Detection of Financial Statement Fraud Using Evolutionary Algorithms

Matthew Alden
Institute of Technology
University of Washington at Tacoma

Daniel Bryan
Milgard School of Business
University of Washington at Tacoma

Brenton Lessley
Institute of Technology
University of Washington at Tacoma

Arindam Tripathy*
Milgard School of Business
University of Washington at Tacoma

September 25, 2012

*Corresponding author

Detection of Financial Statement Fraud Using Evolutionary Algorithms

Abstract

In this paper, we use a Genetic Algorithm (GA) and MARLEDA – a modern Estimation of Distribution Algorithm (EDA) – to evolve and train several fuzzy rule-based classifiers (FRBCs) to detect patterns of financial statement fraud. We find that both GA and MARLEDA demonstrate a better ability to classify unseen corporate data observations than those of a traditional logistic regression model and provide validity for detecting financial statement fraud with Evolutionary Algorithms (EAs) and FRBCs. Using 10-fold cross-validation, the GA and MARLEDA yield average training classification accuracy rates of 75.47 percent and 74.26 percent, respectively, and average validation accuracy rates of 63.75 percent and 64.46 percent, respectively.

Keywords: evolutionary algorithm; fuzzy rule-based classifier; financial statement fraud detection; SAS No. 99

I. INTRODUCTION

Due to widespread corporate accounting fraud in the early 2000s (e.g., Enron, WorldCom, Adelphia, and RiteAid), the Accounting Standards Board (ASB), in 2002, issued *Statement on Auditing Standards* (*SAS*) *No. 99* that significantly redefined both fraud detection procedures and a set of risk factors, or "red flags", which can aid auditors in identifying patterns of fraud (Krishnan and Visvanathan 2011). In order to facilitate the usefulness and effectiveness of *SAS No. 99*, the ASB categorized each red flag based on its relation to one of the following elements of the Fraud Triangle: *pressure*, *opportunity*, and *rationalization*. Due to the potential ramifications of financial statement fraud, it is imperative that auditors and regulators develop and employ intelligent systems that can help detect patterns of fraud that are not perceived by human analyses of *SAS No. 99* red flags. These computational tools can assist auditors with the evaluation of audit risks and support stakeholders with their investment decisions. Our study demonstrates validity for detecting financial statement fraud using *evolutionary algorithms* (EAs) and *fuzzy rule-based classifiers* (FRBCs), and provides insight for auditors and stakeholders in identifying fraud-like characteristics in firms.

A large body of prior literature develops fraud detection systems by means of function-based *machine learning* models, such as logistic regression, linear discriminant analysis, and artificial neural networks. With exception to the recent application of *genetic algorithms* (GAs) by Chai et al. (2006) and Hoogs et al. (2007) that examine the time-based patterns of fraud, little research attempts to detect patterns of financial statement fraud via EAs. EAs have demonstrated success in performing optimization tasks within complex non-fraud problem domains (e.g., bankruptcy prediction and portfolio optimization) that cannot be efficiently or effectively explored by function-based models. By maintaining a flexible solution representation, EAs can be utilized to conveniently generate sets of fuzzy logic rules that

characterize patterns of fraud in a linguistic format that is difficult to emulate in traditional function-based algorithms. In accordance with the language of *SAS No. 99*, this linguistic format models the *degree* to which fraud risk factors (red flags) are present or absent within a pattern of fraud. Based on a literature review, no prior financial statement fraud studies have specifically applied EAs to the development of these fuzzy logic rule sets.

Using financial statement information, we extend prior literature by using both a standard GA and the *Markovian Learning Estimation of Distribution Algorithm* (MARLEDA) to generate FRBCs that classify whether a corporation was the recipient of an Accounting and Auditing Enforcement Release (AAER), which, similar to Skousen et al. (2009) and Dechow et al. (2011), serves as a proxy for the presence of financial statement fraud. An FRBC is represented by a set of if-then fuzzy logic rules, each consisting of a conjunction of one or more if-propositions that are used to infer a then-proposition. Each if-proposition contains the associated fuzzy set of a financial variable (e.g., "*Return on Assets* is *Very Low*"), while the then-proposition represents the predicted classification (i.e., *Fraudulent* or *Non-Fraudulent*). Through the GA and MARLEDA learning processes, a population of FRBCs is iteratively evolved and evaluated to identify the sets of fuzzy logic rules that simultaneously maximize classification accuracy and minimize rule set complexity.

Another important aspect of this study is the application of MARLEDA to the development of the FRBCs. Since MARLEDA addresses several important deficiencies that predecessor GAs and *estimation of distribution algorithms* (EDAs) are known to possess, the results of this study can help corroborate the theoretical discoveries of previous research (Alden 2007) and assess the model's practical ability to solve a difficult classification learning task.

After conducting a suite of classification experiments, the GA and MARLEDA models demonstrate success in classifying financial statement fraud with average training accuracy rates of 75.47 and 74.26 percent, respectively; and average validation accuracy rates of 63.75 and 64.46 percent, respectively.

The remainder of this article is organized as follows. Section two surveys the financial statement fraud classification literature that is relevant to this research. Section three outlines the theory of the classification models that are utilized in this study. Section four enumerates the procedures that were employed to construct and pre-process a sample of financial data, while section five outlines the training processes of the classification models. Section six discusses the results of a suite of classification experiments. Finally, section seven presents a set of concluding remarks and introduces potential avenues of future research.

II. RELATED LITERATURE

An extensive amount of prior research has been conducted to investigate the types of entities, financial variables, and other information that are indicative of fraudulent financial reporting. In this section, we briefly review the key findings of research that utilized artificial intelligence techniques to detect financial reporting fraud and/or subsequent SEC enforcement.

Function-based Models

Function-based models attempt to learn a function (or decision boundary) that partitions an attribute space into distinct classes (or regions), such that the separation between the classes is maximized and the aggregate classification error is minimized. A large body of fraud literature applies function-based machine learning models to the detection and classification of fraudulent and non-fraudulent financial reporting. Persons (1995), Summers and Sweeney (1998), and Dechow et al. (2011) identify sets of financial variables that can successfully detect financial

statement fraud using logistic regression classifiers. Additionally, during the past decade, logistic regression models have been utilized to assess the effectiveness of *SAS No. 99* red flag variables. Skousen and Wright (2006) and Skousen et al. (2009) survey extant accounting research to identify financial ratios and other publicly-available corporate information that could serve as proxies for the pressure, opportunity, and rationalization red flags of *SAS No. 99*. The logistic regression results of the latter study reveal that rapid asset growth, increased cash needs, and external financing are positively-related to the likelihood of fraud (or the issuance of an AAER). Furthermore, the cumulative percentage of outstanding common stock owned by insiders and the control expressed by the board of directors are also linked to an increased incidence of financial statement fraud.

Research utilizing more-sophisticated, non-linear function-based models has also been conducted. Green and Choi (1997), Fanning and Cogger (1998), and Feroz et al. (2000) use multi-layered artificial neural networks to identify patterns of fraudulent behavior within a dataset of AAER and non-AAER data records. Kaminski et al. (2004) conduct a linear discriminant analysis (LDA) classification experiment to provide empirical evidence that financial ratios may possess a limited ability to detect and/or predict fraudulent financial reporting. Employing a meta-learning approach, McKee (2009) combines the outputs of logistic regression, neural network, and decision tree models to predict financial statement fraud. Likewise, Ravisankar et al. (2011) use a suite of five non-linear function-based classification models (e.g., artificial neural networks and genetic programming) to identify companies that resort to financial statement fraud. Finally, Tsaih et al. (2011) utilize the Growing Hierarchical Self-Organizing Map (GHSOM) neural network architecture to implicitly conclude that

relationships can exist between common fraud techniques (exogenous variables) and financial input variables.

Evolutionary Algorithm Models

Evolutionary algorithms (EAs) are a class of stochastic search algorithms that are suitable for performing optimization and classification tasks within complex attribute spaces. One of the first applications of evolutionary algorithms to financial statement fraud classification was initiated by Hoogs et al. (2007). This study uses a genetic algorithm (GA) to detect temporal, or time-based, patterns that are indicative of financial statement fraud. The work of Chai et al. (2006) complements the genetic algorithm approach of Hoogs et al. (2007) by assigning to each learned pattern phrase a fuzzy score representing the degree to which a company's financial data matches the conditions (antecedents) of the phrase.

Within extant literature, no studies have applied estimation of distribution algorithms (EDAs) to the classification of financial statement fraud. Alden (2007) demonstrates that the Markovian Learning Estimation of Distribution Algorithm (MARLEDA) performs as well or better than predecessor GA and EDA models on a suite of optimization tasks, such as function minimization and RNA secondary structure prediction. The utilization of MARLEDA in our research will represent the first application of the model to a real-world classification task.

Fuzzy Logic and Fuzzy Set Theory Techniques

Fuzzy logic and/or fuzzy set theory techniques have been applied to a diverse range of business-related tasks, including the prediction of stock index values (Kim et al. 2004), the generation of profitable trading strategies (Ghandar et al. 2008), the measurement of management performance (Ammar et al. 2000), the identification of firms that need going concern modifications in an audit (Lenard et al. 2000), and the modeling of cost variances

(Zebda 1984).Within the financial fraud domain, Comunale et al. (2010) demonstrate one of the first applications of fuzzy rule-based classification by developing an expert system that assesses the risk of financial statement fraud. First, an auditor inputs data (e.g., a binary indicator) regarding the presence or absence of each *SAS No. 99* red flag variable. Then, the system uses the principles of fuzzy logic to both evaluate the degree to which each red flag is present and computes the fraud risk associated with the various types of financial statement fraud. The expert system in this study requires the manual specification of the fuzzy logic rules, whereas the FRBCs (rules) in our study are automatically generated via supervised learning procedures (GA and MARLEDA).

Textual Detection Techniques

Recently, Glancy and Yadav (2011) investigate the effectiveness of textual, or linguistic, variables in detecting patterns of fraud within annual 10-K reports. The authors employ clustering algorithms to accurately group textual terms (words) from the Management Discussion & Analysis section of 10-K reports into fraudulent and non-fraudulent clusters.

III. THEORY OF EVOLUTIONARY FUZZY RULE-BASED CLASSIFICATION

In this study, EA models are used to generate FRBCs that classify the fraudulent state of a corporation. The following subsections review these EA models and outline the FRBC classification technique that was employed in this study.

Evolutionary Algorithms

When a pattern classification hypothesis space becomes too complex (e.g., slope discontinuity, non-linearity, or large problem domain), traditional analytic and gradient search techniques can be ineffective and computationally slow at converging to a hypothesis (function) that partitions data patterns into distinct classes (Eiben and Smith 2003). *Evolutionary*

algorithms (EAs) are a class of stochastic search algorithms that are suitable for performing optimization tasks within complex hypothesis spaces. Figure 1 illustrates the primary steps of a typical EA model. Two of the most notable EAs are genetic algorithms (GAs) and estimation of distribution algorithms (EDAs). These algorithms are discussed in the following subsections.

Insert Figure 1 Here

Genetic Algorithms

Based on the fundamental principles of natural selection and survival of the fittest, genetic algorithms (GAs) have been successfully applied to a diverse range of optimization tasks, such as sports scheduling, traveling salesman tours, computer circuitry design, portfolio optimization, and protein secondary structure prediction (Eiben and Smith 2003; Alden 2007; Ghandar et al. 2008). A typical GA "evolves" a *population* of candidate solutions, or *chromosomes*, according to a *fitness function* that indicates the quality of a chromosome. Each chromosome consists of a set of *genes*, each representing a parameter of an optimization task. The value of each gene, known as an *allele*, represents one of the possible values for the associated task parameter.

During each algorithm iteration, or *generation*, *mutation* and *crossover* variation operators are applied to *parent* chromosomes with the intent of generating *offspring* chromosomes of higher fitness and diversity. This variation allows GAs to explore uncharted areas of the solution space and avoid becoming trapped in local minima (maxima) regions (Eiben and Smith 2003). The performance and search time of a GA largely depend on the algorithm parameters (e.g., population size, mutation and recombination rates, and chromosome representation) and their suitability for the problem domain (Eiben and Smith 2003). Without

properly tuning these parameters, the search process can become computationally expensive and fail to identify global optimum solutions, if such solutions exist (Larrañaga and Lozano 2002; Eiben and Smith 2003). An additional drawback of GAs is that they are based on the *building block* assumption that high-fitness solutions are located "near" other high-fitness solutions in the search space (Larrañaga and Lozano 2002; Eiben and Smith 2003; Alden 2007). Thus, GAs might not be as effective when this assumption is violated, such as when specific dependencies among genes dictate higher fitness values.

Estimation of Distribution Algorithms

In response to the notable limitations of GAs, *estimation of distribution algorithms* (EDAs) were introduced in the late 1990s (Larrañaga and Lozano 2002). Operating much like a GA, an EDA generates a new population of candidate solutions by statistically sampling from a gene-wise probability distribution, which is estimated from the selected solutions of the previous generation. Due to this statistical feature, an EDA is able to explicitly model the interrelations, or building blocks, among the genes of a chromosome solution. Furthermore, the sampling procedure has historically replaced the mutation and crossover operations within the evolution process.[1]

Many notable EDA implementations maintain directed, acyclic graph (DAG) structures (refer to Figure 2), which define a natural dependency ordering among genes and can be efficiently evaluated via standard graph traversal algorithms (Levitin 2007). Recent research has experimentally verified that undirected graph models may perform better than their directed counterparts on many optimization tasks (Alden 2007). Since undirected graph models lack a natural gene-dependency ordering, it is costly to learn and sample from them. To overcome this

[1] Hybrid EDAs utilize a mutation operator and have demonstrated success in conducting more-efficient searches than mutation-less EDAs, within the space of feasible chromosomes (Larrañaga and Lozano 2002; Pelikan et al. 2006; Alden 2007).

bottleneck, existing undirected EDAs have reduced the complexity of the graph models and/or partially converted them into simpler, directed structures. However, these approaches can potentially reduce the model flexibility and the effectiveness of the model learning procedure (Alden 2007).

Insert Figure 2 Here

MARLEDA

The *Markovian Learning Estimation of Distribution Algorithm* (MARLEDA) is a recently-introduced EDA that learns and samples from a *Markov random field* (MRF) probability model to address the constraints posed by undirected graph structures of predecessor EDAs (Alden 2007). Modeling discrete (nominal) genes $\{x_1, \cdots, x_n\}$ with the random variables $\{X_1, \cdots, X_n\}$, MARLEDA learns the statistical dependencies between random variable pairs by means of Pearson's chi-square nonparametric hypothesis test. If a dependency is significant, then the two genes become *neighbors*; conversely, if the two genes began as neighbors and the similarity is not significant, then they should become non-neighbors. Based on this concept of gene similarity, MARLEDA constructs a MRF *neighborhood system*. To generate new chromosomes, MARLEDA samples the MRF model via a *Markov chain Monte Carlo* process, which iteratively proceeds for a specified number of iterations or until the allele distribution of a new chromosome reaches a steady-state.

Fuzzy Rule-based Classifiers

We classify corporate data observations with *fuzzy rule-based classifiers* (FRBCs), which are generated via the GA and MARLEDA models. First conjectured in 1965, *fuzzy set theory* is

a logical system that assigns to a proposition a degree of truth in the range $[0, 1]$. Consider the proposition "*Return on Assets* is *Very Low*." The goal of fuzzy set theory is to measure the degree to which the *fuzzy variable* (or *linguistic variable*) "*Return on Assets*" is a member of the *fuzzy set* (or *linguistic value*) "*Very Low*." The *Return on Assets* fuzzy variable may possess one or more fuzzy sets, such as *Very Low*, *Low*, *Medium*, *High,* and *Very High*. A fuzzy set $\mathcal{F}$ defined on a domain of input values $\mathcal{U}$ of a fuzzy variable V is characterized by a *membership function* $\mu_{\mathcal{F}}$ that maps an input value $x \in \mathcal{U}$ to the interval $[0, 1]$, indicating x's degree of membership in $\mathcal{F}$.

An FRBC is composed of a set of *fuzzy logic rules*, or logical implications. Each rule is expressed in an if-then format, in which each if-proposition (*antecedent*) contains a logical expression and the then-proposition (*consequent*) indicates the distinct *class* that is issued upon fulfillment of the antecedent. A hypothetical fuzzy logic rule is presented in Figure 3.

Insert Figure 3 Here

More formally, in a *d*-dimensional pattern classification task with M classes, an FRBC rule, R_q, takes the following form:

$$\text{Rule } R_q\text{: } \textbf{if } x_1 \text{ is } A_{q1} \text{ and } \cdots \text{ and } x_n \text{ is } A_{qd} \textbf{ then Class } C_q \text{ with weight } CF_q, \quad (1)$$

where $\boldsymbol{x} = \langle x_1, \cdots, x_d \rangle$ is an *d*-dimensional input data pattern (observation), A_{qi} is the fuzzy set of the i^{th} fuzzy variable, $C_q \in M$ is the consequent class, and $CF_q \in [0,1]$ is a *rule* weight representing the degree of certainty of the classification C_q.

Supervised Learning (Training) and Validation

Given a *training* data matrix $\boldsymbol{D} = \{\boldsymbol{x}_p\}_{p=1}^{m}$ with m observations and a matrix $\boldsymbol{Y} = \{y_p\}_{p=1}^{m}$ with the associated class labels (*Fraudulent or Non-Fraudulent*) of the m training observations, both MARLEDA and the GA engage in *supervised learning* to evolve FRBCs that 1) model the patterns of fraud within the training data and 2) accurately predict the actual class of each observation.

Following the training phase, each FRBC is *validated* by classifying the previously-unseen observations of a validation data matrix $\boldsymbol{T}$ ($\boldsymbol{T} \neq \boldsymbol{D}$). The statistical accuracy results of this phase reveal the generalization ability of the classifiers. In this research, each FRBC is trained and validated by means of a *cross-validation* procedure that partitions a dataset of $n > m$ observations into k approximately-equal folds and performs k classification experiments (EA model runs), where in each experiment one *validation* fold, $\boldsymbol{T}$, is held-out for testing and the remaining $k - 1$ folds are used for training (i.e., the union of the training folds is $\boldsymbol{D}$). During the validation phase, only FRBCs with a training accuracy rate greater than 50 percent are used to classify the observations in $\boldsymbol{T}$.

IV. DATA AND SAMPLE SELECTION

Due to the extant corporate fraud in the early 2000s, the SEC has correspondingly increased its role in fraud detection and prevention through the issuance of Accounting and Auditing Enforcement Releases (AAERs) to corporations, employees, and auditors who knowingly or materially violate generally accepted accounting principles (GAAP). Since the SEC issues an AAER with the belief that a company engaged in fraudulent financial reporting (e.g., violations of GAAP), such a citation is highly suitable for financial fraud classification tasks (Feroz et al. 2000; Skousen and Wright 2006; Skousen et al. 2009; Dechow et al. 2011).

Given the public availability of AAER citations and prior research, AAERs were utilized in this study to serve as proxies for the presence of fraudulent financial reporting. AAERs, however, only identify the material fraud identified by the SEC and do not capture undetected fraud or fraud reported in other sources but not issued an AAER. As a result, our non-fraud sample may contain some fraud firms that negatively affect the training and validation of our models.

The complete set of financial data was collected and pre-processed as part of two primary phases. First, the citation documents of an initial set of 654 corporate AAERs, issued from 1989 to 2010, were extracted from LexisNexis and the SEC EDGAR system, and manually reviewed to acquire two pieces of information: time period of fraudulent activity and reason for the issuance of the citation.[2] Given this information, the AAER list was filtered according to the following criteria:

1. All firms with a fiscal first year of fraudulent activity from 1989 to 2010 are included in the AAER list. This rule was implemented because the effective date of *SAS No. 53* commenced in 1989 and the Compustat database provides broader access to financial data and variables post-1988.
2. Since we are using financial statement variables to detect fraud, if the reason for citation of a corporation was not directly related to any financial accounts (e.g., the failure to disclose notes to a financial statement or the use of illegal bribes and kickbacks) or not material, then the corresponding record was removed from the AAER list.
3. If an AAER citation was only related to financial statement fraud in a Form 8-K or quarterly Form 10-Q report, then the citation record was removed from the AAER list.

During these procedures, 321 corporations were removed, leaving a net total of 333 records in our list of AAER firms.

The second phase of the data collection process consisted of identifying a set of financial variables that would represent a subset of the *SAS No. 99* red flags in the fraud classification task. Using prior literature (Feroz et al. 2000; Skousen and Wright 2006; Skousen et al. 2009),

[2] While preparing the initial set of AAER records, corporations that had been issued multiple AAER citations within a fraud time period were temporarily segregated. If several citations were was related to the same financial reporting violation, then all but one citation record was removed from the AAER list.

we identify 18 financial variables that are related to the *SAS No. 99* red flags and fraud detection. Each of the 18 variables is derivable from quantitative financial statement information. Initially, several additional variables were considered that are not customarily contained in financial statements (e.g., audit committee size, executive stock ownership, percentage of outside board members, and executive turnover). However, the extraction of data for these variables, via the RiskMetrics and ExecuComp databases, would significantly reduce the sample size and limit our analysis to smaller sample with a bias in terms of larger firms with available data. Furthermore, three binary variables related to the external auditor of a firm (qualified audit opinion, change in auditor and use of a Big 4 auditor) were not used because of their heavy skewedness toward one particular value. Preliminary experiments with the GA and MARLEDA revealed that the skewed data of the binary variables was negatively impacting both the learning and classification procedures of the algorithms, and reducing the effectiveness of the constructed fuzzy sets.

From this literature we use the following variables: Change in Sales (*SCHANGE*), Return on Assets (*ROA*), Cash Flows to Earnings Growth (*CATA*), Return on Equity (*ROE*), Leverage (*LEV*), Change in Assets (*ACHANGE*), Free Cash Flow (*FREEC*), Demand for Financing (*FINANCE*), Change in Receivables (*RECEIVABLE*), Change in Inventory (*INVENTORY*), Market Domination (*MARKETDOM*), Difficult to Audit Transactions (*DIFFAUD*), Gross Profit Margin (*GPM*), Unusual Return on Equity (*UROE*), Total Accruals (*TACC*), Current Accruals (*CACC*), Inventory to Sales (*INVSAL*), and Sales to Total Assets (*SALTA*). Based on the *SAS No. 99* red flag descriptions, these variables serve as proxies for 10 red flags. Red flags that involved internal corporate data and or unquantifiable events (e.g., ineffective accounting and information systems, and frequent auditor disputes) were not represented because of their inaccessibility to the public. The reader is referred to Table 1 for definitions of the financial variable set.

Insert Table 1 Here

After compiling the set of financial variables for the fraud classification task, the Compustat financial database was utilized to extract annual, Form 10-K data for the AAER corporate records. First, all incomplete AAER data records were removed from the dataset. An incomplete data record is one in which Compustat was unable to extract non-empty values for every financial variable in the record. This procedure eliminated 104 AAER records. Secondly, to generate and obtain the non-AAER corporate data records, an algorithm was implemented to match each AAER data record with a single, non-AAER firm, based on the following three criteria: first year of fraud, industry, and beginning of the year total assets, which is a proxy for firm size.[3] This procedure was applied to each of the AAER records and yielded a final dataset size of 458 records (229 AAER records and 229 non-AAER records).

V. TRAINING FUZZY RULE-BASED CLASSIFIERS TO DETECT FRAUD

For each of the ten cross-validation runs of an EA model, the complete dataset of 458 observations is partitioned into ten folds: nine for training and one for validation. During the training phase of each run, the following steps are performed. Initially, fuzzy sets are constructed for each financial variable using the training data, and an initial set of 100 parent FRBCs (chromosomes) are created and evaluated to determine how well they classify the training observations as fraudulent or non-fraudulent. Then, for several generations (iterations), the following two operations are conducted. First, 100 offspring FRBCs are either derived from

[3] Although Compustat has over 300 annual data items available for matching, accounting research only uses a subset of these variables. Following Green and Choi (1997), Skousen and Wright (2006), Glancy and Yadav (2011), Krishnan and Visvanathan (2011), and Alam and Petruska (2012), among others, we match our sample based on three factors to control for time period, industry, and firm size, which each have been shown to affect the measurement of variables and firm oversight.

parent FRBCs (via the GA) or from a sampling procedure (via MARLEDA). Second, of the combined pool of parents and offspring, a subset of top-performing FRBCs (in terms of a fitness measure) is selected and advanced to the next generation. These two operations are repeatedly performed to improve the ability of the FRBCs to detect fraud. After the training phase is complete, the validation fold is used to test the classifiers. Figure 4 displays a schematic of the training and validation phases. We discuss the implementation of the FRBCs and the training processes in more detail below.

Insert Figure 4 Here

Chromosome Representation of FRBCs

GA and MARLEDA both evolve chromosomes (candidate solutions) in the form of FRBCs. In this research, an FRBC chromosome, S, consists of 20 fuzzy logic rules, each of the form presented in equation (1), where $\boldsymbol{x} = \langle x_1, \cdots, x_{18} \rangle$ is a corporate data observation with 18 financial variables, $A_{q\mathrm{i}} \in \{$*Very Low*, *Low, Medium, High, Very High*$\}$ is the fuzzy set of the i^{th} variable, and $C_q \in \{$*Fraudulent, Non - Fraudulent*$\}$ is the consequent class of the rule. The *Very Low* and *Very High* fuzzy sets are each characterized by a ramp membership function, while the three inner-most sets are each defined by a triangular membership function[4]. For each financial variable, V_i, these five polygonal fuzzy sets are constructed from training data by means of a custom algorithm that attempts to evenly partition the attribute space of V_i. Figure 5 illustrates the five fuzzy sets of an example financial variable.

[4] The five fuzzy sets were selected because of their ability to linguistically model financial data and sufficiently represent the distribution of attribute values. Furthermore, ramp membership functions are used to model the *Very Low* and *Very High* sets because of their ability to encapsulate extreme data observations, without necessitating additional triangular functions (e.g., *Extremely Low* or *Extremely High*), which would increase the complexity of the fuzzy logic rules and rule sets.

Insert Figure 5 Here

Figure 6 displays an FRBC in the form of a matrix; this representation embodies a chromosome within the GA and MARLEDA models. Each rule, R_q, $1 \leq q \leq m$, maintains an *Active* bit, indicating whether the rule can be used to classify any of the data patterns. Each financial variable, V_i, $1 \leq i \leq n$, is associated with a pair of values that represent, respectively, 1) the fuzzy set of which the variable is a member and 2) the *active* status of the variable.

Insert Figure 6 Here

If V_i is active, then its rightmost value is set to 1 (0, otherwise) and the variable can be used in the classification task. Additionally, each fuzzy set $s \in M = \{$*Very Low*, *Low*, *Medium*, *High*, *Very High*$\}$ is encoded with an integer $i \in \{0, 1, 2, 3, 4\}$, indicating its zero-based index, or position, in M. Note that the consequent class of each rule is not included in the chromosome representation; hence, this feature is not subject to variation in the reproduction phase of the EA models.

It should be noted that in a given rule, only an attribute value's degree of membership in one fuzzy set is utilized, despite the fact that the value can possess membership in multiple fuzzy sets, via the overlapping property of the sets. If, for example, the attribute value of an observation is a partial member of both the *Low* and *Medium* fuzzy sets, then the value could be covered by two rules, in which one rule represents the attribute with the *Low* fuzzy set and the other represents the attribute with the *Medium* fuzzy set. The dynamics of the evolutionary algorithm and supervised learning processes (see following subsection) will determine which of

the possible fuzzy sets should be represented in the rules. For instance, the *Medium* fuzzy set in the foregoing example may not be indicative of fraud when considered in combination with the other active attributes and their corresponding fuzzy sets. Thus, this fuzzy set may not be included in an active rule, despite being partially representative of an attribute value.

Supervised Learning (Training) Phase

Given the training data matrix $\boldsymbol{D}$ and actual class matrix $\boldsymbol{Y}$ of a 10-fold cross-validation run i ($1 \leq i \leq 10$), both MARLEDA and the GA engage in supervised learning each generation to evolve FRBCs with ever-increasing classification accuracy. This learning procedure is utilized by each FRBC to 1) infer the consequent class of each component fuzzy logic rule, and 2) classify each training observation as fraudulent or non-fraudulent. These two steps are discussed as follows.

Inference of Rule Consequents

Prior to entering the classification phase, each rule, R_q, of an FRBC must be processed to infer the consequent class $C_q \in \{\textit{Fraudulent, Non - Fraudulent}\}$ that possesses the maximum *confidence* with R_q's active variables, $d > 0$, and corresponding fuzzy sets, $\boldsymbol{A}_q = (A_{q1}, \cdots, A_{qd})$, where A_{qi} is the fuzzy set of the i^{th} active variable. The following measure of confidence is used to determine C_q:

$$\boldsymbol{c}(\boldsymbol{A_q} \Rightarrow \boldsymbol{C_q}) = \boldsymbol{max}\{\boldsymbol{c}(\boldsymbol{A_q} \Rightarrow \textit{Fraudulent}), \boldsymbol{c}(\boldsymbol{A_q} \Rightarrow \textit{Non-Fraudulent})\}, \quad (2)$$

in which $c(\boldsymbol{A}_q \Rightarrow Class\ q) = \sum_{y_p = Class\ q} \mu_{\boldsymbol{A}_q}(\boldsymbol{x}_p) / \sum_{p=1}^{m} \mu_{\boldsymbol{A}_q}(\boldsymbol{x}_p)$ for $\boldsymbol{x}_p \in \boldsymbol{D}$ and $y_p \in \boldsymbol{Y}$; and

$$\boldsymbol{\mu_{A_q}(x_p)} = \left[\boldsymbol{\mu_{A_{q1}}(x_{p1})} + \cdots + \boldsymbol{\mu_{A_{qd}}}\right] / \boldsymbol{d} \quad (3)$$

is a weighted *compatibility* grade, where $\mu_{A_{qi}}(\cdot)$ is the membership function of the fuzzy set A_{qi}. The class with the maximum compatibility with $\boldsymbol{A}_q$ becomes the consequent class of rule R_q. If

$c(\boldsymbol{A}_q \Rightarrow \textit{Fraudulent})$ equals $c(\boldsymbol{A}_q \Rightarrow \textit{Non-Fraudulent})$, then R_q is set to *inactive* (i.e., C_q becomes irrelevant). Ishibuchi et al. (1996), Ishibuchi and Yamamoto (2004), Ishibuchi and Nojima (2005), and Alcalá et al. (2008) detail variants of this method.

Rule Classification

During each iteration of MARLEDA and the GA, and for each FRBC chromosome**,** S, a *single-winner* approach is employed to classify each training observation $\boldsymbol{x}_p \in \boldsymbol{D}$. In this procedure, $\boldsymbol{x}_p$ is classified with the active rule $R_{q*} \in S$ that has the maximum product of the compatibility grade in equation (3) and the rule weight CF_q:

$$\boldsymbol{max}_{\boldsymbol{R_q} \in \boldsymbol{S}} \left\{ \boldsymbol{\mu}_{\boldsymbol{A_q}}(\boldsymbol{x_p}) \cdot \boldsymbol{CF_q} \right\}, \tag{4}$$

where $CF_q \in [0,1]$ represents the degree of certainty of the classification (consequent class) C_q[5]. If a winning rule is identified, then C_q is used to classify $\boldsymbol{x}_p$ on behalf of S. Otherwise, $\boldsymbol{x}_p$ is considered to be unclassifiable and omitted from accuracy calculations. Finally, CF_q was held constant at 1, since the weight factor did not significantly influence the selection of the winner rule throughout multiple runs of the EA models. The reader is referred to (Ishibuchi and Yamamoto 2004; Ishibuchi and Nojima 2005; Ishibuchi et al. 2008) for a review of other rule classification schemes.

[5] Traditionally, the single-winner approach utilizes the multiplicative compatibility grade $\mu_{\boldsymbol{A}_q}(\boldsymbol{x}_p) = \prod_{i=1}^{d} \mu_{\boldsymbol{A}_{qi}}(\boldsymbol{x}_{pi})$. However, in preliminary experiments, this approach consistently prompted degenerate and less-accurate FRBCs, since a single membership function value of zero makes $\mu_{\boldsymbol{A}_q}(\boldsymbol{x}_p) = 0$ and prevents rule R_q from becoming the single-winner of observation $\boldsymbol{x}_p$. By summing the membership function values, each rule is afforded a greater chance to achieve a positive compatibility grade, despite the presence of one or more zero-valued membership function terms. This sum is then divided by the number of active variables in R_q to prevent more-complex rules from unfairly achieving a higher compatibility grade than less-complex rules, or those with fewer active variables.

Following the rule classification phase, the fitness of each FRBC is measured along two objectives: classification accuracy rate, *Accuracy*, and ratio of active fuzzy logic rules, *ActiveRules*, where

$$Accuracy = \frac{\text{\# of correctly classified patterns}}{\text{\# of total patterns - \# of unclassified patterns}} \tag{5}$$

and

$$\boldsymbol{ActiveRules} = \frac{\text{\# of active rules}}{\text{\# of total rules}}. \tag{6}$$

The goal of the evolutionary process is to evolve chromosomes that maximize *Accuracy* and minimize *ActiveRules*. These two objectives are linearly-combined into a weighted fitness function

$$fitness = \omega_1 \cdot Accuracy - \omega_2 \cdot ActiveRules, \tag{7}$$

where ω_1 and ω_2 are arbitrarily-assigned weights from the interval $[0,1]$. Note that the second term of fitness is subtracted from the first term to simulate the minimization of *ActiveRules*. Hence, the evolutionary algorithms attempt to maximize *fitness*. To maintain accurate FRBCs in the population and prevent negative fitness scores, ω_1 and ω_2 were initialized to 0.99 and 0.01, respectively, for the classification experiments[6].

VI. RESULTS AND DISCUSSION

Both the GA and MARLEDA attempt to evolve accurate and comprehensible FRBCs by combining the *Accuracy* and *ActiveRules* objectives into a single, weighted fitness value. Using 10 fold cross-validation, each model is run for a specific number of generations and maintains

[6] Different combinations of weight values for ω_1 and ω_2 were also considered, such as (1.00, 0.00), (0.95, 0.05), and (0.90, 0.10). The weight pair (1.00, 0.00) yielded similar fitness results as that of the implemented pair (0.99, 0.01). However, the latter two weight pairs compromised a significant amount of classification accuracy by overemphasizing the deactivation of logic rules.

parameter settings that are tuned during preliminary trial runs. The most-notable technical specifications of the GA and MARLEDA models are as follows:

- **GA**: generations = 4000, population size = 100, population selection ratio = 0.75, uniform crossover, mutation rate = 0.05, crossover rate = 0.15, and tournament size = 3.
- **MARLEDA**: generations = 4000, population size = 100, population selection ratio = 0.75, tournament size = 3, mutation rate = 0.01, and Markov chain Monte Carlo sampling iterations = 7400.

In addition to *Accuracy* and *ActiveRules*, the performance of each EA model is assessed by the following measures: *ActiveVars*, *Sensitivity*, *Specificity*, *Precision*, and *Recall*. *ActiveVars* represents the average number of active variables per active rule in an FRBC, and is computed by dividing the total number of active variables across all active rules of an FRBC by the total number of active rules. *Sensitivity* indicates the percentage of actual fraudulent observations that were correctly classified as fraudulent by an FRBC. *Specificity* measures the proportion of actual non-fraudulent observations that were correctly classified as non-fraudulent by an FRBC. *Precision* represents the percentage of all fraudulent-classified observations that were correctly identified. *Recall* specifies the proportion of all non-fraudulent classifications that were correctly made.

Furthermore, each of the *Accuracy*, *ActiveRules*, and *ActiveVars* measures in this research are calculated with respect to the corresponding values of the most-accurate validation FRBC from each cross-validation run (ten FRBCs in total, one from each validation run). The *Sensitivity*, *Specificity*, *Precision*, and *Recall* measures each assess the collective validation performance of these ten most-accurate FRBCs. Accordingly, the reported classifications of the first validation dataset were made by the most-accurate FRBC of the first dataset, the classifications of the second validation run were made by the most-accurate FRBC of the second

dataset, etc. The sum of all the reported validation classifications – 458 in total – is used to compute the latter four measures.

Experimental Results

Table 2 shows the primary training and validation classification statistics for the GA and MARLEDA in terms of the *Accuracy*, *ActiveRules*, and *ActiveVars* performance measures. Over the ten cross-validation iterations, both EA models demonstrate an ability to detect patterns of fraud. The GA yields an average training accuracy rate of 75.47 percent and an average validation accuracy rate of 63.75 percent, with 292 of 458 data observations correctly classified and a maximum validation accuracy of 75.56 percent for classification of the eighth validation dataset. Demonstrating a similar potential of detecting fraud, MARLEDA yields an average training accuracy of 74.26 percent and an average validation accuracy of 64.46 percent, with 295 of 458 data observations correctly classified and a maximum validation accuracy of 73.33 percent for the classification of the ninth validation dataset. Many financial statement fraud and misstatement studies, including Skousen et al. (2009) and Dechow et al. (2011), use logistic regression models. In comparison, using logistic regression with our data set and financial variables results in an accuracy rate of 58.30 percent.

Insert Table 2 Here

Table 3 presents supplemental validation statistics for the GA and MARLEDA. For the *Sensitivity* measure, the GA and MARLEDA correctly classify the actual fraudulent observations 66.38 and 68.12 percent of the time, respectively. The *Specificity* measure reveals that the GA and MARLEDA correctly classify the actual non-fraudulent observations 61.14 and 60.70 percent of the time, respectively.

Insert Table 3 Here

Figure 7 shows the progression of *Accuracy* while Figure 8 reveals the progression of the *ActiveRules* and *ActiveVars* ratios, as the GA and MARLEDA both attempt to extract as much accuracy as possible from a small set of active rules and variables. While increasing the average training *Accuracy* (over the ten most-accurate training FRBCs, one from each validation run) by approximately 18 percentage points from the 25th generation to the 4000th generation, MARLEDA also simultaneously reduced *ActiveRules* by 20 percentage points and *ActiveVars* by approximately two percentage point during the same time period. The GA exhibited a similar phenomenon, but was able to reduce the average *ActiveRules* ratio even further to 63.5 percent, or a decrease of 29 percentage points. These reductions in classifier complexity were largely influenced by both the 0.01 weight factor that was assigned to the *ActiveRules* term in the fitness function and the weighting factor that was applied to the compatibility score of each rule in the classification phase. Since the weight of the *Accuracy* term is set at a disproportionately high value of 0.99 in the fitness function, the evolutionary processes naturally favor more-accurate classifiers and, thus, a point is reached at which fewer rules can be deactivated without sacrificing significant classification accuracy. Hence, for new FRBCs to increase their fitness and accuracy, only minor variations in rule set complexity are instigated (i.e., *ActiveRules* and *ActiveVars* each tend to oscillate about a plateau value). This effect is most evident in the final 2,000 generations of the MARLEDA model.

Insert Figures 7 and 8 Here

Among the ten GA validation runs, the first run yields a thirteen-rule classifier that possesses the maximum validation accuracy rate of 75.56 percent. Likewise, during the fourth validation run of the MARLEDA experiment, a fourteen-rule classifier was evolved that yields a maximum validation accuracy rate of 73.33 percent. The most-accurate active rule from each of these FRBCs is displayed in a human-readable format in Figure 9. An analysis of the two FRBCs reveals that MARLEDA evolved a more-complex rule set than the GA, with a larger number of active rules and multiple conjunctions of fuzzy propositions (active variables). This observation is corroborated by the experimental training results, which indicate that, compared to the GA, MARLEDA possesses a greater *ActiveRules* ratio and similar *ActiveVars* ratio, on average.

Furthermore, a review of the ten most-fit validation FRBCs (one from each validation run) reveals that, for both the GA and MARLEDA, the *TACC* and *SCHANGE* variables are frequently activated in accurate rule classifications (true positives and true negatives). Additionally, *CACC*, *FREEC*, and *ROE* are highly-prevalent within successful MARLEDA rules, while *LEV*, *RECEIVABLE*, and *UROE* commonly appear in accurate GA rules. Interestingly, these same variables are also the most-prevalent variables within inaccurate rule classifications (false positives and false negatives) of both models. This finding suggests that several validation data observations may be either mislabeled or possess highly-similar characteristics (patterns) with the opposite class, which can hinder the supervised learning procedure and depress the overall classification accuracy rate.

Insert Figure 9 Here

Sensitivity Analysis

The results from Table 3 suggest that the GA and MARLEDA are both more successful at classifying fraudulent firms (*Sensitivity*) than non-fraudulent firms. Prior literature suggests that some distressed firms attempt to avoid default or bankruptcy by using more aggressive accounting policies or estimates. These firms' financial data may appear similar to fraudulent firms, making it difficult to both train the FRBCs and distinguish non-fraudulent firms from the fraudulent firms, resulting in a lower *Sensitivity* and *Specificity*. Accordingly, as a sensitivity test, we rerun our EA models after removing nine pairs of firms, where each pair consists of a non-fraudulent firm that subsequently filed for bankruptcy and its matching fraudulent firm that was cited with an AAER. Overall, both the GA model and MARLEDA improved their performance, specifically in the ability to correctly classify non-fraudulent firms (*Specificity*), by which the accuracy increased to 65.45 percent and 68.64 percent for GA and MARLEDA, respectively. While MARLEDA's ability to classify fraudulent firms (*Sensitivity*) remained stationary, the *Sensitivity* of the GA increased to 71.36 percent.

VII. CONCLUDING DISCUSSION AND FUTURE RESEARCH

Throughout our modern economic history, the need to detect fraudulent behavior within a corporation has captured the attention of regulators, auditors, and investors alike. Fostering this widespread interest has been the development of machine learning models that computationally discover patterns of fraud that humans would be unable to manually identify. This study extends prior financial statement fraud detection research by generating FRBCs via two EA models: the MARLEDA and a standard GA.

Both our EA models demonstrate an ability to classify unseen corporate data observations in the ten validation datasets, with MARLEDA producing average and maximum validation

accuracy rates of 64.46 percent and 73.33 percent, respectively; and the GA yielding average and maximum validation accuracy rates of 63.75 percent and 75.56 percent, respectively. These results are better than those of a traditional logistic regression model for our financial variables and observations. Additionally, as part of a sensitivity analysis, we demonstrate that the removal of both non-fraudulent firms that subsequently filed for bankruptcy protection and their matching fraudulent firms increases the average validation classification accuracy rates for both the GA and MARLEDA.

These results suggest that there is validity for detecting financial statement fraud using EAs, such as the GA and MARLEDA. Our analysis provides insight for auditors and stakeholders in identifying fraud-like characteristics in firms. This will assist auditors with evaluating risks in audits of firms and help stakeholders identify risk when forming their investment decisions.

Our analysis also indicates that both of the EA models perform better at correctly classifying fraudulent firms (*Sensitivity*), as compared to classifying non-fraudulent firms (*Specificity*). This would result in the auditors evaluating a higher risk (compared to the actual) for some non-fraudulent firms; however, it has no impact on the effectiveness of the audit. This higher risk can lead to additional audit costs from performing unnecessary audit procedures, but can prevent litigation costs when the fraud is properly detected. Accordingly, the auditor should perform a cost-benefit analysis to determine the monetary threshold at which it is more beneficial to conduct potentially-unnecessary investigations (i.e., absorb the false positive errors) than to pay the legal costs of not detecting or reporting a fraudulent client (i.e., avoid the false negative errors). For a more detailed discussion, see Persons (1995) and Dechow et al. (2011).

Future Research

Our study uses a 50/50 distribution of fraudulent and non-fraudulent firms that may not reflect the real world distribution of fraudulent firms. This one-to-one matched dataset of fraudulent and non-fraudulent firms can exert both positive and negative influences on the classification performance and supervised learning phase. As with prior studies, we have the same number of fraudulent and non-fraudulent firms (and a similar number of fraud and non-fraud FRBCs in our models); and with randomly assigning of classification (or FRBCs), one can achieve 50 percent accuracy. However, with a one-to-one matched sample, the GA and MARLEDA can learn to generalize patterns of fraud from a balanced mix of examples, without bias toward any one particular class, such as non-fraudulent firms. This condition may improve the ability of the models to evolve FRBCs that accurately discriminate the fraudulent firms from the non-fraudulent firms. However, if one of the classes contains several firms that are either mislabeled or possess characteristics of the opposite class, then the classification performance can decline, especially with respect to false positives or false negatives.

With a more realistic distribution containing a disproportionate number of non-fraudulent companies (e.g., 5 fraudulent firms to 95 non–fraudulent firms), it would seem plausible that most of the rule antecedents in an FRBC would be mathematically more-compatible with the *Non-Fraudulent* class. For a rule to be assigned to the *Fraudulent* class, the small number of fraudulent companies would have to be highly compatible with the antecedent of the rule (i.e., possess a higher aggregate compatibility score than the non-fraudulent companies). This rule would then likely make accurate *Fraudulent* classifications whenever it is presented with a true fraudulent observation, while also yielding a low number of false-positives. However, because there will likely be fewer *Fraudulent* rules, a true fraudulent observation may possess a lower

aggregate compatibility score with the *Fraudulent* class, leading to the observation's misclassification as *Non-Fraudulent* (false negative).

To address this problem, the inference calculations could be altered so that each rule is assigned the consequent class that possesses the maximum average compatibility with the rule. This new measure is more-invariant to class distribution than equation (2) and engenders a more-balanced mixture of both *Fraudulent* and *Non-Fraudulent* rules. It could be further enhanced by utilizing a weighted-sum fitness function that allows the auditor to assign weights (e.g., 50 percent each) for the *Sensitivity* and *Specificity* measures based on the relative costs of misclassifying actual fraudulent and non-fraudulent clients. So, if it is more costly to misclassify an actual fraudulent client as non-fraudulent (false negative), then a disproportionate weight should be assigned to *Sensitivity*. This will allow the auditor to trade off the accuracy of predicting fraud with false positives. While this is beyond the scope of the current study, it could be interesting for future research.

Another insightful avenue of future research would be to compare the classification performance of the GA and MARLEDA to that of several other non-linear-regression machine learning models from related literature, using our dataset as the input for each model. Fanning and Cogger (1998) use neural networks to develop a model for detecting management fraud with a data sample size of 102 SEC enforcement releases and a matched sample. They utilize a combination of financial statement and corporate governance variables, and find their model to be better at detecting fraud (63 percent accuracy) than logistic regression (50 percent accuracy). Feroz et al. (2000) further illustrate the application of neural networks by testing the ability of a few SAS No. 53 red flags to predict SEC investigation. However, the authors use a very limited sample size of 42 fraudulent firms (and 90 non-fraudulent firms) with seven variables and obtain

an 81 percent average accuracy rate at detecting fraud, compared to an average of 70 percent for logistic regression. Green and Choi (1997) also use neural networks with a limited sample of 86 fraudulent firms (and 86 non-fraudulent firms) and eight variables. They discuss their findings in terms of classifications being better than the accuracy of a fair coin toss. Ravisankar et al. (2011) use a dataset of 101 fraudulent (and 101 non-fraudulent) Chinese companies and test the effectiveness of several data mining techniques at detecting financial statement fraud using 18 variables. They find their models to be substantially superior to logistic regression in the context of their data. Finally, McKee (2009) use a meta-learning approach that incorporates a three-layer stacked generalization model of neural network, logistic regression, and classification tree algorithms to predict financial fraud. The authors use 50 fraudulent firms (and 50 non-fraudulent firms) with 15 variables from 1995 to 2002 and find that the classification accuracy increased from 71 percent to 83 percent for the neural network by stacking the results first into a logistic regression model and then into a classification tree model.

As discussed above, several prior studies use alternative non-logistic-regression models; however, a direct comparison with these studies will be inaccurate due to significant differences between the nature, composition, and size of the datasets. Some of the studies utilized datasets of only Asian-oriented corporations, while others trained their machine-learning models with lower-dimensional datasets, in terms of the number of corporations and/or input attributes. These less-diverse datasets can hide real-world "noise" that appears in larger, more-diverse datasets and, thus, may yield artificially high training and test accuracy rates, which prevent meaningful comparisons with our results. Given the foregoing reasons, we restricted our comparative analysis to the two evolutionary algorithm models and the standard logistic

regression model, each of which were tested with our dataset and 18 variables. Future research could look at using the same dataset to enable a direct comparison between the models.

Finally, we acknowledge that, as is the case of any artificial intelligence system, our algorithms are limited in their ability to learn and observe. To that extent, there is an opportunity to modify and enhance our algorithms to become more effective. Future research and modifications could focus on applying heuristics, such as adaptable crossover and mutation, in the later stages of the training to spur evolutionary change and improve the effectiveness of the algorithms. These dynamic variation operators could be modified depending on whether the fitness of solutions is stagnant or monotonically increasing. Furthermore, the EA models could be seeded with an initial set of hand-crafted FRBCs that are designed from human knowledge of financial statement fraud patterns. This knowledge could partially be acquired by manually studying the attribute values for a large subset of the financial dataset, and documenting "rules of thumb" that are prevalent for cited and non-cited corporations.

VIII. REFERENCES

Alam, P. and K. A. Petruska. 2012. Conservatism, SEC investigation, and fraud. *Journal of Accounting & Public Policy 31*(4): 399-431.

Alcalá, R., J. Alcalá-Fdez, M. J. Gacto, and F. Herrera. 2008. On the usefulness of MOEAs for getting compact FRBSs under parameter tuning and rule selection. In A. Ghosh, S. Dehuri, and S. Ghosh (Eds.), *Multi-Objective Evolutionary Algorithms for Knowledge Discovery from Databases*, 91–107. Berlin, Germany: Springer-Verlag.

Alden, M. 2007. *MARLEDA: Effective distribution estimation through markov random fields*. Ph.D. thesis, The University of Texas at Austin, Austin, Texas.

Ammar, S., R. Wright, and S. Selden. 2000. Ranking state financial management: A multilevel fuzzy rule-based system. *Decision Sciences* 31 (2): 449-481.

Chai, W., B. K. Hoogs, and B. T. Verschueren. 2006. Fuzzy ranking of financial statements for fraud detection. In *2006 IEEE International Conference on Fuzzy Systems*, 152–158.

Comunale, C. L., R. L. Rosner, and T. R. Sexton. 2010. The auditor's assessment of fraud risk: A fuzzy logic approach. *Journal of Forensic & Investigative Accounting* 3 (1): 149–194.

Dechow, P. M., W. Ge, C. R. Larson, and R. G. Sloan. 2011. Predicting material accounting misstatements. *Contemporary Accounting Research* 28 (1): 17–82.

Eiben, A. E. and J. E. Smith. 2003. *Introduction to Evolutionary Computing*. Berlin, Germany: Springer-Verlag.

Fanning, K. M. and K. O. Cogger. 1998. Neural network detection of management fraud using published financial data. *International Journal of Intelligent Systems in Accounting, Finance & Management* 7: 21–41.

Feroz, E. H., T. M. Kwon, K. J. Park, and V. Pastena. 2000. The efficacy of red flags in predicting the SEC's targets: An artificial neural networks approach. *International Journal of Intelligent Systems in Accounting, Finance & Management* 9 (3): 145–157.

Ghandar, A., Z. Michalewicz, M. Schmidt, T. Tô, and R. Zurbrugg. 2008. Evolving trading rules. In A. Yang, Y. Shan, and L. Bui (Eds.), *Success in Evolutionary Computation*, Volume 92 of *Studies in Computational Intelligence*, 95–119. Berlin, Germany: Springer Berlin / Heidelberg.

Glancy, F. H. and S. B. Yadav. 2011. A computational model for financial reporting fraud detection. *Decision Support Systems* 50: 595–601.

Green, B. P. and J. H. Choi. 1997. Assessing the risk of management fraud through neural network technology. *Auditing: A Journal of Practice & Theory* 16 (1): 14–28.

Hoogs, B., T. Kiehl, C. Lacomb, and D. Senturk. 2007. A genetic algorithm approach to detecting temporal patterns indicative of financial statement fraud: Research Articles. *International Journal of Intelligent Systems in Accounting, Finance & Management* 15: 41–56.

Ishibuchi, H., I. Kuwajima, and Y. Nojima. 2008. Evolutionary multi-objective rule selection for classification rule mining. In A. Ghosh, S. Dehuri, and S. Ghosh (Eds.), *Multi-Objective Evolutionary Algorithms for Knowledge Discovery from Databases*, Volume 98 of *Studies in Computational Intelligence*, 47–70. Berlin, Germany: Springer Berlin / Heidelberg.

Ishibuchi, H., T. Murata, and H. Tanaka. 1996. Construction of fuzzy classification systems with linguistic if-then rules using genetic algorithms. In S. K. Pal and P. P. Wang (Eds.), *Genetic Algorithms for Pattern Recognition*, 227–251. Boca Raton, FL: CRC Press.

Ishibuchi, H. and Y. Nojima. 2005. Multiobjective formulations of fuzzy rule-based classification system design. In *Joint 4th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT 2005) and the 11th Rencontres Francophones sur la Logique Floue et ses Applications (LFA 2005)*, 285–290. Boca Raton, FL: CRC Press.

Ishibuchi, H. and T. Yamamoto. 2004. Fuzzy rule selection by multi-objective genetic local search algorithms and rule evaluation measures in data mining. *Fuzzy Sets and Systems* 141 (1): 59 – 88.

Kaminski, K., T. Wetzel, and L. Guan. 2004. Can financial ratios detect fraudulent financial reporting? *Managerial Auditing Journal* 19 (1): 15–28.

Kim, M. J., I. Han, and K. C. Lee. 2004. Hybrid knowledge integration using the fuzzy genetic algorithm: Prediction of the Korea Stock Price Index. *International Journal of Intelligent Systems in Accounting, Finance and Management* 12 (1): 43-60.

Krishnan, G. V. and G. Visvanathan. 2011. Is there an association between earnings management and auditor-provided tax services?. *Journal of the American Taxation Association 33*(2): 111-135.

Landry, Jr., R. M., P. Lin, and G. D. Moyes. 2005. Raise the red flag: A recent study examines which SAS No. 99 indicators are more effective in detecting fraudulent financial reporting. *Internal Auditor* 62 (5): 47–51.

Larrañaga, P. and J. A. Lozano. (Eds.) 2002. *Estimation of Distribution Algorithms*. Boston, MA: Kluwer Academic Publishers.

Lenard, M. J., P. Alam, and D. Booth. 2000. An analysis of fuzzy clustering and a hybrid model for the auditor's going concern assessment. *Decision Sciences* 31 (4): 861-884.

Levitin, A. 2007. *Introduction to the Design and Analysis of Algorithms* (Second ed.). Upper Saddle River, NJ: Pearson Education, Inc.

McKee, T. E. 2009. A meta-learning approach to predicting financial statement fraud. *Journal of Emerging Technologies in Accounting* 6: 5–26.

Pelikan, M., A. Hartmann, and K. Sastry. 2006. Hierarchical BOA, cluster exact approximation, and Ising spin glasses. Technical Report 2006002, Missouri Estimation of Distribution Algorithms Laboratory (MEDAL) - University of Missouri in St. Louis, St. Louis, MO.

Persons, O. 1995. Using financial statement data to identify factors associated with fraudulent financial reporting. *Journal of Applied Business Research* 11 (3): 38–46.

Ravisankar, P., V. Ravi, G. R. Rao, and I. Bose. 2011. Detection of financial statement fraud and feature selection using data mining techniques. *Decision Support Systems* 50: 491–500.

Skousen, C. J., K. R. Smith, and C. J. Wright. 2009. Detecting and predicting financial statement fraud: The effectiveness of the fraud triangle and sas no. 99. In M. Hirschey, K. John, and A. K. Makhija (Eds.), *Corporate Governance and Firm Performance (Advances in Financial Economics)*, 53–81. New York, NY: Emerald Group Publishing Limited.

Skousen, C. J. and C. J. Wright. 2006. Contemporaneous risk factors and the prediction of financial statement fraud. *Journal of Forensic Accounting* 9 (1): 37–62.

Summers, S. L. and J. T. Sweeney. 1998. Fraudulently misstated financial statements and insider trading: An empirical analysis. *The Accounting Review* 73 (1): 131-146.

Tsaih, R., W. Lin, and S. Huang. 2011. The exogenous issue of feature extraction. Working paper that was presented at the 2011 American Accounting Association Annual Meeting.

Zebda, A. 1984. The investigation of cost variances: A fuzzy set theory approach. *Decision Sciences* 15 (3): 359-388.

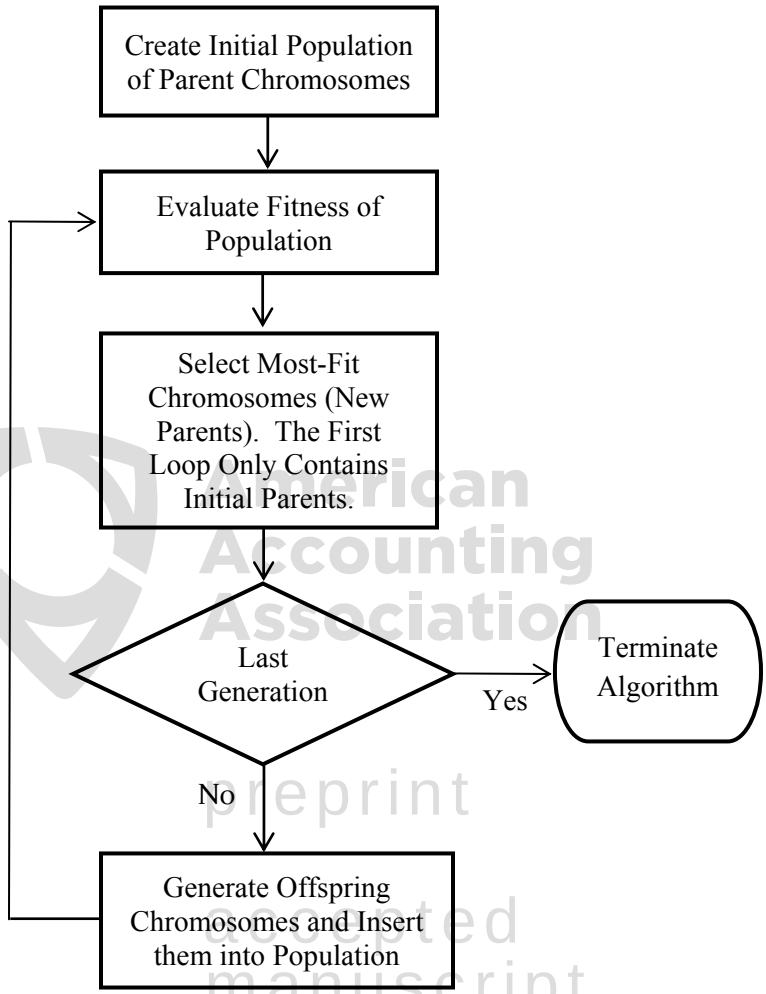


Figure 1: The primary steps of an Evolutionary Algorithm (EA).

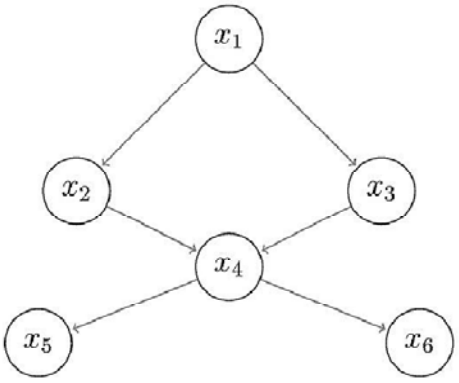


Figure 2: A multivariate gene-dependency graph of an EDA, applied to a hypothetical optimization task with six parameters. Each numerically-labeled gene x_i represents the i^{th} task parameter and is connected to one or more parent and child genes.

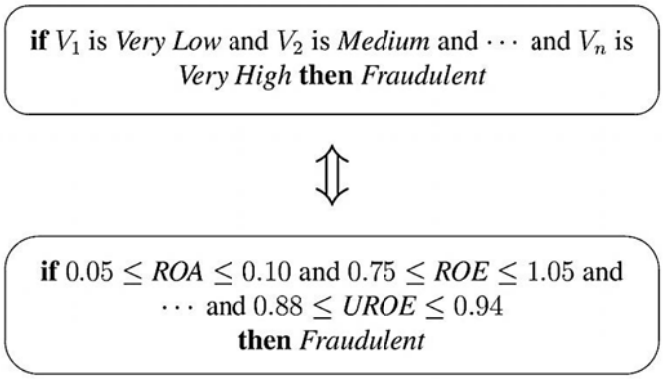


Figure 3: Example of an FRBC rule in two equivalent formats. The second format is obtained by mapping each fuzzy set from the first format to its hypothetical real-value interval (defined by minimum and maximum values), and replacing each variable with its actual name.

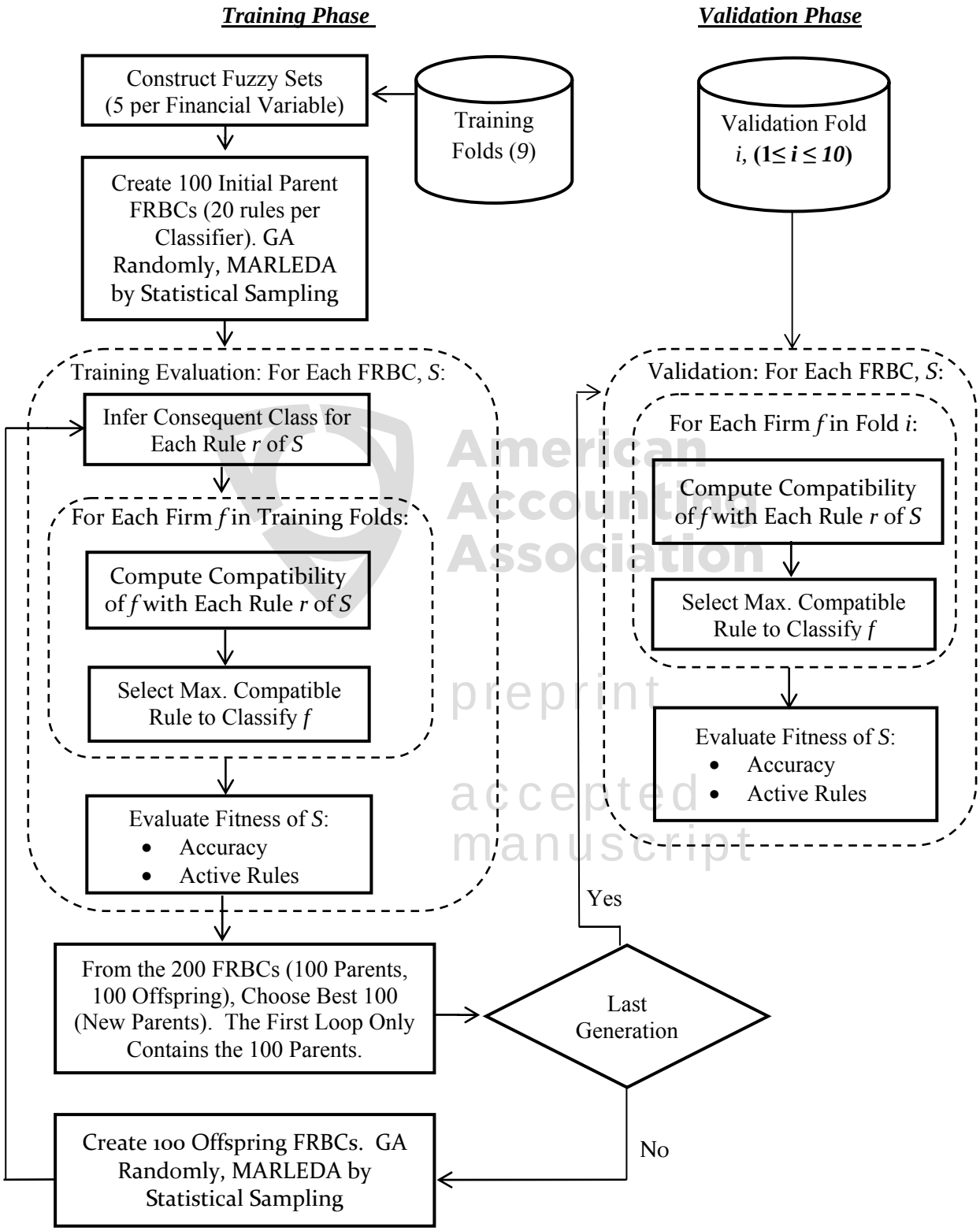


Figure 4: Training and validation phases of the GA and MARLEDA for the i^{th} cross-validation run.

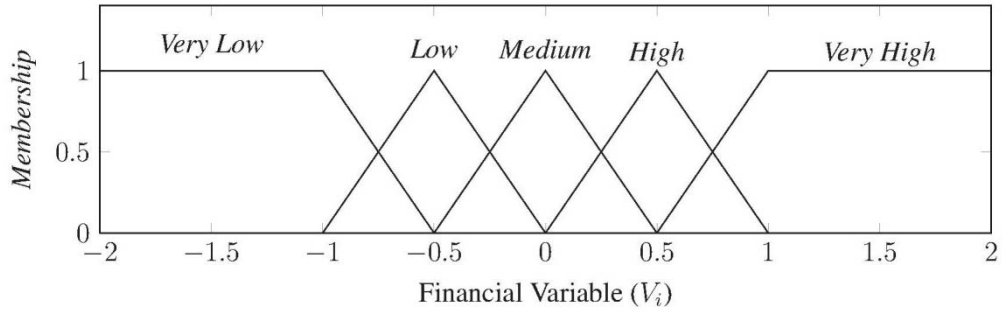
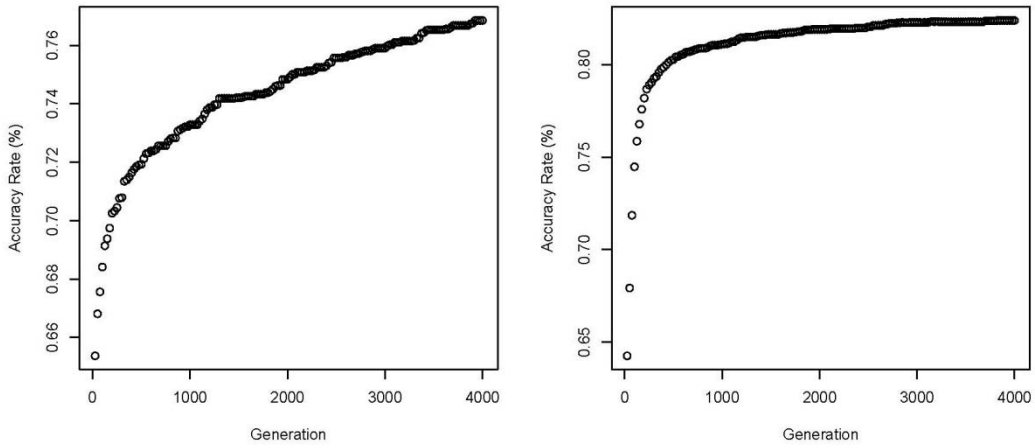


Figure 5: Illustration of the membership functions for the five fuzzy sets of a financial variable V_i, $1 \leq i \leq 18$.

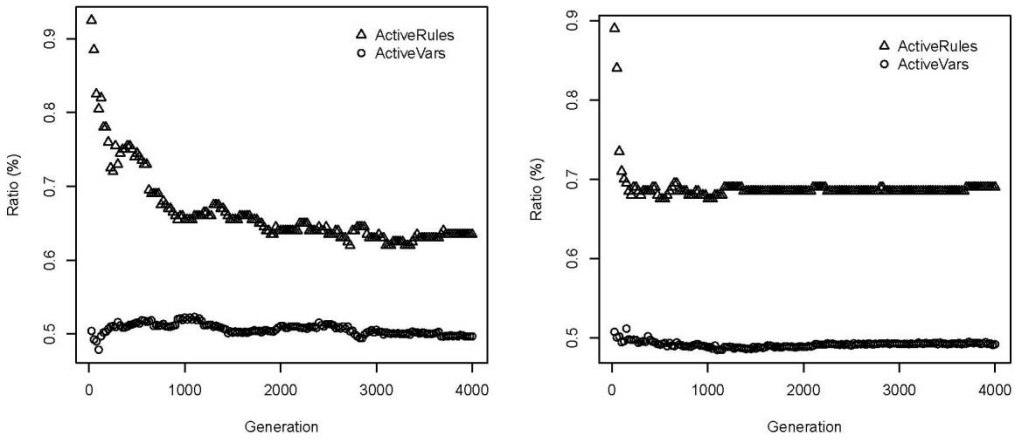
	Active	V_1		V_2		$\cdots$	V_n	
R_1	1	4	0	1	1	$\cdots$	0	1
R_2	0	0	1	2	1	$\cdots$	4	1
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
R_m	1	1	0	3	0	$\cdots$	3	0

Figure 6: Example of the internal representation of an FRBC rule set.



(a) Genetic Algorithm (b) MARLEDA

Figure 7: Scatter plots of the average generational *Accuracy* rate for the ten most-accurate training FRBCs (one from each validation run) that were evolved GA and MARLEDA.



(a) Genetic Algorithm (b) MARLEDA

Figure 8: Scatter plots of the average generational *ActiveRules* and *ActiveVars* ratios for the ten most-accurate training FRBCs (one from each validation run) that were evolved by the GA and MARLEDA.

Active Rule #12:	83.72% accuracy over 43 classifications
if:	$0.108 \leq$ *ROA* ≤ 1.687 (*Very High*) and
	$0.677 \leq$ *SALTA* ≤ 1.202 (*Medium*) and
	$0 \leq$ *INVSAL* ≤ 0.0364 (*Very Low*) and
	$-1.115 \leq$ *ROE* ≤ 0.028 (*Low*) and
	$-0.170 \leq$ *ACHANGE* ≤ 0.080 (*Low*) and
	$-0.005 \leq$ *TACC* ≤ 0.232 (*High*) and
	$0.236 \leq$ *LEV* ≤ 0.518 (*Very High*) and
	$-0.121 \leq$ *CATA* ≤ 0.027 (*Low*) and
	$2.214 \times 10^{-5} \leq$ *MARKETDOM* $\leq 4.116 \times 10^{-4}$ (*Low*)
then:	*Fraudulent*

(a) GA

Active Rule #4:	100% accuracy over 13 classifications
if:	$0.114 \leq$ *ROE* ≤ 0.406 (*High*) and
	$0.854 \leq$ *ACHANGE* $\leq 1{,}701$ (*Very High*) and
	$-0.249 \leq$ *TACC* ≤ -0.052 (*Low*) and
	$-673.750 \leq$ *CACC* ≤ -0.062 (*Very Low*) and
	$0.062 \leq$ *FREEC* ≤ 0.377 (*Very High*) and
	$0.012 \leq$ *CATA* ≤ 0.080 (*Medium*) and
	$-0.014 \leq$ *RECEIVABLE* ≤ 0.017 (*Medium*) and
	$-0.015 \leq$ *INVENTORY* ≤ 0.003 (*Medium*) and
	$-9{,}366.080 \leq$ *SCHANGE* ≤ -67.981 (*Very Low*) and
	$2.040 \times 10^{-5} \leq$ *MARKETDOM* $\leq 4.775 \times 10^{-4}$ (*Low*) and
	$3.047 \leq$ *UROE* $\leq 11{,}600.100$ (*Very High*)
then:	*Non-Fraudulent*

(b) MARLEDA

Figure 9: The most-accurate rules from the most-accurate validation FRBCs that were evolved by the GA and MARLEDA during the first and fourth cross-validation runs, respectively.

Table 1: Financial variable definitions. The assignment of the financial variables and *SAS No. 99* red flags into Fraud Triangle categories was based on the work of Feroz et al. (2000) and Skousen et al. (2009).

Variable	Definition	Red Flag(s) (Fraud Triangle Category)	Significant *p*-value in Literature
SCHANGE	Change in Sales	Rapid growth or unusual profitability (Pressure)	--
ROA	Return on Assets	High degree of competition or declining profit margins	Summers and Sweeney (1998)
CATA	Cash Flows to Earnings Growth	Operating losses (Pressure)	--
ROE	Return on Equity	High degree of competition or declining profit margins (Pressure)	--
GPM	Gross Profit Margin	High degree of competition or declining profit margins (Pressure)	--
LEV	Leverage	Need to obtain additional debt or equity financing	Persons (1995)
ACHANGE	Change in Assets	Rapid growth or unusual profitability	Skousen et al. (2009)
FREEC	Free Cash Flow	Recurring negative cash flows from operations	Skousen et al. (2009)
FINANCE	Demand for Financing	Need to obtain additional debt or equity financing	Skousen et al. (2009)
INVSAL	Inventory to Sales	Declines in customer demand (Pressure)	--
SALTA	Sales to Assets	High degree of competition or declining profit margins	Persons (1995)
RECEIVABLE	Change in Receivables	Accounts based on significant estimates; Significant, unusual, or highly complex transactions	Dechow et al. (2011)
INVENTORY	Change in Inventory	Accounts based on significant estimates; Significant, unusual, or highly complex transactions	Dechow et al. (2011); Summers and Sweeney (1998)
MARKETDOM	Market Domination	A strong financial presence or ability to dominate a certain industry sector (Opportunity)	--
DIFFAUD	Difficult to Audit Transactions	Significant, unusual, or highly complex transactions (Opportunity)	--
UROE	Unusual Return on Equity	A strong financial presence or ability to dominate a certain industry sector (Opportunity)	--
TACC	Total Accruals	Aggressive or unrealistic forecasts; Interest by management in employing inappropriate means to minimize reported earnings for tax (Rationalization)	--

CACC	Current Accruals	Aggressive or unrealistic forecasts; Interest by management in employing inappropriate means to minimize reported earnings for tax (Rationalization)	--

Table 1: Continued.

Notes:
(1) The statistical significance of a variable was set at $p < 0.05$ for a one-tailed *t*-test. Each of the denoted studies from extant literature employed a form of logistic regression to detect financial statement fraud. Thus, the financial variables are only significant with respect to their corresponding logistic regression experiment(s).

(2) Financial variable definitions:

$$SCHANGE = \Delta Sales - \Delta Avg.Revenue^{Industry};$$

$$ROA = \frac{Net\ Income_t}{Assets_{t-1}};$$

$$CATA = \frac{OIBDP - OperatingNetCash}{Assets}$$, where $OIBDP$ is operating income before depreciation;

$$ROE = \frac{Net\ Income_t}{Equity_{t-1}};$$

$$GPM = \frac{Revenue - CostofGoodsSold}{Revenue};$$

$$LEV = \frac{LongTerm\ Debt_t}{Assets_t};$$

$$ACHANGE = \frac{Assets_{t-1} - Assets_{t-2}}{Assets_{t-2}};$$

$$FREEC = \frac{OperatingCash - CashDividends - CapitalExpenditures}{Assets};$$

$$FINANCE = \frac{OperatingCash_t - Avg.CapitalExpenditures_{t-2\ to\ t}}{Assets_{t-1}};$$

$$INVSAL = \frac{Inventory}{Revenue};$$

$$SALTA = \frac{Revenue}{Assets};$$

$$RECEIVABLE = \frac{Receivables_t}{Revenue_t} - \frac{Receivables_{t-1}}{Revenue_{t-1}};$$

$$INVENTORY = \frac{Inventory_t}{Revenue_t} - \frac{Inventory_{t-1}}{Revenue_{t-1}};$$

$$MARKETDOM = \frac{Avg.Revenue^{Industry}}{Revenue};$$

$$DIFFAUD = \frac{Receivables}{Revenue};$$

$$UROE = \frac{Avg.ROE^{Industry}}{ROE};$$

$TACC = \frac{\Delta Receivables_t + \Delta Inventory_t + \Delta CurrentAssets_t - \Delta AccountsPayable_t - \Delta TaxesPayable_t - \Delta CurrentLiabilities_t - Depreciation_t}{Assets_{t-1}}$;

$CACC = \frac{OIBDP - OperatingNetCash}{Revenue}$, where $OIBDP$ is operating income before depreciation.

Table 2: Summary statistics for the GA and MARLEDA. Each statistic is stated as a percentage.

Model	**Training**					**Validation**		
	Accuracy			*ActiveRules*	*ActiveVars*	*Accuracy*		
	Max	Mean	Min			Max	Mean	Min
GA	77.24	75.47	73.37	60.50	50.90	75.56	63.75	58.70
MARLEDA	77.43	74.26	71.22	66.50	49.11	73.33	64.46	52.08

Table 2: Continued.

Notes:
Summary statistics definitions:

$$Accuracy = \frac{\text{\# of correctly classified observations}}{\text{\# of total observations - \# of unclassified observations}},$$

$$ActiveRules = \frac{1}{10}\sum_{i=1}^{10} \frac{\text{\# of active rules in the most-accurate validation FRBC, } S, \text{ of fold } i}{\text{\# of total rules in } S};$$

$$ActiveVars = \frac{1}{10}\sum_{i=1}^{10} ActiveVars_j, \text{ where}$$

$ActiveVars_j = \frac{1}{d}\sum_{k=1}^{d} \frac{\text{\# of active variables in active rule } k \text{ of FRBC } j}{\text{\# of total variables in } j}$ is the active variables ratio for the FRBC, j, which contains d active rules and is the most-accurate validation classifier of fold i.

Table 3: Supplemental classification statistics for the GA and MARLEDA. The following statistics are stated as percentages: *Sensitivity*, *Specificity*, *Precision*, and *Recall*.

Model	**Validation (Total – 10 Folds; 458 Observations)**			
	Sensitivity	*Specificity*	*Precision*	*Recall*
GA	66.38	61.14	63.07	64.52
MARLEDA	68.12	60.70	63.41	65.57

Table 3: Continued.

Notes:
Supplemental classification statistics definitions:

$$\textit{Sensitivity} = \frac{\text{\# of correctly classified } \textit{Fraudulent} \text{ observations}}{\text{\# of actual } \textit{Fraudulent} \text{ observations}};$$

$$\textit{Specificity} = \frac{\text{\# of correctly classified } \textit{Non-Fraudulent} \text{ observations}}{\text{\# of actual } \textit{Non-Fraudulent} \text{ observations}};$$

$$\textit{Precision} = \frac{\text{\# of correctly classified } \textit{Fraudulent} \text{ observations}}{\text{\# of classified } \textit{Fraudulent} \text{ observations}};$$

$$\textit{Recall} = \frac{\text{\# of correctly classified } \textit{Non-Fraudulent} \text{ observations}}{\text{\# of classified } \textit{Non-Fraudulent} \text{ observations}}.$$