A TIME FREQUENCY BASED SUSPICIOUS ACTIVITY DETECTION FOR ANTI-MONEY LAUNDERING

Murali Ponaganti

Associate Professor & HOD, Department Of MCA, Sree Chaitanya College of Engineering, Karimnagar

ABSTRACT: Money laundering is the crucial mechanism utilized by criminals to inject proceeds of crime into the financial system. The primary responsibility of the detection of suspicious activity related to money laundering is with the financial institutions. Most of the current systems in these institutions are rule-based and ineffective (over 90 % false positives). The available data science-based anti-money laundering (AML) models to replace the existing rule-based systems work on customer relationship management (CRM) features and time characteristics of transaction behaviour. Due to thousands of possible account features, customer features, and their combinations, it is challenging to perform feature engineering to achieve reasonable accuracy. Aiming to improve the detection performance of suspicious transaction monitoring systems for AML systems, in this article, we introduce a novel feature set based on time-frequency analysis, that uses 2-D representations of financial transactions. Random forest is utilized as a machine learning method, and simulated annealing is adopted for hyperparameter tuning. The designed algorithm is tested on real banking data, proving the results' efficacy in practically relevant environments. It is shown that the time-frequency characteristics are discriminatory features for suspicious and non-suspicious entities. Therefore, these features substantially improve the area under curve results (over 1%) of the existing data science-based transaction monitoring systems. Using time-frequency features alone, a false positive rate of 14.9% has been achieved, with an F-score of 59.05%. When combined with transaction and CRM features, the false positive rate is 11.85%, and the F-Score is improved to 74.06%.

I. Introduction:

Money laundering (ML) is the umbrella under which the legitimization of the proceeds of crime is attempted while laundered money can be both re-inserted into the legitimate economy and re-used to fuel further criminal activities. All major criminality such as drug and human trafficking, terrorism, extortion, kidnap-for-ransom, bribery, embezzlement, tax evasion, corruption and a multiplicity of other offenses (also known as predicate offenses) are connected through ML. Even though it is impossible to provide an accurate estimate of the size of such a complex underground market, the International Monetary Fund (IMF) indicates that every year, up to 2 trillion USD is laundered through financial systems globally, making ML one of the world's largest markets. To tackle this issue, most countries following the Financial Action Task Force (FATF) recommendations set up an anti-money laundering (AML) structure, as shown in Fig.1. It is the responsibility of the financial institutions to report suspicious activities to the Financial Intelligence Unit (FIU). The FIU collects intelligence from all different financial institutions within and outside the jurisdiction, which are later reported to the law enforcement agencies (LEA) as necessary. The police, using this intelligence, builds a case to the judicial system, and if ordered the Asset Recovery Bureau (ARB), recovers the suspicious assets for the public, closing the loop. As the initiator of the whole process, identifying the suspicious activity by the financial institutions is very critical. While technology is essential for the processing and identification of suspicious transactions given the volume of data that needs to be filtered, technology adoption in an AML-context needs to be carefully balanced against the various stakeholders in the AML chain of investigation. Also, most of the existing proposed software' solutions' are rule based, and only 25 % of respondents have already implemented Artificial Intelligence (AI) and expressed their main business drivers for machine learning in AML as anomaly detection, segmentation, and model tuning . However, the use of AI instead of rule-based engines for new cases detection is infrequent. The rule-based systems have three significant problems. First, any such software solution depends on a human workforce with varying performance and experience. Instead of enabling AML-analysts and FIU to make more meaningful decisions about what cases should be pursued, they create an unmanageable volume of data. With employees being bombarded by a constant stream of noise from technology-based alerts, it is no surprise that negative repercussions are experienced within financial institutions, which also propagate to the FIUs and the ARBs. As the number of false-positive alerts is over 90% of all alerts, AML experts are consumed by clearing false positives and confirming the non-suspicious nature of cases. This

contingency creates difficult work conditions for many AML employees. Often, employees that experience such a continuous stream of false positives will be desensitized towards actual suspicious cases. Second, most of the suspicious transactions does not even generate and alert since the rules are exposed to criminals from various channels. The exposure can be in the form of insider threats, employees collaborating with money launderers, reverse engineering of software path-dependencies, published Financial Action Task Force (FATF) typologies that are translated into threshold-based rules or contain specific behavioural traits that can be avoided). Third, the design of rules against new methods of laundering remains a reactive and lengthy process. According to the United Nations Office on Drugs and Crime (UNODC), just 0.2% of the activities can be detected [2]. Despite advances in computation, ML detection remains challenging as a complex behavioural, computational, socio-economic, and managerial problem. These problems resulted in the introduction of new methods of transaction monitoring using data science and machine learning techniques. However, most of the machine learning techniques are as successful as the quality of the input features. There are hundreds of potential features that can be used, such as ATM withdrawals, SWIFT transactions, online transfers, age, occupation. There are also combinations of features that can be created per channel, per time interval, per currency. The complete list of 237 transactional candidate features (related to the similar field of credit card fraud) has been shown in [3]. As a result, feature engineering (feature creation and selection) for AML is an essential yet very challenging and a time-consuming problem, as specified in [4]–[6]. It can take many weeks, if not months, to determine a useful combination of features out of thousand potential features to be employed. In this study, we propose a novel and a generalized solution using time-frequency (TF) analysis as a feature extraction method, so that with a handful of features, high-level accuracy can be achieved. Time-frequency features improve the accuracy of machine learning results compared to using transaction features alone. The proposed feature set can be utilized as a standard in suspicious transaction detection in order to shorten the feature engineering stage. There are three key contributions in this work at different stages: feature engineering methodology, model implementation and tests with current real banking data. The first one is a novel methodology for feature extraction in order to build data science models for AML in time and frequency, significantly reducing feature engineering workload. The second contribution is implementing 2D time-frequency features in building data science models for detection, improving the model precision. The third contribution is testing the models in real banking data and proving the improvements in the detection of suspicious activity. The remainder of the paper proceeds as follows. In the next section, we examine state of the art. In Section III, we present the proposed approach and the time-frequency features. Experiment details and the experimental results are given in Sections IV and V, respectively. Finally, we discuss the results and suggest possible future works.

1.1 Objective of the project:

Money laundering is the crucial mechanism utilized by criminals to inject proceeds of crime into the financial system. The primary responsibility of the detection of suspicious activity related to money laundering is with the financial institutions. Most of the current systems in these institutions are rule-based and ineffective (over 90 % false positives). The available data science-based anti-money laundering (AML) models to replace the existing rule-based systems work on customer relationship management (CRM) features and time characteristics of transaction behaviour. Due to thousands of possible account features, customer features, and their combinations, it is challenging to perform feature engineering to achieve reasonable accuracy. Aiming to improve the detection performance of suspicious transaction monitoring systems for AML systems, in this article, we introduce a novel feature set based on time-frequency analysis, that uses 2-D representations of financial transactions. Random forest is utilized as a machine learning method, and simulated annealing is adopted for hyperparameter tuning. The designed algorithm is tested on real banking data, proving the results' efficacy in practically relevant environments. It is shown that the time-frequency characteristics are discriminatory features for suspicious and non-suspicious entities. Therefore, these features substantially improve the area under curve results (over 1%) of the existing data science-based transaction monitoring systems. Using time-frequency features alone, a false positive rate of 14.9%

has been achieved, with an F-score of 59.05%. When combined with transaction and CRM features, the false positive rate is 11.85%, and the F-Score is improved to 74.06%.

II. Literature Survey:

"Estimating Illicit Financial Flows Resulting From Drug Trafficking and Other Transnational Organized Crimes"
This study's review of relevant reports concludes that the best estimate for the amount of criminally obtained money laundered by transnational organized crime is approximately 2.7 percent of the global gross domestic product (GDP) in 2009, which amounted to U.S. $1.6 trillion. The largest income for transnational organized crime apparently comes from the sale of illicit drugs, which accounts for 20 percent of all crime proceeds. This study's estimate of gross profits from global cocaine sales in 2009 is U.S. $84 billion, compared with approximately U.S. $1 billion paid to farmers in the Andean region. Most of the gross profits (retail and wholesale) were generated in North America (U.S. $35 billion) and in West and Central Europe (U.S. $26 billion). This report reminds readers that investments of illicit money into licit economies can cause problems that range from distortions of resources allocation to the "crowding out" of legitimate economic sectors. The largest outflows of illicit proceeds for laundering occur from countries in North America, South America, and Europe. These regions together account for 95 percent of all cocaine profit-related outflows worldwide. In terms of net outflows (outflows less inflows), the study model suggests that the main destination outside the regions where the profits were generated would be the Caribbean, with net inflows of approximately U.S. $6 billion. Extensive tables and figures and appended text of relevant sections of international legal instruments.

''Predicting credit card transaction fraud using machine learning algorithms,''

Credit card fraud is a wide-ranging issue for financial institutions, involving theft and fraud committed using a payment card. In this paper, we explore the application of linear and nonlinear statistical modeling and machine learning models on real credit card transaction data. The models built are supervised fraud models that attempt to identify which transactions are most likely fraudulent. We discuss the processes of data exploration, data cleaning, variable creation, feature selection, model algorithms, and results. Five different supervised models are explored and compared including logistic regression, neural networks, random forest, boosted tree and support vector machines. The boosted tree model shows the best fraud detection result (FDR = 49.83%) for this particular data set. The resulting model can be utilized in a credit card fraud detection system. A similar model development process can be performed in related business domains such as insurance and telecommunications, to avoid or detect fraudulent activity.

''Representation learning: A review and new perspectives,''

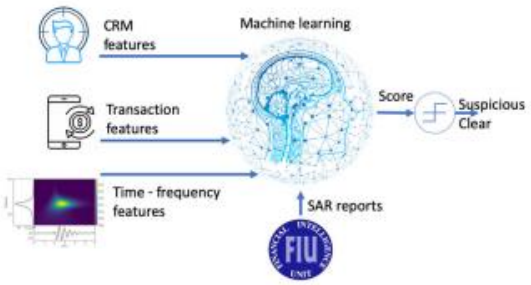
The success of machine learning algorithms generally depends on data representation, and we hypothesize that this is because different representations can entangle and hide more or less the different explanatory factors of variation behind the data. Although specific domain knowledge can be used to help design representations, learning with generic priors can also be used, and the quest for AI is motivating the design of more powerful representation-learning algorithms implementing such priors. This paper reviews recent work in the area of unsupervised feature learning and deep learning, covering advances in probabilistic models, autoencoders, manifold learning, and deep networks. This motivates longer term unanswered questions about the appropriate objectives for learning good representations, for computing representations (i.e., inference), and the geometrical connections between representation learning, density estimation, and manifold learning.

''**An empirical analysis of feature engineering for predictive modeling,''**

Machine learning models, such as neural networks, decision trees, random forests and gradient boosting machines accept a feature vector and provide a prediction. These models learn in a supervised fashion where a set of feature vectors with expected output is provided. It is very common practice to engineer new features from the provided feature set. Such engineered features will either augment or replace portions of the existing feature vector. These engineered features are essentially calculated fields, based on the values of the other features. Engineering such features is primarily a manual, time-consuming task. Additionally, each type of model will respond differently to different types of engineered features. This paper reports on empirical research to demonstrate what types of engineered features are best suited to which machine learning model type. This is accomplished by generating several datasets that are designed to benefit from a particular type of engineered feature. The experiment demonstrates to what degree the machine learning model is capable of synthesizing the needed feature on its own. If a model is capable of synthesizing an engineered feature, it is not necessary to provide that feature. The research demonstrated that the studied models do indeed perform differently with various types of engineered features.

III. System Analysis

System Architecture



3.1 Existing System

When just time-frequency characteristics are used, the F-score increases to 59.05% and the false positive rate drops to 14.9 percentage points. The false positive rate drops to 11.85% and the F-score rises to 74.06% when transaction and CRM features are taken into account.

Disadvantages of Existing System:

1. Less Prediction.
2. Security is less.

3.2 Proposed System

The term "money laundering" (ML) refers to the process by which unlawful earnings are made to appear more legitimate so that they can be utilized for lawful purposes again. All significant criminal activity is linked through ML, including drug and human trafficking, terrorism, extortion, kidnap-for-ransom, bribery, embezzlement, tax fraud, corruption, and a wide variety of other offences (sometimes called predicate offences). Although it is hard to give an exact assessment of the scale of such a complicated underground industry, the International Monetary Fund (IMF) estimates that up to $2 trillion USD is laundered annually via financial institutions throughout the world, making ML one of the world's largest marketplaces.

Advantages of Proposed System:

1. Security is more.

2. More Prediction.

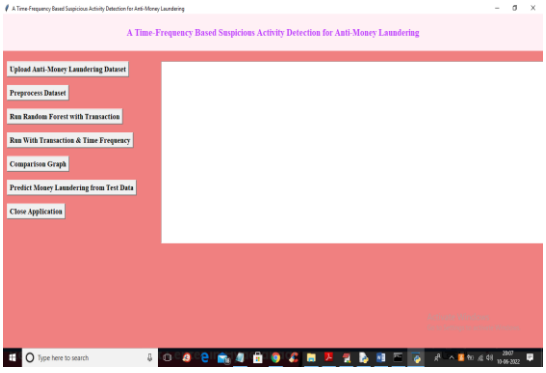
IV. Modules Information:

To implement this project, we have designed following modules

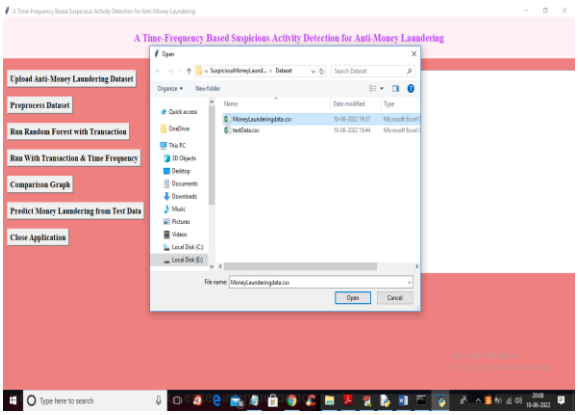
1) Upload Anti-Money Laundering Dataset: using this module we will upload dataset to application and then plot graph with normal and money laundering transaction count.
2) Preprocess Dataset: dataset contains missing and non-numeric data so Random Forest will accept only numeric data so by applying Label Encoder class we are converting non-numeric data into numeric data. Label encoders assign unique ID to each non-numeric data.
3) Run Random Forest with Transaction: processed data will be split into train and test where application will used 80% dataset to train Random Forest on transaction data and then this trained model will be applied on 20% test data to calculate FSCORE, True Positive and Negative Rate.
4) Run With Transaction & Time Frequency: using this module we will convert transaction data into Time Frequency by applying FFT algorithm and then this Time Frequency transaction data get trained with Random Forest to get FSCORE, TPR and FPR values.
5) Comparison Graph: using this module we will plot FSCORE Random Forest graph between Transaction data and Time Frequency FFT data.
6) Predict Money Laundering from Test Data: using this module we will upload test data and then apply Random Forest model to predict whether test data contains normal transaction features or Money-Laundering features.

V. Screen Shots

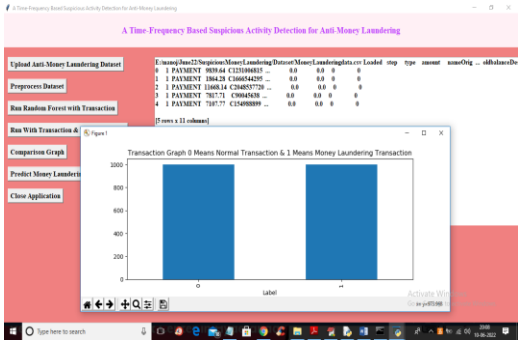
To run project double click on 'run.bat' file to get below screen



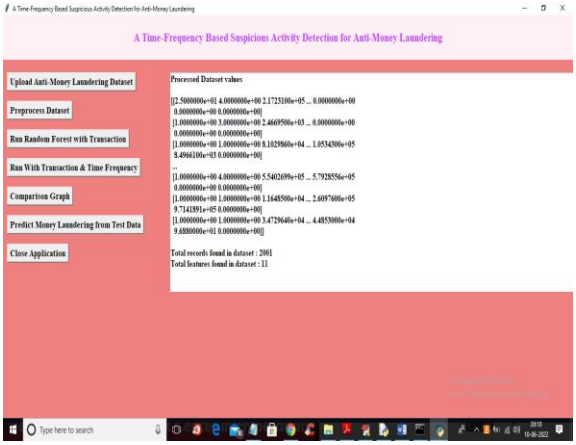
In above screen click on 'Upload Anti-Money Laundering Dataset' button to upload dataset and get below output



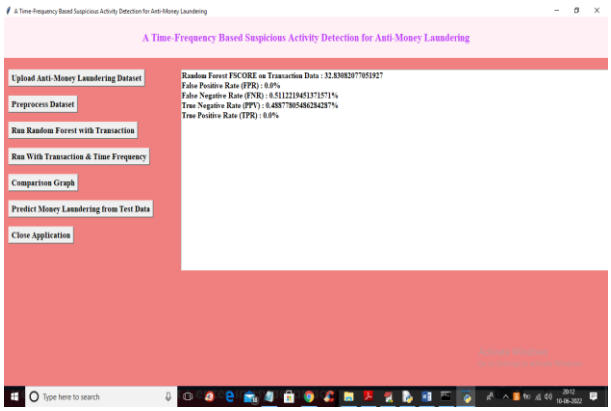
In above screen selecting and uploading 'Money Laundering' dataset and then click on 'Open' button to upload dataset and get below output



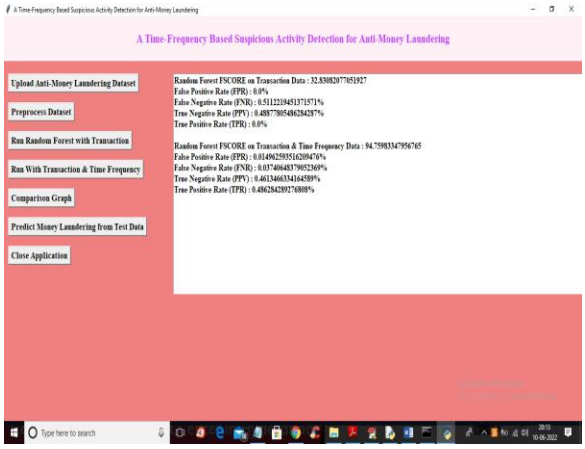
In above screen we can see dataset loaded and dataset contains some numeric and non-numeric data so we need to Preprocess data to convert into numeric and in above graph x-axis represents labels where 0 means Normal transaction and 1 means Money-Laundering transaction and y-axis represents counts. Now close above graph and then click on 'Preprocess Dataset' button to clean dataset and get below output



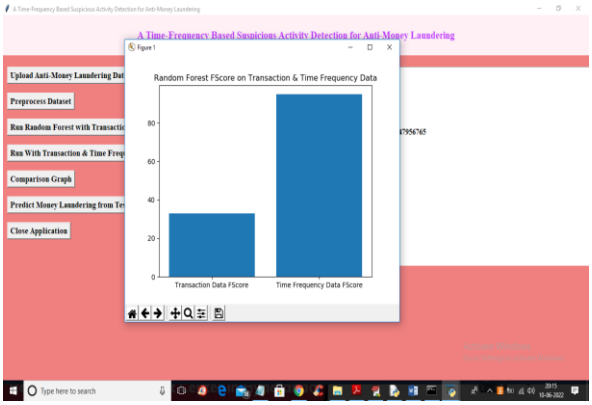
In above screen we can see entire dataset converted to numeric format and in last two lines we can see dataset contains 2001 records and each record contains 11 features and now dataset is ready and now click on 'Run Random Forest with Transaction' button to train Random Forest on transaction dataset and get beloww output



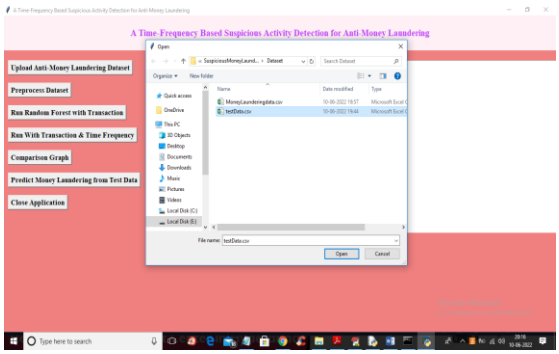
In above screen with Random Forest on transaction data we got FSCORE as 32% and True Positive Rate (TPR) as 0. So transaction data features will not allow Random Forest to train perfectly and now click on 'Run With Transaction & Time Frequency' button to convert Transaction data into Time Frequency and then retrain Random Forest to get below output



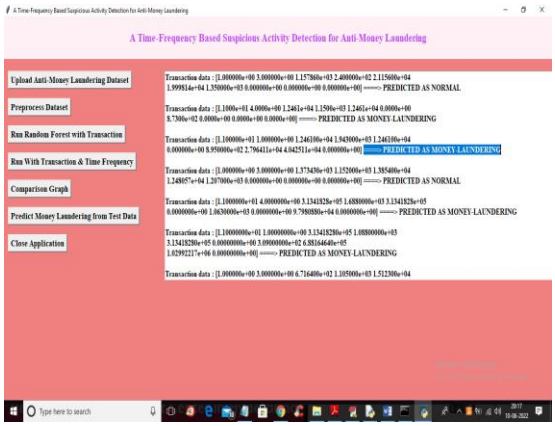
In above screen Random Forest with Time Frequency Transaction data got 94% FSCORE and TPR as 45% so by converting transaction data into time frequency we can build accurate prediction model. Now click on 'Comparison Graph' button to get below graph



In above graph x-axis represents technique name and y-axis represents FSCORE and in both techniques Time Frequency got high FSCORE and now close above graph and then click on 'Predict Money Laundering from Test Data' button to upload test data and get prediction output



In above screen selecting and uploading 'testData.csv' file and then click on 'Open' button to get below output



In above screen in square bracket we can see transaction test data and after arrow symbol =➔we got prediction output as 'Normal or Money Laundering'

VI. Conclusion:

In this paper, we have shown that adding time-frequency features, simplifies the feature selection process and improves the quality of the data science model. Time-frequency features such as mean, variance, Kurtosis, and skewness have been used for the first time in machine learning model training for suspicious transaction detection. Therefore, the feature engineering stage can be shortened by calculating the proposed time-frequency feature set. This potentially saves many person-months of modeling studies for the financial institutions. The proposed solution has been implemented in Python and the high-level of accuracy has been proven on real financial data. The generalized solution can easily be adapted to detect suspicious transactions in various organizations. An analysis of actual customer data indicates that time-frequency features can distinguish between suspicious and clear cases, improving AUC and the efficiency of the transaction monitoring system. Among different time-frequency characteristics, Kurtosis provided the maximum differentiation in the model. The gains in accuracy and the capability of detecting money laundering cases that were not detectable before can save financial institutions from regularity fines and HR cost in the order of millions of USD. In this work, only a low complexity Fourier transformbased approach is utilized for frequency domain analysis. As a future work, the time-frequency analysis can be accomplished with other types of linear and non-linear transforms. There are also potential gains in comparing multiple window lengths, increment sizes and making the analysis in multiple banking channels (such as ATM, Branch, Web). Also, the same analysis can be extended to investigate the characteristics of networks rather than single entities. In particular, when a customer has multiple accounts in multiple banks, the whole picture can only be analyzed by the FIUs. Therefore, repeating this study with additional FIU data would be beneficial as well. Hence, time-frequency features have numerous potential future uses in the area of financial behaviour analysis.

REFERENCES:

[1] T. Sausen and A. Liegel, ‘‘AI in AML: The shift is underway,’’ NICE Actimize, Hoboken, NJ, USA, Tech. Rep., Jan. 2020. [Online]. Available: https://www.niceactimize.com/Documents/aml_ai_in_aml_insights_report.pdf

[2] Estimating Illicit Financial Flows Resulting From Drug Trafficking and Other Transnational Organized Crimes, U. N. O. Drugs and Crime, Vienna, Austria, 2011.

[3] J. Gao, Z. Zhou, J. Ai, B. Xia, and S. Coggeshall, ‘‘Predicting credit card transaction fraud using machine learning algorithms,’’ J. Intell. Learn. Syst. Appl., vol. 11, no. 3, pp. 33–63, 2019.

[4] Y. Bengio, A. Courville, and P. Vincent, ‘‘Representation learning: A review and new perspectives,’’ IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[5] J. Heaton, ‘‘An empirical analysis of feature engineering for predictive modeling,’’ in Proc. SoutheastCon, Mar. 2016, pp. 1–6.

[6] A. Coates, A. Ng, and H. Lee, ‘‘An analysis of single-layer networks in unsupervised feature learning,’’ in Proc. 14th Int. Conf. Artif. Intell. Statist., JMLR Workshop Conf., 2011, pp. 215–223.

[7] R. C. Watkins, K. M. Reynolds, R. Demara, M. Georgiopoulos, A. Gonzalez, and R. Eaglin, ‘‘Tracking dirty proceeds: Exploring data mining technologies as tools to investigate money laundering,’’ Police Pract. Res., vol. 4, no. 2, pp. 163–178, Jun. 2003.

[8] T. E. Senator, H. G. Goldberg, J. Wooton, M. A. Cottini, A. U. Khan, C. D. Klinger, W. M. Llamas, M. P. Marrone, and R. W. Wong, ‘‘Financial crimes enforcement network AI system (FAIS) identifying potential money laundering from reports of large cash transactions,’’ AI Mag., vol. 16, no. 4, p. 21, 1995.

[9] J. S. Zdanowicz, ‘‘Detecting money laundering and terrorist financing via data mining,’’ Commun. ACM, vol. 47, no. 5, pp. 53–55, May 2004.

[10] T. Zhu, ‘‘An outlier detection model based on cross datasets comparison for financial surveillance,’’ in Proc. IEEE Asia–Pacific Conf. Services Comput. (APSCC), Dec. 2006, pp. 601–604.