A Framework for Web Based Detection of Journal Entries Frauds using Data Mining Algorithm

Awodele O.*[1], Akinjobi J.**[2], Akinsola J. E. T.*[3]

*Department of Computer Science, Babcock University, Ilishan-Remo, Ogun State, Nigeria.
**Department of Computer Science, Crawford University, Igbesa, Ogun State, Nigeria.

Abstract *— Fraud detection has been a major challenge in the financial industry. The fraud menace has made a lot of organization lost billions of naira especially in multi-divisional and multi-branch enterprises. Hence, there is a need for pragmatic approach to proffer solution to this challenge. The methodological approach used in this work involves the development of framework which includes the remote extraction of financial Journal Entries (JE) from each branch location of multi-divisional multi-branch enterprises and integrated into an SQL Server database using a standard data format through SQL Server Management Studio. The extracted data is thereafter used to build a central data warehouse that is transmitted to the auditors at the corporate headquarters of the enterprise using web applications tools through Internet. A Decision Tree data mining algorithm constructed is applied by the auditors at the corporate headquarters on the data warehouse to detect possible financial JE fraud. The tasks are guided by the concept of a three tier client/server architecture in which the data extraction and data warehouse construction tasks constitute a data/backend tier, the transmission of the data warehouse through the web services constitutes the application/ middle tier while the decision tree data mining algorithm application for fraud detection through a user interface program of Active Server Pages (ASP.NET) constitute a presentation tier. Therefore, supervised predictive machine learning was employed in this study because the classes for fraud detection are user defined, they are ensured to conform to the classification hierarchy of the Journal Entries (JE). The use of training data improves the ability to differentiate between classes with similar journal profiles using the methods that are more reliable and produce more accurate results.*

Keywords: *Data Mining, Data Warehouse, Decision Tree, Journal Entries, Machine Learning, Web Services.*

I. INTRODUCTION

In today business world, data and fund transfer are part of primary business activities. So prevention of fraud, abuse and data alteration has become a major concern of many organizations. Data analysis techniques used before were primarily oriented toward extracting quantitative and statistical data characteristics.

Data mining is known as gaining insights and identifying useful patterns from the huge amount of data stored in large databases in such a way that the patterns and insights are statistically reliable, previously unknown, and understandable [13], [19], [5], and [29]. Data mining is also defined as a process that uses statistical, mathematical, artificial intelligence and machine learning techniques to extract and identify useful information and subsequently gaining knowledge from a large database [36] and [18].

According to [14] the existence of multi-locations increases audit-risk because transactions data are not centralized but scattered in systems existing in such remote locations. This situation cannot allow frauds to be easily and completely detected. Each location not visited increases the risk of errors and irregularities such as local fraud, non-existent assets, and phantom employees. [2] the auditor must understand the unique risks of the multi-location audit and the specific characteristics of the individual client. He should thereafter select audit procedures that guarantee an effective audit and reduce the audit risk to an acceptable level.

In order to detect and consequently prevent financial frauds particularly those relating to journal entries, it becomes necessary to automate the audit tasks through data mining techniques [3]. In order to achieve some cost savings from travelling and labour, improve effectiveness and productivity, the audit tasks can then be carried out remotely using web applications and interface programs [21].

DATA MINING PARADIGMS

[15] Data mining is the process of discovering actionable information from large sets of data. Data mining uses algorithms to derive patterns and trends that exist in data. Typically, these patterns cannot be discovered by traditional data exploration tools because the relationships and interdependencies may be too complex or because there is too much data. These patterns and trends can be collected, defined,

trained and analyzed as a data mining model in order to bring out the desired values and insights. Building a mining model is part of a larger process that includes everything from asking questions about the data and creating a model to answer those questions, deploying the model into a working environment. The time taken to build a model and precision (accuracy) is a factor on one hand; while kappa statistic and Mean Absolute Error (MAE) is another factor on the other hand [26].

JOURNAL ENTRIES PROCESS

Modern multidivisional enterprises commonly measure and monitor their financial activities through journal entries. These journal entries are summarized into a general ledger that is eventually used to produce financial statements that include trial balances, profit and loss accounts, balance sheet (see figure 1), using relevant paper-based information technology [28].

These financial statements are however recorded after a while when the financial transactions generating them particularly from the branches in distant locations, have been concluded. This process results in a delayed detection of fraud that may have been perpetuated through the journal entries. Financial fraud detection is more effective if audit consideration is given to frauds originating in journal entries as the primary level of all accounting statement preparation [27].

Many occurrences of journal entries fraud include the following common symptoms:

i. Posted journal entries with repetitive and unique account sequences.
ii. Journal entry postings made on weekends and holidays.
iii. Postings made in prior year and after fiscal year end.
iv. Postings made to suspense accounts
v. Journal entries errors corrected.
vi. Postings involving round multiples of 10,000,100,000 and 1,000,000.
vii. Postings made below/above approval limits.
viii. Unauthorized postings.
ix. Postings in which debits less credits do not net to zero.

THE JOURNAL ENTRIES (JE) FRAUD DETECTION CONCEPT USING DECISION TREE DATA MINING ALGORITHM

Decision Tree is a supervised data mining algorithm. Decision Trees (DT) are trees that classify instances by sorting them based on feature values. Each node in a decision tree represents a feature in an instance to be classified, and each branch represents a value that the node can assume. Instances are classified starting at the root node and sorted based on their feature values [22]. Decision tree learning, used in data mining and machine learning, uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. More descriptive names for such tree models are classification trees or regression trees [16]. Decision tree classifiers usually employ post-pruning techniques that evaluate the performance of decision trees, as they are pruned by using a validation set. Any node can be removed and assigned the most common class of the training instances that are sorted to it [22].

DT consists of three elements: the root node, interior node and a leaf node. The top most element is the root node. The leaf node is the terminal element of the structure and indicates the value of the target attribute (class) of JE data set. A JE data set is a collection of related journal entries such as a set of journal entries errors corrected or a set of unauthorized journal entries' postings (also see the list above for other types) [4]. The major objectives of grouping journal entries into data sets are efficient retrieval, searching, sorting and recognition.

CONCEPTUAL MODEL OF THE JOURNAL ENTRIES FRAUD DETECTION

The conceptual model is hinged on the following:

a. Each of the branches consolidate its operational activities with its division while all divisions are consolidated into the corporate accounts.

b. The existence of multi branches and divisions with their individual characteristics creates decentralized transactional data in different formats thereby increasing audit risks and costs.

c. Data centralization/aggregation entails remote data extraction, transformation and loading from each branch into a centralized data warehouse using SQL server integration services and data transmission facilities of web applications and services [30].

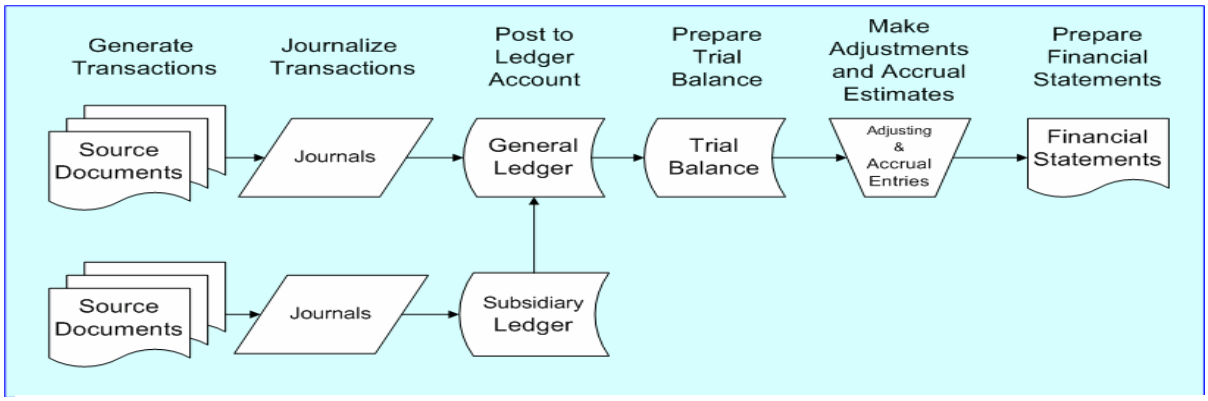

Figure 1: Overview of the Accounting Process [6]

The paper focuses on multidivisional, multi-branch enterprises. Such enterprises could be conglomerates, institutional, commercial or governmental establishments that are subdivided into smaller dependent operational units as subsidiaries, branches, campuses or any other form of sub systems that guarantee effectiveness and reliable operational control. Each sub operating unit has its own independent and unique business functions and goals but ultimately committed to achieving the overall objective of the establishment.

The remaining part of this work is arranged as follows: Section 2 presents the literature review discussing machine learning, data mining and classification, and Enterprise Branches Audit Processes; section 3 presents the methodology used, section 4 discusses the implementation of the work while section 5 gives the conclusion and recommendation for further works.

II. Literature Review

A. *Machine Learning*

Machine learning (ML), a sub-field of artificial intelligence (AI), focuses on the task of enabling computational systems to learn from data about how to perform a desired task automatically. Machine learning has many applications including decision making, fore- casting or predicting and it is a key enabling technology in the deployment of data mining and big data techniques in the diverse fields of healthcare, science, engineering, business and finance [1] The goal of machine learning is to enable a system to learn from the past or present and use that knowledge to make predictions or decisions regarding unknown future events. In the most general terms, the workflow for a supervised machine learning task consists of three phases: build the model, evaluate and tune the model, and then put the model into production [23]. ML algorithms require precision, accuracy and minimum error to have supervised predictive machine learning [26]

B. *Data Mining and Classification*

Data mining is the discovery of "models" for data [24] Big data caused an unprecedented explosion in the use of more reliable extensive data mining techniques, partially because the size of the information is much larger and because the information tends to be more varied and extensive in its very nature and content [7].

According to [1], data mining normally denotes the automation of pattern discovery and prediction from huge volumes of data using ML techniques. Data mining can also be used to denote an Online Analytical Processing (OLAP) or Structured Query Language (SQL) queries that entails retrospectively searching a large database for a specific query [9]. The method of extracting useful information or knowledge from the structured/ unstructured data and databases (relational and non-relational), using Data. Mining and ML techniques, is called knowledge discovery, sometimes collectively called KDD (Knowledge Discovery in Databases). Classification can be used to build up an idea of the type of customer / subscriber by describing multiple attributes to identify a journal class (entry).

C. *Enterprise Branches Audit Processes*

The current practice entails the internal audit team visiting the audit sites and running through a checklist of audit tasks and procedures in an effort to validate the accuracy of data, and the existing internal controls and business processes. At the preliminary planning stage the auditor examines financial reports existing in various locations. Statements of Accounting Standards (SAS), recommends detailed level audit of each enterprise location because doing so generally have a greater chance of detecting financial frauds [14]

Previous works on remote auditing included using Information and Communication Technology (ICT) facilities by internal auditors to interact with different departments and functions of the firm and third parties over long distances using remote communication technology, such as web

conferencing, electronic document management systems, and currently, remote access to information system clouds [35]

format. Other systems interact with the web services in a manner prescribed by its description using SOAP messages, typically conveyed using Hypertext Transmission Protocol (HTTP) with an

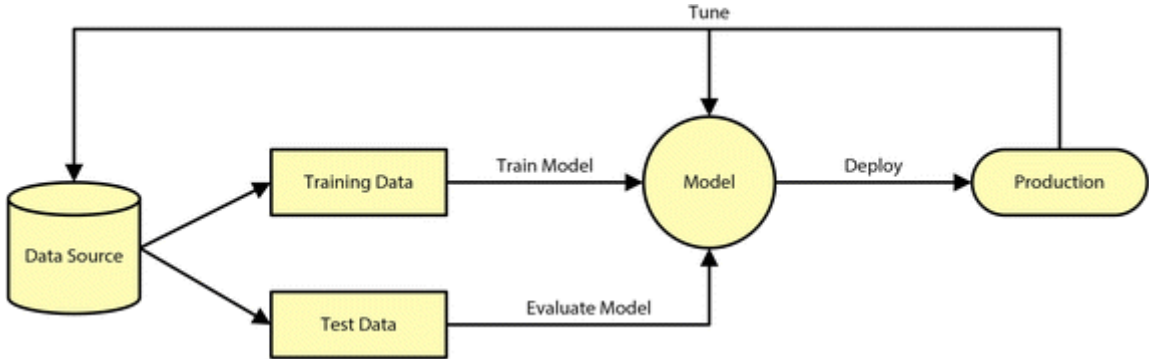

Figure 2: Supervised Machine learning workflow [23]

Some collaboration and communication tools are also run remotely on distributed systems over the Internet.

The remote audit practices mentioned, enable centralized evidence gathering and coordination within the audit team [8].

Table 1: Data Analysis Methods [25]

Type	Methods	Description	Example Methods
Machine Learning	Supervised Learning (SL)	The SL algorithms predict the events from learning models that are trained using labelled data points. The SL models are trained using labelled data points and tested with leave-one-out, cross validation, and 5-fold validation methods. The SL models are widely used for data classification and clustering. However, the SL algorithms have the limitations to handle concept drift (detecting changes) in the data streams	1. Neural Networks (Rojas, 2013) 2. Decision Trees (Barros, Basgalupp, De Carvalho, & Freitas, 2012) 3. Bayesian Networks (S. H. Chen &Pollino, 2012)

D. Application of Web Services in Remote Audit

Before data can be said to be remotely extracted for auditing, a web service has to serve as the intermediary between the data source and the target data warehouse source [20]. A web service by World Wide Consortium (W3C) definition is a software system designed to support interoperable machine to machine interaction over a network and has an interface described in a machine process-able

Extensible Mark-up Language (XML) serialization in conjunction with other web-related standards. It is therefore, a method of communication between two electronic devices over a network.

(Representational State Transfer. http://en.m.wikipedia.org/wiki/ Wikipedia.com.2014; http://en.m.wikipedia.org/wiki/Web_API).

III. METHODOLOGY

The Remote Audit Processes

These comprise the data extraction, data transformation, data loading, data warehousing, data transmission and data mining to detect fraud patterns and trends [20]. Details are illustrated below.

A. Architecture of the Remote Data Mining Audit of Journal Entries

A proposed conceptual three-tier architecture is used as the framework on which the enterprise remote audit tasks are modeled. It consists of the Front End otherwise called Presentation tier, the Connection string otherwise called Application/Business Logic/Middle tier and the Back End otherwise called the Data tier. See figure 3

B. Backend/ Data Retrieval Layer/Data Tier

1) Data Extraction, Transformation and *Loading*

Data extraction is the process of retrieving data out of unstructured or poorly structured data sources for further analyses and data processing or storage as in a data warehouse environment. The process in general is carried out at the data source using the

data source formats.

The goal of data extraction process is to bring all source data into a common, consistent format to be made ready for further processing. Looking at the figure 3 from the bottom up, the journal entries data mining process commences with the identification of the journal entries data, the format and their sources in each of the branches [31]. The identified data is then extracted, cleansed, transformed and loaded into an intermediate staging area with a uniform data format using Microsoft SQL Server Management Studio. This is to ensure that the data is correct and reliable for the construction of the Data Warehouse. The data warehouse is thereafter constructed using Microsoft SQL [34] Server Integration Services (SSIS) and ready to be transmitted to the corporate headquarters.

2) Building the Data Warehouse

Journal entries transaction input consists of the following data attributes. This is used to build the Journal entry tables for the data mining task as shown in Table 2.

It will be verified that the content types of the journal entry table (see Table 2) were assigned correctly, and any that is not will be modified.
The selected journal entries' data attributes and variables consist of both discrete and continuous attribute types. Since the algorithm to be implemented processes the dataset in number of iterations, the continuous type attributes will increase the load on the algorithm and thereby decreasing the performance.

Hence the continuous type attributes are converted to discrete values by applying the Supervised Discretized method module available in SQL Server Integration Services (SSIS) software. This method also helps to remove the outliers not completely removed during the data cleaning phase.

C. Middle Tier/ Business Logic Layer

1) Data Transmission

The transmission is done using web applications and services. Before data can be said to be remotely extracted for auditing, a web service has to serve as the intermediary between the data source and the target data warehouse source [11]. Branches data

Table 2: Journal Entries Data Attributes Transformation / Encoding

Input Variables	Data Type	ContentType	Usage
Journal Entry Number	Numeric	Discrete	
Key	Alphanumeric	Discrete	
Posting date	Alphanumeric	Discrete	
General Ledger Acct Type	Alphabetic	Discrete	
(e.g., suspense, receivable, revenue, expense, etc.)			
Entry date/time	Alphanumeric	Continuous	
Posting Period	Alphanumeric	Discrete	
General Ledger Account code	Alphanumeric	Discrete	
Journal Entry Serial Number	Numeric	Discrete	
Journal Entry Prepared by	Alphabetic	Discrete	
Journal Entry Adjusted by	Alphabetic	Discrete	
Journal Entry Approved by	Alphabetic	Discrete	
Approval Limit	Currency	Continuous	
Journal Entry Acct Descr	Alphabetic	Discrete	
Journal Entry Amount	Currency	Continuous	
Journal Entry transact descr	Alphabetic	Discrete	
Financial Statement Mapping	Binary	Discrete	
Debit presence indicator	Binary	Discrete	
Credit presence indicator	Binary	Discrete	
Journal Entry Line number	Numeric	Discrete	
Currency Sign	Text	Discrete	

The cleansing and transformation process would detect whether a column is discrete (categorical) or continuous by sampling and analyzing the source data and choosing an appropriate content type. If a continuous type is determined and a selected algorithm does not support continuous columns, the content type will be specified as DISCRETIZED (the continuous values broken into discrete ranges).

information and the divisional warehouse data are accessed from a Web server and loaded into the corporate data warehouse. The Web Services connection enables access to the divisional journal entries data warehouses existing as a single corporate data warehouse ready for the data mining audit using a decision tree algorithm [12]. The tasks here are to extract data from the SQL-Server hosting

the divisional data warehouse and transform it into an "understandable" form to call a web service [32]. A web service defined by a Web Service Description Language (WSDL), is called using the stored-procedures programmed in Visual Basic.NET within the SQL-Server.

and the application files are published to the webserver.

3) The Data Mining Algorithms for Constructing the Decision Tree

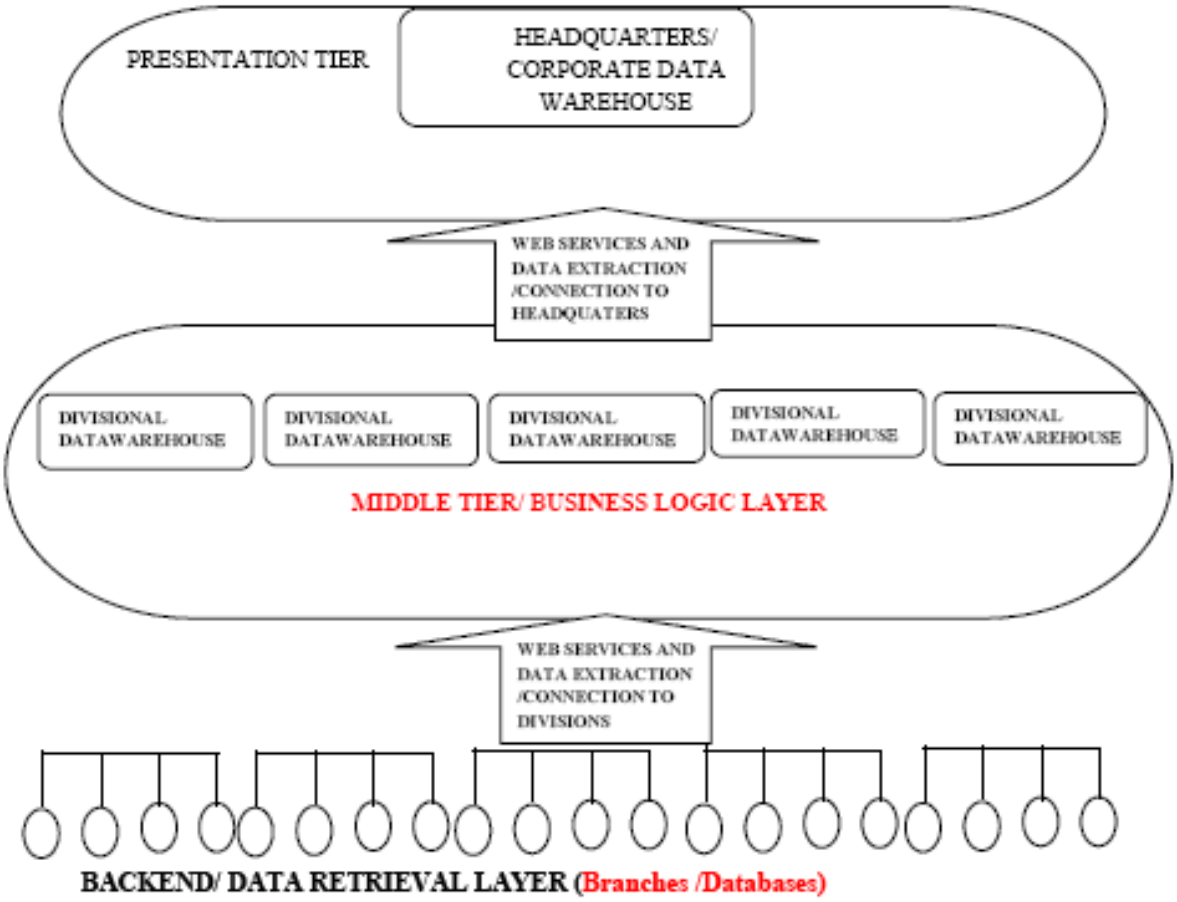

Figure 3: Architecture of the remote data mining detection of journal entries fraud

2) Presentation Tier

A) Data Warehouse at the Data Presentation Tier

As shown in the figure 3, this tier occupies the top level as the user interface. This tier communicates with other tiers by sending results to the browser and other tiers in the network. The transmittable journal entries data warehouse are accessed via a web browser enabling the auditor to commence the data mining audit and initiate further investigation to detect anomalous and potentially fraudulent journal entries [33] [10] [17].

B) Auditing Tasks at the Presentation Layer in the Corporate Headquarters

Setting Up the Data Mining Audit Environment at the Presentation Layer Setting up the application environment requires first that the database is created on the server. Database tables are created

The algorithm for the construction process consists of five steps as follow:

a. Create a whole JE training data set node.

b. If all the JE training data set are ‘non frauds’, then create a ‘non fraud’ node as the terminal node with whole of JE training data set and stop.

c. If all JE training data set are ‘frauds’, create a ‘fraud’ node with JE training data set as the terminal node with whole of JE training data set and stop.

d. Select an attribute (such as General Ledger Account type with values such as suspense, and receivables) and partition the JE training data set into subsets according to their attribute values.

Accordingly, create either ‘fraud’ nodes or ‘non fraud’ nodes with JE training data set as their parent and the attribute values as the label of the branch from JE training data set to its subsets.

e. For each JE training data set, replace the data set

with subsequent subsets of the JE training data set then go back to step b above.

The following journal entries transaction type is used as a sample for the decision tree data mining detection of JE entries fraud.

4) JE Summarized By Serial Numbers

Model showing the production rules employed in this work is in Figure 4

If JE has serial number missing, contains amounts and description, but not approved then label as non-fraud
If JE has serial number missing, but without amounts and description, then label as non-fraud
If JE has serial number missing, contains amounts field and description but approved then label as non-fraud
If JE has serial number missing, has the JE occurring more than once with amount and description then label as fraud
If JE does not have missing serial numbers and such serial number occurs once then label as non-fraud
If JE does not have serial number missing but same serial number occurs more than once without amount and description then label as non-fraud

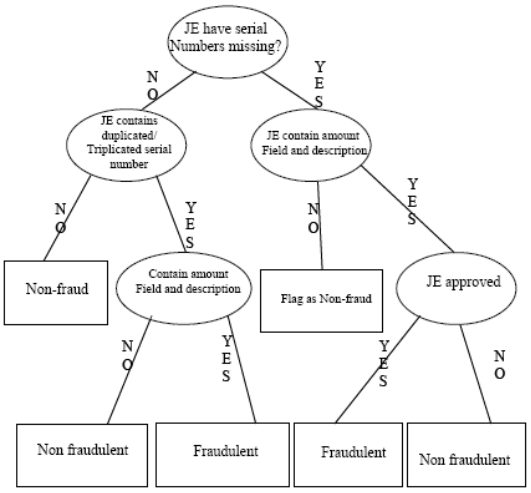

Figure 4: The Constructed Decision Trees Models

This is JE Summarized by serial numbers. This algorithm identifies JE with missing serial number, missing amounts and description

IV. IMPLEMENTATION

Microsoft Visual Studio is an IDE (Integrated Development Environment) for developing windows and web based software applications proprietary to Microsoft Corporation [34]. The visual studio dot net IDE has a number of programming languages and these are: C#.net (called C Zap.net), VB.net, and J#.net (called J zap dot net).

VB.net is used to implement this tool, because of its flexibility, simplicity and interoperability. The web application portion of the Visual Studio IDE is used because multiple branches are involved and they need to connect to the data warehouse server through web browsers.

Dot Net Framework serves as an application server and IIS served as a web server in this case. Without these two, there is no way a web application developed could be built and tested locally. When the application is finally deployed, the server must be provided with a server operating system, dot net Framework and IIS.

This tool was evolved through the use of vb.net aspx programming, HTML, CSS, Javascript and JQuery. The architecture of the system is depicted in Figure 5.

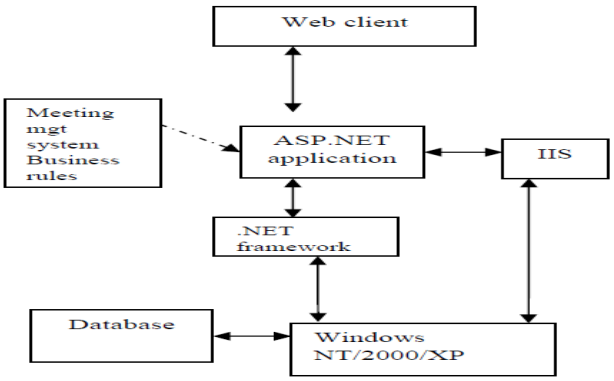

Figure 5: System Architecture for financial fraud Detection Software

The database was created in Microsoft SQL Server, This is because, Microsoft SQL Server is highly secured and highly scalable database system and this will definitely take care of our need.

An appointed systems administrator in a branch is provided with security credentials to log in and access the system and is required to post all journal entries in the branch.

Similarly, an auditor at the enterprise headquarters is provided with security credentials of a user name and password to login and gain access into the system to carry out the audit test.

The audit test uses the decision tree data mining algorithms to detect fraud in journal entries posted from the branches.

The final result of all the intervening data mining processes is shown in Figure 6.

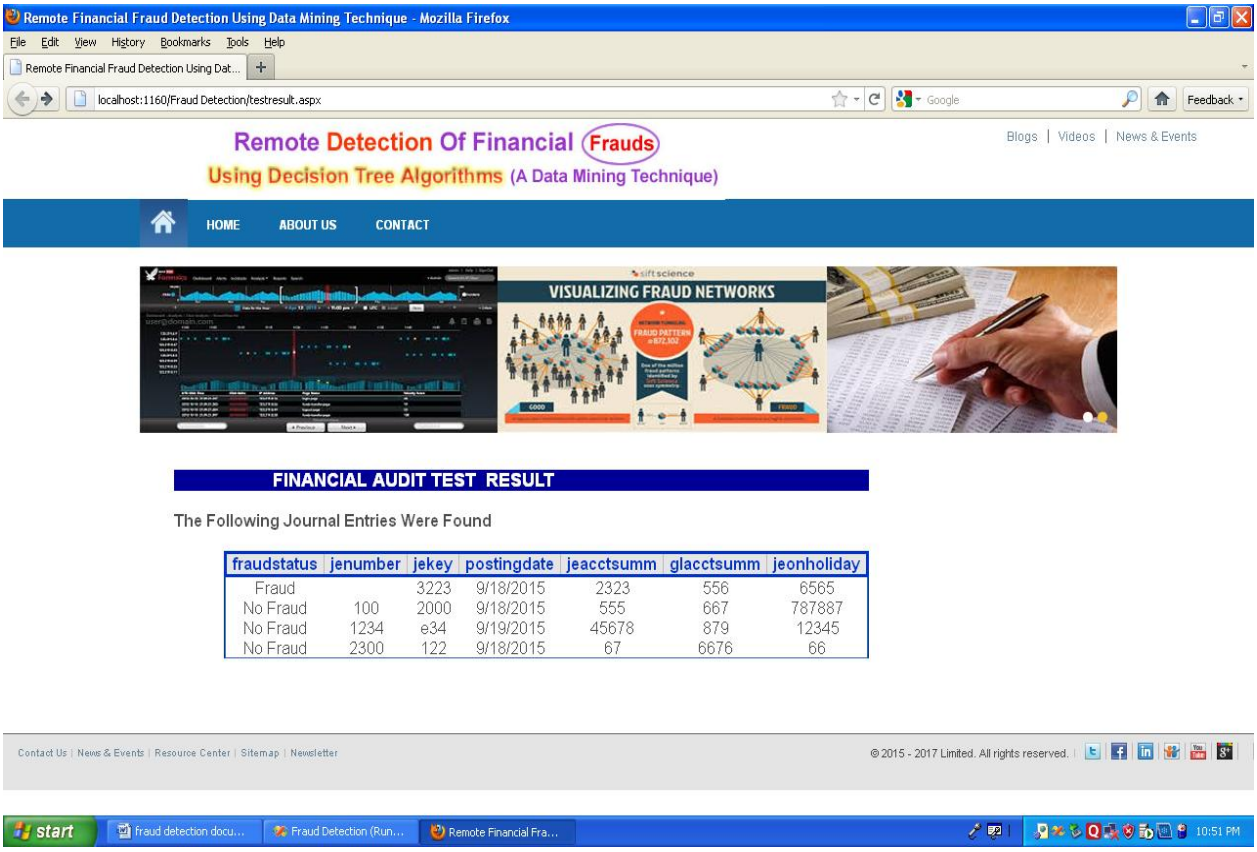

Figure 6: Results of the Decision Tree Analysis window

V. SUMMARY, CONCLUSION AND RECOMMENDATIONS FOR FURTHER STUDIES

Summary

Web applications and services that enabled the transmission of the journal entries from the remote branch locations to the enterprise headquarters were utilized and the data mining analysis of the data warehouse using a decision tree algorithm for detecting and predicting the journal entries frauds was carried out at the headquarters by the auditors. The concepts of front end layer at the headquarters, the applications/logic/middle layer constituting the web services and transmission of the data warehouse from branches to the enterprise headquarters, and the data/backend layer comprising data extraction processes at the locations, were also carried out.

Conclusion

This framework includes the remote extraction of financial journal entries from each branch location of multi-divisional multi-branch enterprises and integrated into an SQL Server database using a standard data format through SQL Server Management Studio. The extracted data is thereafter used to build a central data warehouse that is transmitted to the auditors at the corporate Headquarters of the enterprise using web applications tools. A decision tree data mining algorithm constructed is applied by the auditors at the corporate headquarters on the data warehouse to detect possible financial journal entries fraud. The tasks are guided by the concept of a three tier client server architecture.

Recommendations for further Research

The following are recommended for future research in order to address the limitations of the current study and to further justify the applicability of the framework developed. Future research could extend this study by replication in other class of enterprises such as multi-campus academic and financial institutions.

The fraud detection algorithms developed could further be subjected to application to other class of journal entries with fraud symptoms in Journal Entries postings. Other data mining algorithms such as Neural Networks, Multi-Layer Perceptron (MLP), Support Vector Machines (SVM) and k- means clustering will extend identification of potential threats or frauds when applied on the class of journal entries used in the framework.

REFERENCES

[1] Anwaar, A., Junaid, Q., Raihan, R., Arjuna, S., Andrej, Z. & Jon C. (2016). Big data for development: applications and techniques. Big Data Analytics, 1 / 2, 1- 24. ISSN: 2058-6345. doi: 10.1186/s41044-016-0002-4

[2] Aral, K.D., Güvenir, H.A., Sabuncuog, T., & Akar, A.R. (2011). A prescription fraud detection model. An Elsvier journal publication

[3] Argyrou, A. (2013). Auditing Journal Entries Using Extreme Value Theory. Proceedings of the 21st European Conference on Information Systems. 22, 00101 Helsinki, Finland,

[4] Argyrou, A. (2013). Developing Quantitative Models for Auditing Journal Entries. A PhD Theisis, Hanken School of Economics.

[5] Ashutosh, K. D., Animesh, K. D., Vipul, A. & Yogeshver, K (2012). "Knowledge Discovery with a Subset-Superset Approach for Mining Heterogeneous Data with Dynamic Support", Conseg-2012.

[6] Bay, S., Kumaraswamy, K., Anderle, M.G., Kumar, R., & Steier, D.M2012). Large Scale Detection of Irregularities in Accounting Data. Center for Advanced Research, PricewaterhouseCoopers

[7] Brown, M. (2012). Data mining techniques. Retrieved from website: https://www.ibm.com/developerworks/library/ba-data-mining-techniques

[8] Center for Audit Quality (CAQ) (2008). Practice Aid for Testing Journal Entries and Other Adjustments Pursuant to AU Section 316. A publication of the Center for Audit Quality.

[9] Chaudhuri, S., & Dayal, U. An Overview of Data Warehousing and OLAP Technology.Microsoft Research, Redmond Hewlett-Packard Labs, Palo Alto.

[10] Chintalapati S., Jyotsna G. (2013). Application of Data Mining Techniques for Financial Accounting Fraud Detection Scheme, *International Journal of Advanced Research in Computer Science and Software Engineering* Volume 3, Issue 11.

[11] Chorba, R. W., & Bommer, M. R. W. (1983).Methodology for the Construction of a Data Warehouse. Cordova, Argentina.

[12] Debreceny, R.S. & Gray, G.L. (2010). Data Mining Journal Entries for Fraud Detection: A Pilot Study

[13] Elkan, C. (2001). Magical Thinking in Data Mining: Lessons from COIL Challenge 2000. Proc. of SIGKDD01, 426-431.

[14] Ghanbari, M. K. & Einakian, M. (2014). Using "Data Mining" to Detect Frauds of Internal Audits. *Proceedings of 9th International Business and Social Science Research Conference*, Dubai, UAE, ISBN: 978-1-922069-41-2

[15] Han, J., & Kamber, M. (2000). Data Mining: Concepts and Techniques. Book published inSimon Fraser University

[16] Hastie, T., Tibshirani, R. & Friedman, J. H (2001). The elements of statistical learning. Data mining, inference, and prediction, 2001, New York: Springer Verlag.

[17] Inmon, W. H. (2005). Building the Data Warehouse (4th ed.). Indianapolis, IN: Wiley Publishing, Inc.

[18] Joseph, M. V. (2013). Significance of Data Warehousing and Data Mining in Business Applications. *International Journal of Soft Computing and Engineering (IJSCE)* ISSN: 2231-2307, Volume-3, Issue-1.

[19] Kai, L. & Peng, L. (2013). "A Selective Fuzzy Clustering Ensemble Algorithm", *International Journal of Advanced Computer Research (IJACR)*, Volume-3, Issue-13, December-2013, pp.1-6.

[20] Kimball, R. (2006). The data warehouse toolkit. John Wiley & Sons.

[21] Kirkos, E., Spathis, C., &Manolopoulos,Y. (2007). Data mining techniques for the detection of fraudulent financial statements. Expert Systems with Applications, 32, 995-1003.

[22] Kotsiantis, S. B. (2007). Supervised Machine Learning: A Review of Classification Techniques. Informatica 31 (2007). Pp. 249 – 268. Retrieved from IJS website: http://wen.ijs.si/ojs-2.4.3/index.php/informatica/article/download/148/140.

[23] Landset, S., Khoshgoftaar, T. M., Richter, A. N. & Hasanin, T. (2015). A survey of open source tools for machine learning with big data in the Hadoop ecosystem. Journal of Big Data (2015) 2 / 24. 1 – 36, doi: 10.1186/s40537-015-0032-1. Retrieved from: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-015-0032-1. Springer Open Journal, Springer Press

[24] Leskovec, J., Rajaraman, A. & Ullman, D. J. (2014). Mining of Massive Datasets. Retrieved from website: http://i.stanford.edu/~ullman/mmds/book.pdf

[25] Mohammad, A. A., Dusit, N., Shaowei, L., Hwee-Pink, T. & Zhu, H. (2016). Mobile Big Data Analytics Using Deep Learning and Apache Spark. Mobile big data analytics using deep learning and apache spark. IEEE Network. 30 / 3, 22-29. Research Collection School of Information Systems. ISS: N0890-8044. Identifier: 10.1109/MNET.2016.7474340I. Retrieved from website: http://ink.library.smu.edu.sg/sis_research/3422. Available at: http://doi.org/10.1109/MNET.2016.7474340. Institute of Electrical and Electronics Engineers (IEEE) Press

[26] Osisanwo F.Y., Akinsola J.E.T., Awodele O., Hinmikaiye J. O., Olakanmi O.,Akinjobi J. (2017). Supervised Machine Learning *Algorithms: Classification and Comparison. International Journal of Computer Trends and Technology (IJCTT)* – Volume 48 Number 3 June 2017 ISSN: 2231-2803 http://www.ijcttjournal.org. Pp. 128 - 138

[27] Ott, J., MacLeod, A., & Fan, K. Mar. (n.d). Computer-assisted Audit Techniques: Value of Data Mining for Corporate Auditors

[28] Ponniah, P. (2001). Data Warehousing Fundamentals: A Comprehensive Guide for IT Professionals. Canada: John Wiley & Sons, Inc.

[29] Preeti, K & Hitesh, G. (2012). "Finding Frequent Pattern with Transaction and Occurrences based on Density Minimum Support Distribution", International Journal of Advanced Computer Research (IJACR), Volume-2, Number-3, Issue-5 September-2012.

[30] Sarka D., (2013). Fraud Detection with the SQL Server Suite Parts1-4. A blog entry posted in SQL Server Blogs, Training

[31] Sarka, D. (2012). Data Mining with SQL Server. SolidQ. Retrieved from http://www.solidq.com/squ/courses/Pages/Data-Mining-with-SQL-Server-2012.aspx

[32] Saxena, G. &Agarwal, B. B. (2014). Data Warehouse Designing: Dimensional Modelling and E-R Modelling. International Journal of Engineering Inventions e-ISSN: 2278-7461, p-ISSN: 2319-6491 Volume 3, Issue 9

[33] Sharma, A., & Panigrahi, P.K. (2012).A Review of Financial Accounting Fraud Detection based on Data Mining Techniques. International Journal of Computer Applications (0975 – 8887) Volume 39.

[34] SQL Server Management Studio Express Microsoft.com (2011); Wikipedia, the free encyclopedia.com.)

[35] Teeter, R. A., Vasarhelyi, M. A., Alles, M. G. (2011).The Remote Audit. *Journal of Emerging Technologies in Accounting* Volume 7 Issue 1, pp. 73-88.

[36] Turban, E., Aronson, J. E., Liang, T. P., & Sharda, R. (2007). Decision Support and Business Intelligence Systems, Eighth edition, Pearson Education, 2007