihood of fraud in revenue recognition.

Responses in the asset misappropriation and corruption areas showed few notable results except for accounting organization respondents that were concerned with the skimming of incoming funds. Employee expense reimbursements were rated the fraud risk with the highest likelihood in the entire reviews for all the industries with Technology industry respondents indicating the highest likelihood of the four industries. This has necessitated the research into the fraud detection techniques because expense reimbursements to employees are an area where internal auditors have to rely on the honesty of individuals.



CHAPTER 3

MATERIALS AND METHODS

This Chapter deals with the method of data collection and how the data is organized for fraud detection analysis. It also contains detailed discussion of the main techniques and a review of the basic methods used for the analysis of the data for fraud in data mining.

3.1 Description of Project Site

The research was carried out from selected multiple Databases in Electricity Company of Ghana (ECG) which includes Accra East, Accra West, Tema, Eastern, Volta , Central , Western, Ashanti East and Ashanti West. Each data set collected contained transaction databases, personnel databases, and accounting databases. The data was used for the fraud detection analysis using adaptive and rule learning algorithms technique to match the data on which application was developed and deployed. Thus, the tasks that needed to be performed in data mining were repeated when data mining algorithms and applications were deployed (refer appendix A).

3.1.1 Adaptive Algorithms

An adaptive Technique thus considering customer's behaviour of payment history was collected from the ECG database system to track transactions that were fraudulent or not. This technique was further used to predict fraudulent transactions by partial matching of the strings of data from the input to the extracted datasets. The table 3.0 shows behaviour a particular customer payment for a period of two years, which was used in technique for the algorithm.

Table 3.0: Payments history of a customer (source -: ECG Payment Platform)

Acct_No	Receipt_No	Amount	Pmt_Date	Trans_ID
9146092001	166421-4971	100.00	8/1/2012	623223
9146092001	166421-4972	50.00	10/2/2012 0:00	63223819
9146092001	166421-4973	30.00	11/3/2013 0:00	63223420
9146092001	166421-4974	150.00	12/4/2012 0:00	63223221
9146092001	166421-4975	80.00	1/5/2013 0:00	6322322
9146092001	166421-4976	30.00	2/6/2013 0:00	6322323
9146092001	166421-9016	240.00	3/6/2013 0:00	6935663
9146092001	159387-1	40.00	8/6/2012 0:00	1272279
9146092001	181.00	100.00	6/11/2012 0:00	875431
9146092001	4295.00	50.00	7/3/2012 0:00	1015006
9146092001	185299-1	100.00	8/3/2012 0:00	1261101
9146092001	3272.00	400.00	2/25/2013 0:00	3509701
9146092001	192317-11421	100.00	10/3/2013 0:00	7470865
9146092001	4416.00	50.00	3/21/2013 0:00	3925067
9146092001	166421-9002	400.00	9/5/2013 0:00	6932551

3.1.2 Supervised Learning Method (Directed Data Mining)

The variables under investigation were split into two groups: explanatory variables and dependent variables. The target of the analysis was to specify the relationship between the explanatory variables and the dependent variables as it is done in the analysis.

All the records available were fraudulent' or 'non-fraudulent'. After building a model (figure 3.3) using these collected data, new cases were classified as either fraudulent or legal. This method is able to detect frauds of a type, which has previously occurred.

To apply directed data mining techniques the values of the dependent variable must be known for a sufficiently large part of the data set.

3.1.3 Unsupervised learning

For this analysis, all the variables are treated in the same way without any a distinction between explanatory and dependent variables records. Rather the accounts of customers, amount, Trans-ID and Receipt No. for example, suspected to be fraudulent are audited.

Sample of suspected accounts of customers from ECG Payment Platform in table 3.1 with their text on deployed application

Table 3.1: Suspected customers account information is shown (source: ECG Payment Platform)

ACCOUNT NO	AMOUNT	TRANS ID	RECEIPTS	PAID DTE
901-0728-001	60.00	97271	913266	21-03-2013
901-0786-002	77.20	1710	924428	21-03-2013
901-0786-002	77.20	1710	924428	21-03-2013
901-1268-001	35.00	1710	915492	21-03-2013
901-1268-001	35.00	1710	915492	21-03-2013
901-1279-001	14.00	97271	910738	21-03-2013
901-1279-001	14.00	97271	910738	21-03-2013
901-1302-004	10.00	90263	925723	21-03-2013
901-1331-002	90.00	90263	925781	21-03-2013
901-1331-002	90.00	90263	925781	21-03-2013
901-1371-002	10.00	1710	922887	21-03-2013
901-1371-002	10.00	1710	922887	21-03-2013
901-1422-001	32.00	1710	914454	21-03-2013
901-1422-001	32.00	1710	914454	21-03-2013
901-1427-001	59.00	1710	914420	21-03-2013
901-1427-001	59.00	1710	914420	21-03-2013
901-1448-002	20.00	1710	907775	21-03-2013
901-1448-002	20.00	1710	907775	21-03-2013
901-1452-001	42.00	90263	925675	21-03-2013
901-1452-001	42.00	90263	925675	21-03-2013
901-1472-001	7.00	90263	925541	21-03-2013
901-1472-001	7.00	90263	925541	21-03-2013

Although the research dealt with large data sets and typically had abundant cases, partially missing values and other data peculiarities sometimes made it impossible to split the data into as many sub datasets as would have been necessary.

Resampling and cross-validation techniques were hence often used in combination with the data and computer intensive methods in the Data Mining.

3.1.4 Process for Algorithms Techniques of Adaptive and Rule Learning Used For Fraud Detection

In particular, seven types of tests of fraud detection algorithms Technique were performed on customer data. From the seven available tests, three were Instantaneous of adaptive behaviour and four were the rule learning, which took some time to compute and completed in the application developed. Each test showed only positive or negative result on the conditions of system deployed. The tests on collected data were always performed before all other tests in the system for fraud detection. These tests check the following:

1. Is the customer a recent customer? If the customer is not sufficiently known in the database, no history exists to access is behaviour. Unknown customer's history or recent customers may pose increased risks to be fraud in system.
2. Is the rate of payment history High? The payment behaviour of the customer (how many payments update in system per month services used, payments made, list of destinations) is checked. A large rate of payments and crediting per month may indicate non-personal payments to several persons by the customer without intention of paying for it.
3. Are payments high? The algorithm checks if the customer is making high payments in his history, especially within a particular month. This is a fast check because only the number of payments and their receipt numbers are used.

4. Payment check - Checks whether the payments limit of the customer is the appropriate by using its invoice and payment history (if available), its demographic data and the history of that customer in database.
5. Payments Pattern – Checks whether the payments is a typical fraudster payments. Fraudsters tend to have very high payments history with receipt number repeating with very high value and changes to transaction id before update.
6. Genuine Payment history in system - Checks if the type of payment is typical of systems history than accept processing, if not than check history of payments of customer before update.
7. Check list of Fraudsters? Checks whether the customer has a similar fraudulent behaviour history already known fraudster. These fraudulent activities are usually obtained by using the receipt numbers, transaction identity and sequence of payments history as shown in Table 3.2 and Table 3.3.

Table 3.2. PARAMETER DIFINITION MODEL (PDM)

ATTRIBUTES	MODEL ID	STAGES
Account Number	A1	Ad, Ar
Receipt Number	Rn	Rs, Rd, Rr
Transaction Id	To	Ts, Ts1, Tda, Tr
Paid Date	Pd	Pd1
Fraud Detect	Fd	Fd1, Fd2 , Fd3

Table 3.3 Stages for Model

SYMBOL	DESCRIPTION IN MODEL REPRESENTATION.
A1	checking for account number exiting with customer name
Ad	checking for duplication of account number in data set for process
Ar	Checking for account number with customer class, tariff class and ready.
Rs	Test for receipt number already exists in system, duplication in data set or with other customer transaction in data set
Rd	Receipt duplicate in the data set to be processed
Rr	Receipt ready for process.
To	Transaction id check process
Ts	Transaction code exists in the database
Tda	Testing for Transaction code and duplication in data set to be processed
Tr	Transaction code ready and good to be processed
Pd	Payment date check in data set
Fd1	Fraud suspected because of Receipt ready matches with account ready in data set but does not match with transactional code in data set.
Fd2	Fraud suspected because of Receipt ready matches with Transaction code in data set but does not match with account number in data set.
Fd3	Fraud suspected because of Receipt ready matches with Transaction code in data set but does not match with account number.
Pup	Process data to system for crediting
Sp	Systems stop or end for feed results.

3.1 5 ADAPTIVE AND RULE LEARNING ALGORITHM GENERATED FOR DEPLOYING THE APPLICATION WITH ACL (ADUIT COMMAND LANGUAGE).

1. START
2. LOAD THE PARAMETERS $A_1, R_n T_o,$
3. CHECK VALIDITY AND FILTRATION OF DATA SET ON $A_1, R_n T_o,$
4. CHECK IF $A_1,$ EXIST IN THE SYSTEM AS GENUINE ACCOUNT NUMBER BEFORE PROCESSING.
5. CHECK IF $A_1,$ EXIST IN DATA SET ($A_d,$) WITH RECEIPT NUMBER DUPLICATION.
6. IF YES , THEN $F_d,$ ELSE $A_r,$
7. CHECK IF RECEIPT NUMBER EXIST IN THE DATABASE(SYSTEM($R_s,$)) WITH RECEIPT IN DATA SET ($RA_d,$) THEN $F_d,$ ELSE $R_r,$
8. CHECK IF TRANSACTION ID($T_o,$) EXIST IN DATABASE($T_s,$) WITH TRANSACTION ID IN DATA SET($T_{da},$) THEN $F_d,$ ELSE $T_r,$
9. IF $A_r = R_r \neq T_r$ THEN F_{d1} GO TO 6.
10. ELSE IF $A_r = T_r \neq R_r$ THEN F_{d2} GO TO 5
11. ELSE IF $R_r = T_r \neq A_r$ THEN F_{d3} THEN GO TO 3
12. ELSE IF $A_r = R_r = T_r$ THEN PUP (PAYMENT PROCESS READY).
13. REPEAT PROCESSES UNTIL DATASET COMPLETE
14. VERIFY FIELDS ACCOUNT_No AMOUNT PAID_DATE RECIEPT_NO TRAN_ID ERRORLIMIT 10 TO SCREEN
15. GAPS ON RECIEPT_NO PRESORT TO SCREEN
16. SEQUENCE ON RECIEPT_NO TRAN_ID ERRORLIMIT 10 TO SCREEN

17. DUPLICATES ON RECIEPT_NO OTHER ACCOUNT_No AMOUNT PAID_DATE

RECIEPT_NO TRAN__ID PRESORT TO SCREEN

18. SET FILTER

19. SET FILTER

20. DUPLICATES ON RECIEPT_NO OTHER ACCOUNT_No AMOUNT PAID_DATE

RECIEPT_NO TRAN__ID PRESORT TO SCREEN

21. OPEN ASH_EAST

22. VERIFY FIELDS ACCOUNT_No AMOUNT PAID_DATE RECIEPT_NO TRAN__ID

ERRORLIMIT 10 TO SCREEN

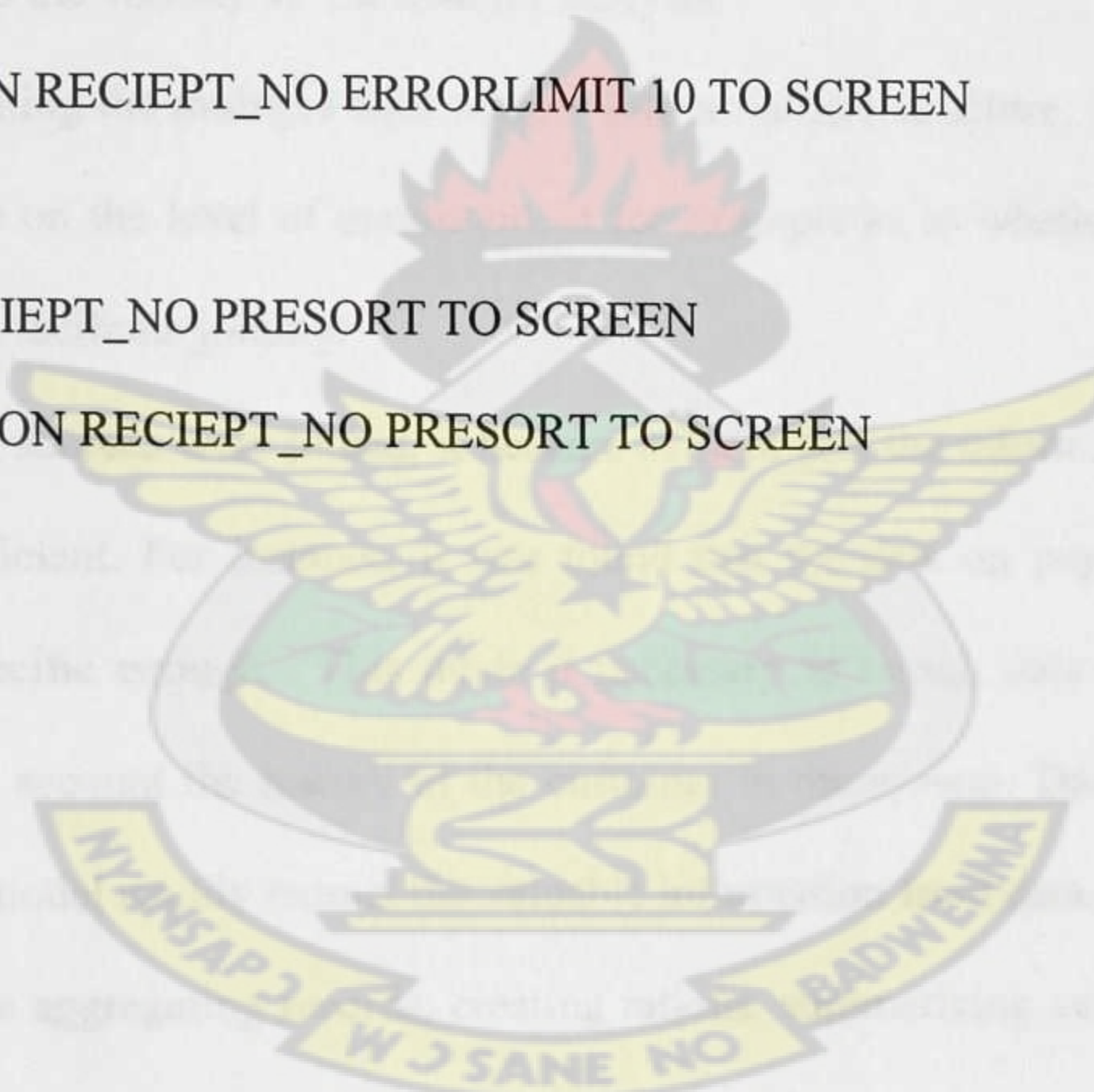
23. SEQUENCE ON RECIEPT_NO ERRORLIMIT 10 TO SCREEN

24. SET FILTER

25. GAPS ON RECIEPT_NO PRESORT TO SCREEN

26. DUPLICATES ON RECIEPT_NO PRESORT TO SCREEN

27. STOP



3.2 Methods for Fraud Detection Analysis (MFDA)

The data was annotated based on the data mining approach (i.e. both supervised and unsupervised learning methods).

A comprehensive picture of transaction of data set was collected by incorporating legacy data with transactions on current systems. Data gaps in the dataset were edited and filtered off to enrich the data. Finally, data mining tools such as filtering, data transformations involving aggregation, arithmetic operations, and other aspects of data preparation were applied to the data on the available field of the data set such account number, receipt numbers and transaction identity which produces the validity of the data for analysis.

Data step involved joining the multiple data sources into a flat file structure. This step required that decisions be made on the level of measurement for example as to whether some data was summarized in order to facilitate joining.

As data from disparate sources were joined, it became evident that the information contained in the records was insufficient. For instance, it was found that the data on payment or platform extraction was not specific enough. This made it necessary to enrich data with external (or other) data taking into account the history of the customer in the system. Data transformations were used to enable a model readily extract the valuable information from data. Examples of data transformations include aggregating records, creating rations, summarizing very granular fields, etc

Below are figures showing verification, filtering and data transformations applied in the MFDA in application.

Figure 3.0 Verification of Feillds for analysis

Project Navigator

oliverproject.ACL

- ACCRA_WEST
- ASH_EAST
- ASH_WEST
- EASTEN
- OLIVER
- oliverproject
- sup
- TEMA
- VOLTA

Welcome | TEMA | Verify

Filter:

	ACCOUNT_No	RECIPT_NO	AMOUNT PAID DATE	TRAN
1			2013	1.
2			2013	2.
3			2013	3.
4			2013	4.
5			2013	5.
6			2013	6.
7			2013	7.
8			2013	8.
9			2013	9.
10			2013	10.
11			2013	11.
12			2013	12.
13			2013	13.
14			2013	14.
15			2013	15.
16			2013	16.
17			2013	17.
18			2013	18.
19			2013	19.
20			2013	20.
21			2013	21.
22			2013	4.
23			2013	10.

Verify

Main | More | Output

Verify Fields...

Name	Title	St...	Cate...
ACCOU...	ACCOUNT_No	1	N
AMOUNT	AMOUNT	17	N
PAID_D...	PAID_DATE	24	C
RECIPT...	RECIPT_NO	11	N
TRAN_ID	TRAN_ID	31	C

OK Cancel Help

Project Navigator

oliverproject.ACL

- ACCRA_WEST
- ASH_EAST
- ASH_WEST
- EASTEN
- OLIVER
- oliverproject
- sup
- TEMA
- VOLTA

Welcome | TEMA | Verify

As of: 02/14/2013 10:58:50

Command: VERIFY FIELDS ACCOUNT_No AMOUNT PAID_DATE RECIPT_NO TRAN_ID ERRORLIMIT 10 TO SCREEN

Table: TEMA

0 data validity errors detected

The screenshot displays the Microsoft Access 'Edit: Filter' dialog box. The 'Available Fields' tab is selected, showing a list of fields from the 'sup' table: ACCOUNT_NO, AMOUNT, PAID_DTE, RECEIPTS, and TRANS_ID. The 'Functions' tab lists various functions like ABS, AGE, ALLTRIM, etc. The 'Variables' tab shows GAPDUP1 and OUTPUTFOLDER. The 'Expression' tab is empty. The background shows a table with columns ACCOUNT_NO, AMOUNT, and TRANS_ID, with rows 1 through 20.

3.3. Data Collection

Data was collected from various ECG regions with information related to the specific target group envisioned in the ECG platform for about two (two) years and above concerning customer's history of payments for the analysis.

Table 3.4 Payment collection for Accra East (source: Accra east database)

ACCOUNT NO	AMOUNT	TRANS ID	RECEIPTS	PAID DTE
11-0728-001	60.00	97271	913266	21-03-2013
11-0728-003	77.20	1710	924428	21-03-2013
11-0728-004	35.00	1710	915492	21-03-2013
11-0728-006	14.00	97271	910738	21-03-2013
11-0728-008	10.00	90263	925723	21-03-2013
11-0728-009	90.00	90263	925781	21-03-2013
11-0728-011	10.00	1710	922887	21-03-2013
11-0728-013	32.00	1710	914454	21-03-2013
11-0728-014	32.00	1710	914454	21-03-2013
11-0728-016	59.00	1710	914420	21-03-2013
11-0728-017	20.00	1710	907775	21-03-2013
11-0728-018	20.00	1710	907775	21-03-2013
11-0728-020	42.00	90263	925675	21-03-2013
11-0728-022	7.00	90263	925541	21-03-2013
11-0728-023	2.00	90263	925697	21-03-2013
11-0728-024	2.00	90263	925697	21-03-2013
11-0728-026	50.00	97271	910948	21-03-2013
11-0728-027	60.00	97271	907982	21-03-2013
11-0728-029	75.10	1710	908665	21-03-2013
11-0728-031	2.00	1710	909054	21-03-2013
11-0728-032	75.10	1710	908665	21-03-2013
11-0728-033	20.00	1710	917988	21-03-2013
11-0728-035	45.00	97271	906069	21-03-2013
11-0728-036	45.00	97271	906069	21-03-2013
11-0728-038	27.00	1710	907453	21-03-2013
11-0728-040	16.00	335	920955	21-03-2013
11-0728-041	18.00	335	912178	21-03-2013
11-0728-043	5.00	335	910836	21-03-2013
11-0728-044	5.00	335	910836	21-03-2013
11-0728-046	15.00	335	913191	21-03-2013

Table3.5: Payment collection for Accra West (Source: Accra west database)

ACCOUNT_No	RECIEPT NO	AMOUNT	PAID_DATE	TRAN_ID
2185785001	23140	179.8	5082013	200345
2185785001	24910	40	5082013	200345
2004815001	792800	100	9082013	200612
2185785001	700220	1712.5	9082013	200345
2185785001	737310	1000	14082013	200345
2009076002	780960	150	14082013	200612
2017230001	818770	200	14082013	200612
2185785001	723380	92.5	2082013	200610
2185785001	327020	2212	6082013	200345
2185785001	330420	1000	6082013	200345
2185785001	369440	1106	6082013	200345
2185785001	371170	1200	6082013	200345
2185785001	56120	1000	15082013	200345
2171197001	59810	100	15082013	159832
2236041001	16740	30	15082013	159832
2223768001	39480	31.5	15082013	200345
2100264001	49610	100	15082013	159832
2049917002	452830	30	7082013	20107
2236561001	22810	80	12082013	200612

Table 3.63: Payment collection for Tema (Source: Tema database)

ACCOUNT_No	RECIEPT_NO	AMOUNT	PAID_DATE	TRAN_ID
3143612001	620860	20	1082013	1. 189219
3140972001	626660	30	1082013	2. 186538
3136101001	627470	50	1082013	3. 186538
3018628001	628750	20	1082013	4. 186538
3020741001	630990	120	1082013	5. 186538
3016997001	639130	40	1082013	6. 186538
3106317001	643090	100	1082013	7. 186538
3117931001	644830	20	1082013	8. 186538
3143975001	645560	60	1082013	9. 186538
3018737001	609400	59	1082013	10. 200064
3083480001	610220	10	1082013	11. 186538
3125523001	611490	55.96	1082013	12. 200064
3006895001	611840	20	1082013	13. 189219
3094506001	612180	50	1082013	14. 186538
3144558001	648790	40	1082013	15. 200064
3127014001	650720	20	1082013	16. 186538
3091439001	652210	30	1082013	17. 186538
3001540001	671930	400	1082013	18. 200979
3096468001	677370	50	1082013	19. 146358
3096468001	677730	500	1082013	20. 146358

Table 3.7: Payment collection for Volta (source: Volta database)

ACCOUNT_No	RECIEPT_NO	AMOUNT	PAID_DATE	TRAN_.ID
4056034001	786980	20	2082013	1. 201237
4099389001	789120	15	2082013	2. 201237
4043886001	790320	25	2082013	3. 201237
4034574001	790870	7	2082013	4. 40183
4125335001	791040	4	2082013	5. 201237
4038609001	791980	28	2082013	6. 201237
4045692001	792300	50	2082013	7. 40183
4142177001	792560	20	2082013	8. 201237
4114256001	792640	20	2082013	9. 40183
4146211001	793410	5	2082013	10. 201237
4143119001	798850	20	2082013	11. 40183
4149017001	799150	22	2082013	12. 201237
4127470001	801730	50	2082013	13. 201237
4113037001	802660	10	2082013	14. 40183
4034080001	820650	50	2082013	15. 40183
4149342001	820730	11	2082013	16. 201237
4098793001	821800	26	2082013	17. 40183
4125354001	822100	10	2082013	18. 201237
4034076001	822590	10	2082013	19. 40183
4031049001	824230	12	2082013	20. 201237

Table 3.8: Payment collection for Eastern (source: Eastern database)

ACCOUNT_No	RECIEPT_NO	AMOUNT	PAID_DATE	TRAN_.ID
5044799001	414940	10	6082013	1. 50110
5048158001	414950	41	6082013	2. 201745
5044760002	414970	16	6082013	3. 50110
5044900001	414990	14	6082013	4. 50110
5044853001	415020	50	6082013	5. 50110
5049403001	415040	2	6082013	6. 50110
5127377001	415050	10	6082013	7. 50110
5044775001	415100	2	6082013	8. 50110
5075670001	415120	26	6082013	9. 50110
5045323001	789040	30	2082013	10. 201745
5167297001	809010	25	2082013	11. 201745
5114365001	872650	5	2082013	12. 50110
5088485001	872960	5	2082013	13. 50110
5088487001	873220	10	2082013	14. 50110
5088510001	873570	2	2082013	15. 50110
5088516001	873750	4	2082013	16. 50110
5114347001	873990	10	2082013	17. 50110
5088502001	874140	5	2082013	18. 50110

Table 3.9: Payment collection for Ashanti West (source: Ashanti West database)

ACCOUNT_No	RECIEPT_NO	AMOUNT	PAID_DATE	TRAN_ID
6104188001	609690	52	1082013	1. 200614
6104236002	609940	5	1082013	2. 200614
6105624001	610200	12	1082013	3. 200614
6208725001	610960	55	1082013	4. 200614
6244561001	611390	50	1082013	5. 200614
6245605001	612330	10	1082013	6. 200614
6269010001	612940	80	1082013	7. 60102
6104568001	613100	40	1082013	8. 200614
6105113001	613550	27.5	1082013	9. 200614
6278602001	614060	10	1082013	10. 200614
6185697001	614710	992.34	1082013	11. 200614
6330914001	615270	10	1082013	12. 60102
6274529001	615760	56	1082013	13. 60102
6138882001	616680	40	1082013	14. 200614
6269360001	617110	30	1082013	15. 60102
6135994001	617520	100	1082013	16. 200614
6185763001	617890	10.2	1082013	17. 200614
6105280001	618500	40	1082013	18. 60102
6105138001	618660	20	1082013	19. 200614

Table 3.10: Payment collection for Ashanti East (source: Ashanti East Database)

ACCOUNT_No	RECIEPT_NO	AMOUNT	PAID_DATE	TRAN_.ID
9241513001	785490	57	2082013	1. 203022
9141819001	785670	147	2082013	2. 203024
9265665001	786020	83.3	2082013	3. 203024
9118827001	786290	42	2082013	4. 203024
9259863001	786570	17	2082013	5. 203024
9295792001	786860	3	2082013	6. 203024
9141484001	786990	34.2	2082013	7. 203022
9295042001	787260	15	2082013	8. 203024
9187189001	787440	50	2082013	9. 203022
9308982001	787610	4	2082013	10. 203024
9122571002	787900	19	2082013	11. 203022
9313703001	788820	20	2082013	12. 203024
9122875001	789110	5	2082013	13. 203022
9278443001	789140	10	2082013	14. 203024
9151575001	789550	15	2082013	15. 203022
9316461001	789760	23	2082013	16. 203024
9122622001	789850	10	2082013	17. 203022
9242323001	789960	40	2082013	18. 201727
9253959001	789990	40	2082013	19. 201727
9215244001	790250	35	2082013	20. 201727

3.5 Modeling Objectives

Prediction algorithms determine models or rules to predict continuous or discrete target values of a given input data. The classification scheme involved trying to determine if transactions represented fraudulent behaviour based on some indicators such as the type of receipt number, account number, amount paid, date paid and transaction identification as in figure 3.3 with parameter definition in table 3.2 and table 3.3 for model fraud detection.

Exploration uncovered dimensionality in input data. Affinity analysis determined which events are likely to occur in conjunction with one another in a particular data set problem involving fraud, the objective was to establish a classification scheme for fraudulent transactions. Adaptive and rule learning techniques were used to test and detect the fraudulent problem. The adaptive and rule learning helped build classification rules and other mechanisms for detecting fraud. Clustering indicating the types of groupings in a given data (based on a number of inputs) were used to exhibit fraud as shown in figure 3.2.

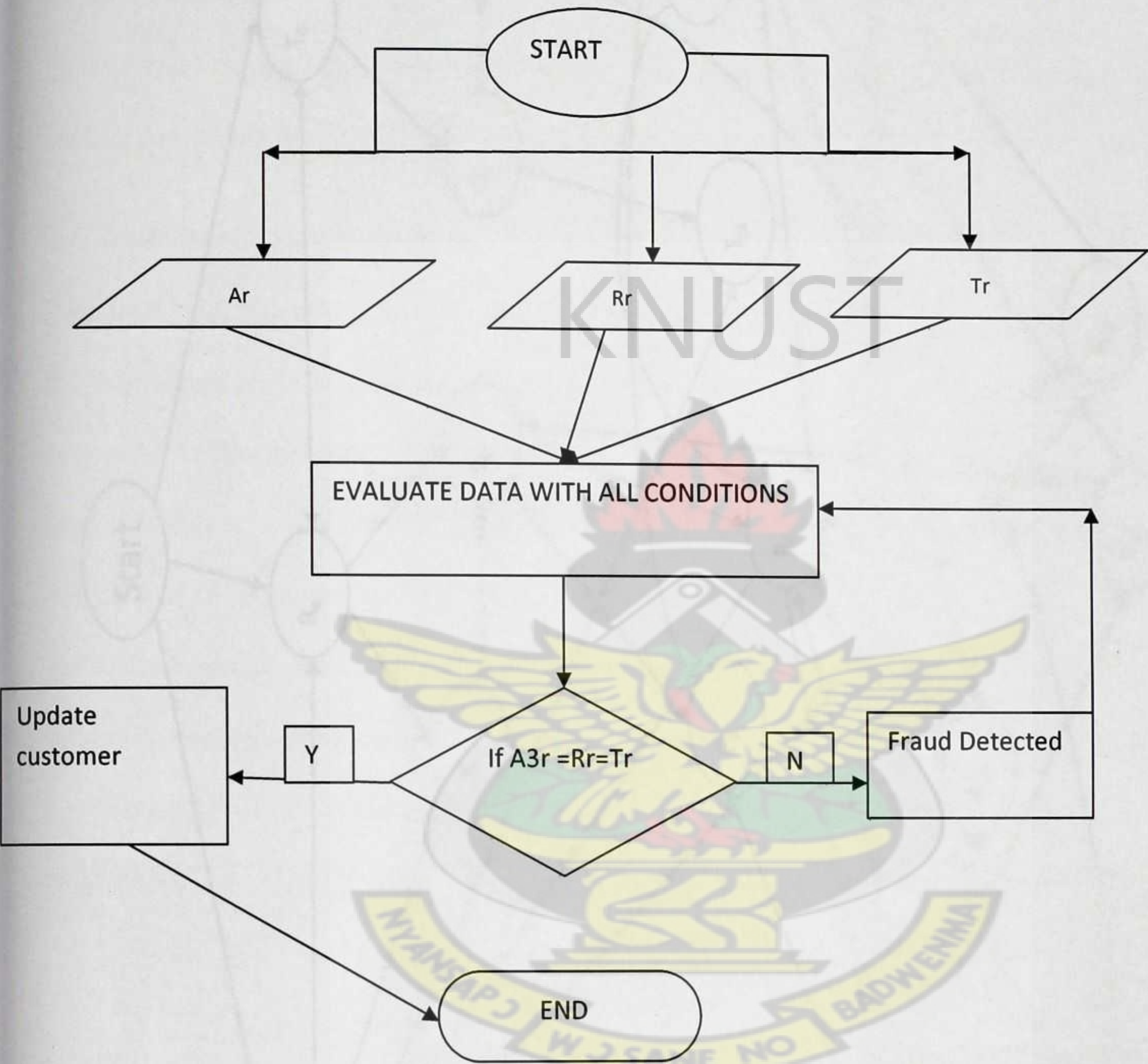
Classification modeling tries to find models (or rules) that predict the values of one or more variables in a data set (target) from the values of other variables in the data set (inputs).

After finding a good model, the data mining tool was applied the model to new data sets that could contain the variable(s) being predicted.

3.6 Data Analysis

An algorithm (appendix A) was developed to analyze all the data set extracted from the various regional data sets in detecting patterns that were similar to trends known to represent fraudulent activities and those that were not. The detection pattern was in two levels: a) detecting suspicious accounts with fraudulent behaviour, b) detecting duplication that was associated with fraudulent activities.

Figure 3.2: payment transaction processing system (PTPS) Flowchart.



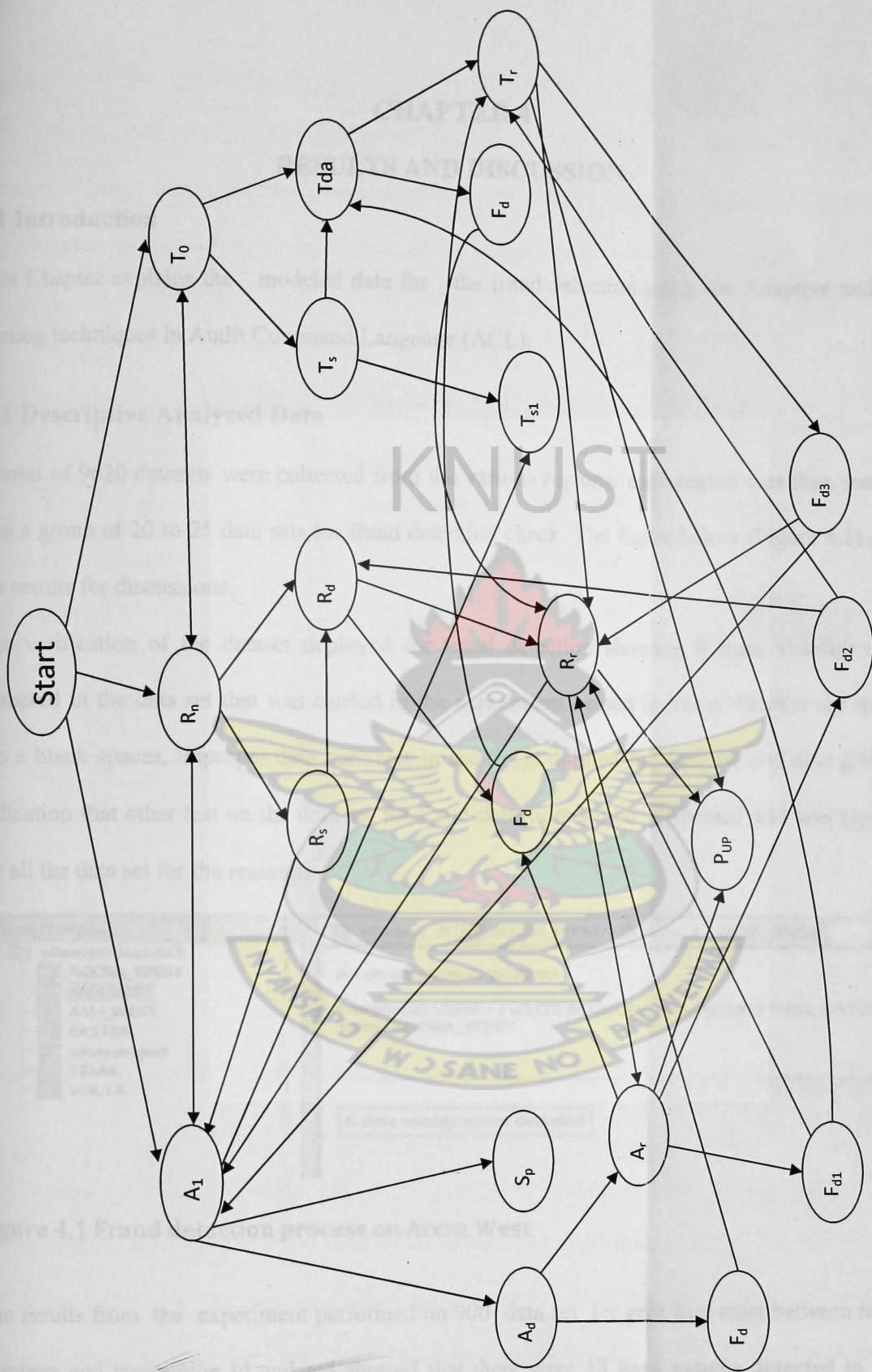


Figure 3.3 Model Diagram for fraud detection system (MDFDS)

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Introduction

This Chapter explains the modeled data for the fraud detection using the Adaptive and rule leaning techniques in Audit Command Language (ACL).

4.1 Descriptive Analyzed Data

A total of 9620 datasets were collected from the various regions; each region was than sampled into a group of 20 to 25 data sets for fraud detection check. The figure below (Figure 4.1) gives the results for discussions.

The verification of the dataset deployed for fraud detection showing **0 data validity error detected** in the data set that was carried in the experiment helped to know whether the dataset has a blank spaces, improper data variables in the fields defined for analysis and also gives an indication that other test on the data set for fraud detection can be performed and was repeated for all the data set for the research.



Figure 4.1 Fraud detection process on Accra West

The results from the experiment performed on 900 data set for gaps that exist between receipt numbers and transaction Id updated showed that there were **18 gaps ranges detected** in their

receipts numbers indicating where exactly the gaps starts and ends in each group receipts numbers. This calls for further investigation into the corresponding payment used to credit the customers' payments in the system and why this gaps. Figure 4.2 below is the snap shot of the experiment performed on detection fraudulent activities of receipts numbers for gaps.

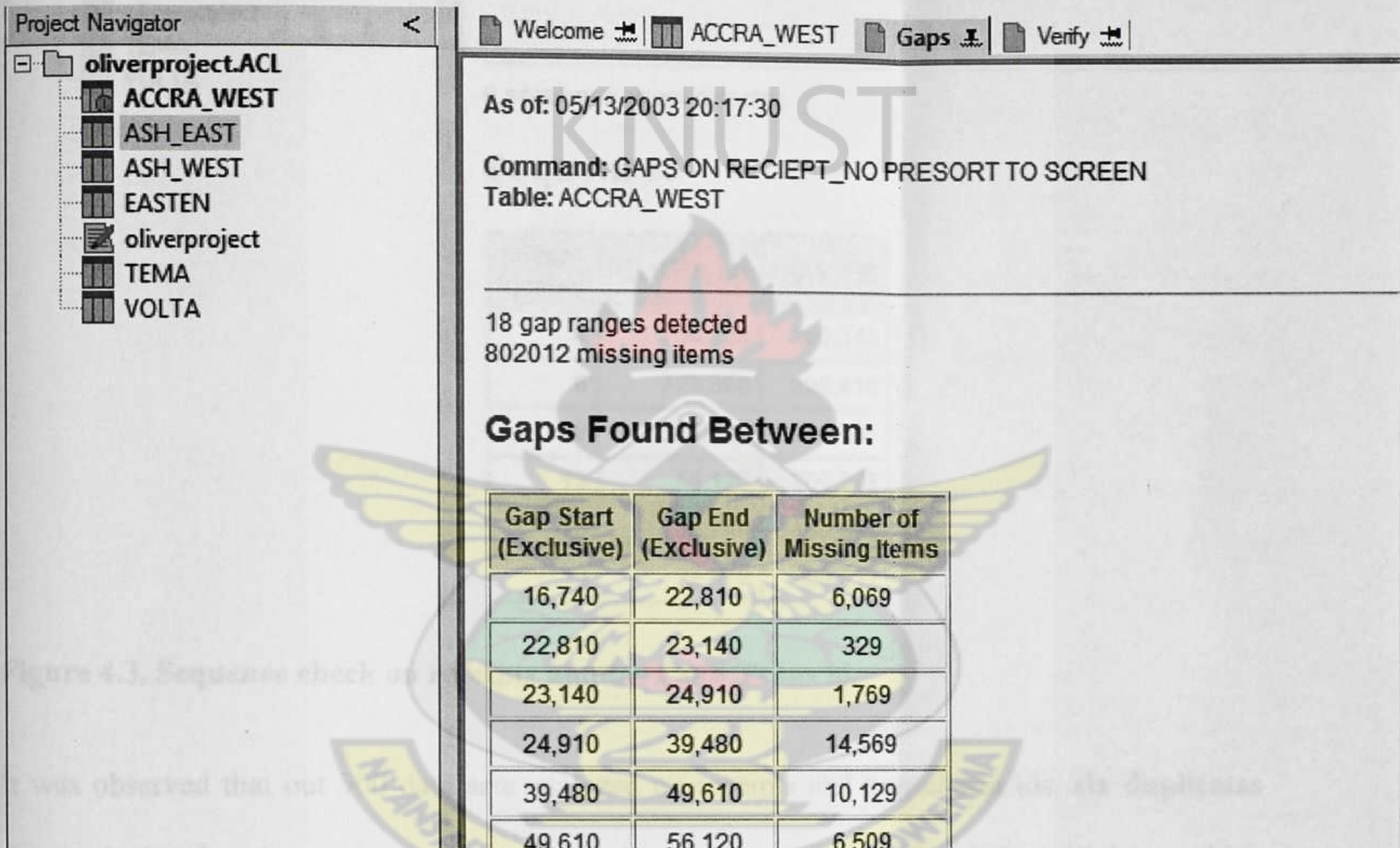


Figure 4.2 Gap checks on receipt number in data set

It was observed that out of about 900 customer payment data updated, the experiment performed on the receipts numbers and transaction Id for sequences, **6 sequence errors detected**, showing the exact records in the data set where the receipts numbers and their trans_id were not sequentially order of payment pattern or had their receipt numbers and transaction identities (ID)

NOT sequentially followed and detected as fraud which should be investigated. These were all of fraudulent activities detected that occurred in an already updated database system of the organization as shown in figure 4.3 below.

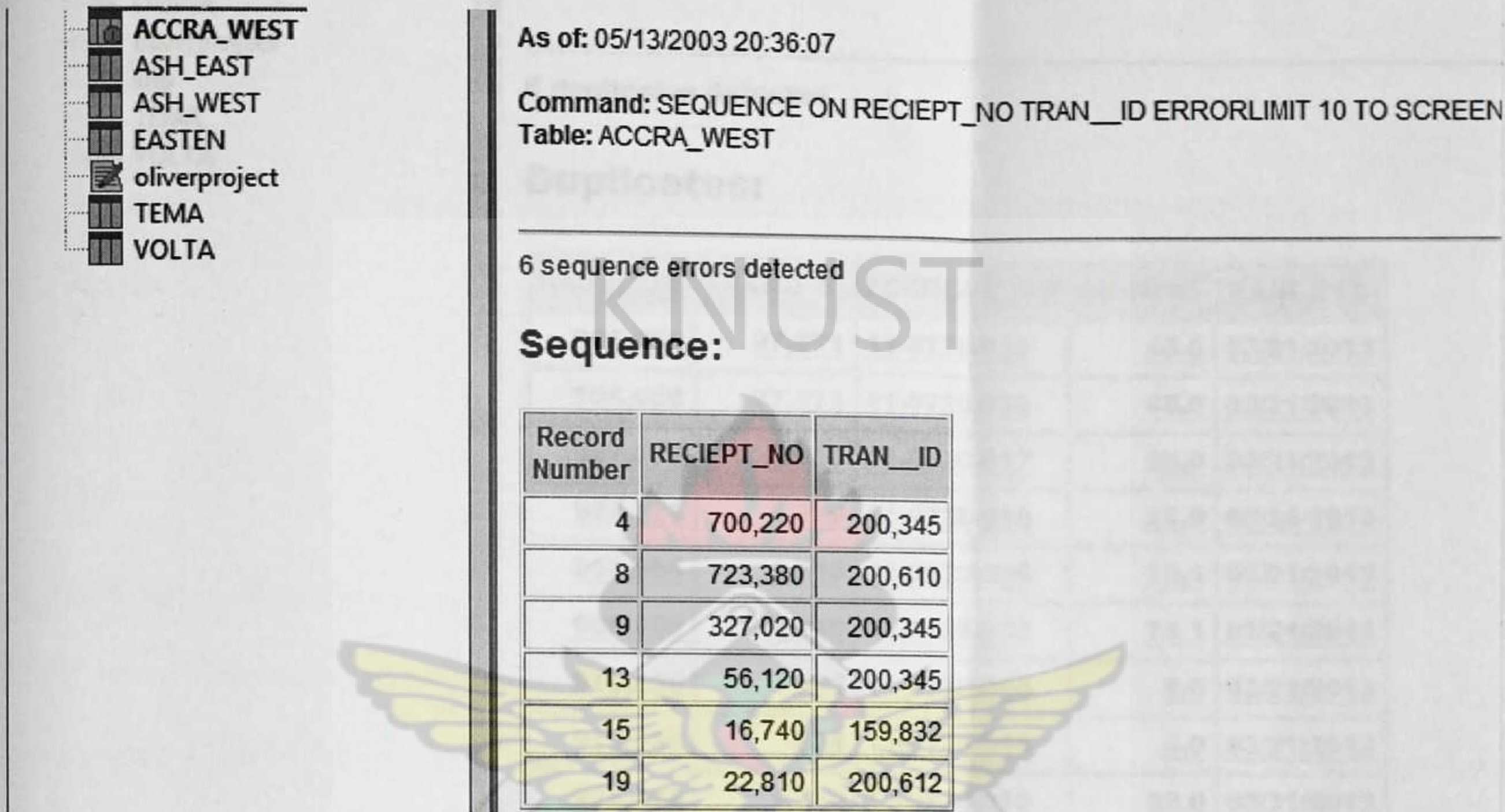


Figure 4.3. Sequence check on receipts numbers and Trans id.

It was observed that out 900 data sets analyzed on receipts and transaction ids, **six duplicates detected**, showing the receipt numbers, Tran's id, the amount involved and the paid dates, which were updated in the system. There were some fraudulent activities detected which indicated that they were been done by the operators of the system who updated the database of the organization and collecting the corresponding cash from the respective customers. (Shown in figure 4.4).

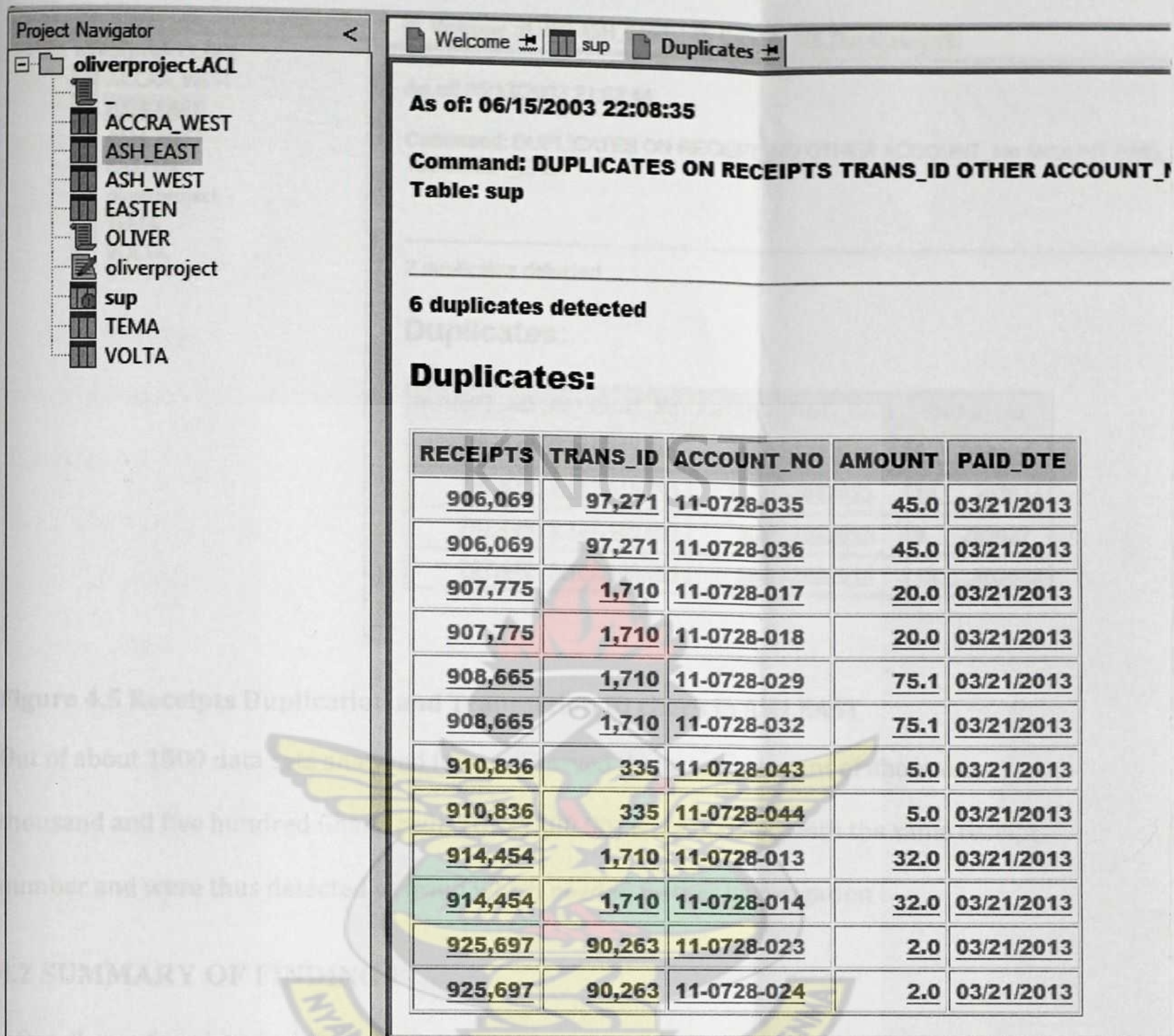


Figure 4.4 Receipts Duplication and Transaction ID check

The process was repeated in all the regions data set collected and it was observed that in ASH EAST, there were some receipt duplications. Figure 4.5 shows the results from ASH EAST.

Project Navigator

- oliverproject.ACL
 - ACCRA_WEST
 - ASH_EAST
 - ASH_WEST
 - EASTEN
 - oliverproject
 - TEMA
 - VOLTA

Welcome | ASH_EAST | Gaps | Duplicates

As of: 05/13/2003 21:03:44

Command: DUPLICATES ON RECIEPT_NO OTHER ACCOUNT_No AMOUNT PAID_
Table: ASH_EAST

2 duplicates detected

Duplicates:

RECIEPT_NO	ACCOUNT_No	AMOUNT	PAID_DATE	TRAN_ID
785,670	9,141,819,001	147.0	2082013	2. 203024
785,670	9,141,819,001	147.0	2082013	14. 203024
787,440	9,187,189,001	50.0	2082013	9. 203022
787,440	9,187,189,001	50.0	2082013	19. 203022

Figure 4.5 Receipts Duplication and Transaction ID check in ASH EAST

Out of about 1000 data sets analyzed it was observed that a total amount of about one thousand and five hundred Ghana cedis (GH¢1500.00) were updated with the same receipt number and were thus detected as fraud which needed further investigation

4.2 SUMMARY OF FINDINGS

After the entire data had been collected and analyzed using adaptive and rule learning techniques, it was observed that in Accra West a total of 900 customer payment data sets updated, 18 gaps existed between the receipt numbers and were detected as fraud with transaction ids contributing to 35 missing receipts and such was its equivalent amount in cash for fraud. In sequential receipt check for fraud, it was observed that 6 records had their receipt numbers and transaction identification NOT sequentially followed which also calls for investigation for detected fraud but contained no receipt duplication.

Volta region data set revealed that, out of 500 considered, no fraud was detected on the entire checks performed. Eastern region data revealed that, 10 gaps were identified out of the 453 data set on the receipts numbers and transaction ids, showing that about 27 receipts numbers were missing in the data set. This was detected as fraud and need to be investigated.

Table 4.1 SUMMARY OF RESULTS FROM ANALYSIS

REGIONS	TOTAL DATA	Gaps (Tr)	Gaps(Rr)	Dup(Rr)	Seq(Tr)	Missing(Rs)
ASH EAST	1000	103	103	10		25
ASH WEST	650	15	10	7	29	
EASTERN	453	10		27		
TEMA	820	12		4		23
VOLTA	500	0	0	0	0	0

CHAPTER 5

CONCLUSIONS AND RECOMMENDATIONS

5.1 CONCLUSION

The findings of this thesis indicate that there is always some amount of fraudulent activities in any organization or institution that handles pools of data in its already updated system or about to be updated which may contribute to the revenue level of that organization.

Fraud behaviour changes frequently as population of customer data increases to adapt to detection technique and fraud detection systems as well.

From the research and the experiment performed it shows that increase in data size affects the execution time for fraud detection, for a data size of 20000, the execution time 3 minutes 45 seconds but as increase to 25000, the execution time is 4 minutes 15 seconds.

However, in order to build a good monitoring system, we must know which aspects of customers' behaviour key in the process to be designed. Historically, determining such aspects has involved a good deal of mathematical analysis of behaviour attributes, building monitors, testing and determining how to combine them in a large amount of data set.

This research presented and demonstrated a framework that automates the process of detecting fraud using any three (3) unique identities in a data set for its algorithm modeled. This framework is not specific to Electricity Company of Ghana, but may be applied to any organization that uses a pool of data set (raw data to be processed or in system, i.e. processed data) for fraudulent detection checks.

In cases of supervised and unsupervised approaches, fraud detection can be used for counterterrorism work, actual monitoring systems and text mining from law enforcement.

The modeled diagram for the adaptive and rule learning algorithm can be calculated using probability ratio of how frequent it occurs as fraud and the time of execution of the algorithm on particular data set can also be searched on using fuzzy logic and neural network in AI.

5.2 RECOMMENDATIONS

- The Electricity Company of Ghana management can use the developed algorithm to test their entire database for detecting of fraudulent activities and contribute to their revenue booster in the company.
- Financial Organizations that make use of pool of customers behaviour in their systems for payments (crediting, debiting and usage of facilities such ATM) can also use this modeled algorithms for fraud detection in their database systems.
- Any organization or institution that makes use of data set with any three (three) unique attributes in their database could implement this modelled algorithm for detection of fraudulent activities.

REFERENCES

1. Abbott, D., Matkovsky, P. & Elder, J. (1998). An Evaluation of High-End Data Mining Tools for Fraud Detection.
2. Aleskerov, E., Freisleben, B. & Rao, B. (1997). CARDWATCH: A Neural Network-Based Database Mining System for Credit Card Fraud Detection. Proc. of the IEEE/IAFE on Computational Intelligence for Financial Engineering.
3. Anderson, K. and Kerr C. (2002) Customer Relationship Management, McGraw-Hill, New York.
4. Aronis, J. and Provost, F. (1996-7). Increasing the efficiency of data mining algorithms with breadth-first marker propagation.
5. Artis, M., Ayuso M. & Guillen M. (1999). Modelling Different Types of Automobile Insurance Fraud Behaviour in the Spanish Market.
6. Bain, T., Benkovich. M., Dewson, R., Ferguson, S., Graves, C., Joubert, T., Lee D., Scott, M., Skoglund, R., Turley, P. and Youness, S. (2001) Professional SQL Server 2000 Data Warehousing with Analysis Services.
7. Barse, E., Kvarnstrom, H. & Jonsson, E. (2003). Synthesizing Test Data for Fraud Detection Systems. Proc. of the 19th Annual Computer Security Applications Conference.
8. Bayerl, S., Bent, G., Lee, J. and Schommer, C. (2001) Mining Your Own Business in Telecoms Using DB2 Intelligent Miner for Data, IBM Corporation, San Jose.
9. Belhadji, E., Dionne, G. & Tarkhani, F. (2000). A Model for the Detection of Insurance Fraud. The Geneva Papers on Risk and Insurance.
10. Bell, T. & Carcello, J. (2000). A Decision Aid for Assessing the Likelihood of Fraudulent Financial Reporting. Auditing: A Journal of Practice and Theory.

11. Beneish, M. (1997). Detecting GAAP Violation: Implications for Assessing Earnings Management among Firms with Extreme Financial Performance. *Journal of Accounting and Public Policy*.
12. Bentley, P. (2000). Evolutionary, my dear Watson: Investigating Committee-based Evolution of Fuzzy Rules for the Detection of Suspicious Insurance Claims. *Proc. of GECCO2000*.
13. Bentley, P., Kim, J., Jung, G. & Choi, J. (2000). Fuzzy Darwinian Detection of Credit Card Fraud. *Proc. of 14th Annual Fall Symposium of the Korean Information Processing Society*.
14. Berry, M. and Linoff, G. (2004) *Data Mining Techniques: for Marketing, Sales, Customer Relationship Management - Second Edition*,
15. Bhargava, B., Zhong, Y., & Lu, Y. (2003). *Fraud Formalisation and Detection*.
16. Bolton, R. & Hand, D. (2001). *Unsupervised Profiling Methods for Fraud Detection. Credit Scoring and Credit Control VII*.
17. Bolton, R. & Hand, D. (2002). *Statistical Fraud Detection: A Review (With Discussion). Statistical Science*.
18. Bonchi, F., Giannotti, F., Mainetto, G., Pedreschi, D. (1999). *A Classification-based Methodology for Planning Auditing*.
19. Dzeroski, S. 1996. *Inductive logic programming and knowledge discovery in databases. Advances in Knowledge Discovery and Data Mining, Menlo Park*.
20. Davis, A. and Goyal, S. (1993). *Management of cellular fraud: Knowledge-based detection, classification and prevention*.
21. Ezawa, K. and Norton, S. (1995). *Knowledge discovery in telecommunication services data using bayesian network models*.

22. Ezawa, K. and Norton, S. (1996.) Constructing Bayesian networks to predict uncollectible telecommunications accounts. IEEE Expert
23. In Proceedings of First International Conference on Knowledge Discovery and Data Mining, U. Fayyad and R. Uthurusamy (Eds.), Menlo Park.

```
import java.io.*;
import java.io.*;
import java.util.*;
import java.util.*;
```

```
import java.util.*;
```

```
public class DuplicateTracker {
```

```
    public String fileName;
    public Workbook workbook;
    public Sheet sheet;
    private int rowNumber = 0;
```

```
    // Constructor for this is the default constructor for the excel reader
    // ...
}
```

```
public DuplicateTracker {
```

```
    // TODO: Add your code here
    // ...
    // ...
}
```

```
public void perform {
```

```
    try {
```

```
        workbook = Workbook.getWorkbook(new File(fileName));
        sheet = workbook.getSheet(0);
        checkForDuplicates(sheet);
```

```
        // printFileContent();
    } catch (Exception e) {
        System.out.println("Error : " + e.toString());
        e.printStackTrace();
        // TODO: handle exception
    }
}
```

KNUST



APPENDICES

APPENDIX A

CODE FOR MODELLED ALGORITHM IN JAVA

```
package gh.edu.knust.idl.mit.duplicator;

import java.io.File;
import java.io.InputStream;
import java.util.ArrayList;
import java.util.List;

import jxl.*;

public class DuplicateTracker {

    public String fileName;
    public Workbook workbook;
    public Sheet sheet;
    private int rowStartFrom = 0;
    /*
     * @ExcelReader this is the default Constructor for the excel reader
     *
     * */

    public DuplicateTracker() {
        // TODO Auto-generated constructor stub
        //setFileName("/home/dev/corenett/project/kn/doc/Donewell220820123.xls");
        setFileName("C:\\Users\\OLIVER\\Documents\\oli2.xls");
        //performExcelRead();
    }

    public void performExcelRead() {
        try {

            workbook = Workbook.getWorkbook(new File(getFileName()));
            setSheet(workbook.getSheet(0));
            checkForDuplicates(false);

            // printFileContent();
        } catch (Exception e) {
            System.out.println("Error : " + e.toString());
            e.printStackTrace();
            // TODO: handle exception
        }
    }
}
```



```

    datePaid = getColumnContent(4, i);
    txnCode = getColumnContent(5, i);

    if (receiptNo.equals(tempReceipt)
    txnCode.equalsIgnoreCase(tempTxnCode)) {
        repeating++;
        System.out.println(repeating + "\t\t" + i + "\t\t" + accountNo + "\t\t\t" +
        amountPaid + "\t\t\t" + payee + "\t\t\t" + receiptNo + "\t\t\t" + datePaid + "\t\t\t" +
        txnCode);
        status = "REPEAT";
    } else if (validateAccountNo(accountNo)) {
        // System.out.println("Account Number is valid");
        status = "VALID";
    } else {
        //System.out.println("Invalid Account No");
        status = "INVALID_ACC";
    }

    transactions.add(new TransactionInfo(accountNo, receiptNo, datePaid,
    txnCode, payee, amountPaid, status));
    tempReceipt = receiptNo;
    tempTxnCode = txnCode;
}
System.out.println("Row Cnt : " + rows);
return transactions;
}

public boolean validateAccountNo(String accNo) {

    //Write Query here to validate the account number

    return true;
}

String getFileName() {
    return fileName;
}

void setFileName(String fileName) {
    this.fileName = fileName;
}

Workbook getWorkbook() {
    return workbook;
}

```

```

void setWorkbook(Workbook workbook) {
    this.workbook = workbook;
}

Sheet getSheet() {
    return sheet;
}

void setSheet(Sheet sheet) {
    this.sheet = sheet;
}

public static void main(String[] arg) {
    DuplicateTracker reader = new DuplicateTracker();
    reader.performExcelRead();
    // String[][] readList = reader.readFileContent(true);
}

public int getRowStartFrom() {
    return rowStartFrom;
}

public void setRowStartFrom(int rowStartFrom) {
    this.rowStartFrom = rowStartFrom;
}
}

```

APPENDIX B

CODE FOR ORACLE EXTRACT FROM DATABASE

```

VERIFY FIELDS ACCOUNT_No AMOUNT PAID_DATE RECIEPT_NO TRAN__ID ERRORLIMIT 10
TO SCREEN
GAPS ON RECIEPT_NO PRESORT TO SCREEN
SEQUENCE ON RECIEPT_NO TRAN__ID ERRORLIMIT 10 TO SCREEN
DUPLICATES ON RECIEPT_NO OTHER ACCOUNT_No AMOUNT PAID_DATE RECIEPT_NO
TRAN__ID PRESORT TO SCREEN
SET FILTER
SET FILTER
DUPLICATES ON RECIEPT_NO OTHER ACCOUNT_No AMOUNT PAID_DATE RECIEPT_NO
TRAN__ID PRESORT TO SCREEN
OPEN ASH EAST
VERIFY FIELDS ACCOUNT_No AMOUNT PAID_DATE RECIEPT_NO TRAN__ID ERRORLIMIT 10
TO SCREEN
SEQUENCE ON RECIEPT_NO ERRORLIMIT 10 TO SCREEN
SET FILTER

```

GAPS ON RECIEPT_NO PRESORT TO SCREEN
DUPLICATES ON RECIEPT_NO PRESORT TO SCREEN
DUPLICATES ON RECIEPT_NO OTHER ACCOUNT_No AMOUNT PAID_DATE RECIEPT_NO
TRAN__ID PRESORT TO SCREEN
OPEN ASH_WEST
SEQUENCE ON RECIEPT_NO ERRORLIMIT 10 TO SCREEN
DUPLICATES ON RECIEPT_NO OTHER ACCOUNT_No AMOUNT PAID_DATE RECIEPT_NO
TRAN__ID PRESORT TO SCREEN

KNUST

