International Journal of Innovative
Computing, Information and Control
Volume **12**, Number **1**, February **2016**

ICIC International ©2016 ISSN 1349-4198
pp. **333–356**

THE ROLE OF IMPUTATION IN DETECTING FRAUDULENT FINANCIAL REPORTING

Stephen Obakeng Moepya[1,2], Sharat Saurabh Akhoury[1]
Fulufhelo Vincent Nelwamondo[1,2] and Bhekisipho Twala[2]

[1]Council for Scientific and Industrial Research
1 Meiring Naudé Road, Brummeria, Pretoria 0184, South Africa
{ smoepya; sakhoury; fnelwamondo }@csir.co.za

[2]Department of Electrical and Electronic Engineering Science
University of Johannesburg
Kingsway Ave and University Road, Auckland Park, Johannesburg 2092, South Africa
btwala@uj.ac.za

Received June 2015; revised December 2015

Abstract. *Financial fraud detection plays a crucial role in the stability of institutions and the economy at large. Data mining methods have been used to detect/flag cases of fraud due to a large amount of data and possible concept drift. In the financial statement fraud detection domain, instances containing missing values are usually discarded from experiments and this may lead to a loss of crucial information. Imputation has been previously ruled out as an option to keep instances with missing values. This paper will examine the impact of imputation in financial statement fraud in two ways. Firstly, seven similarity measures are used to benchmark ground truth data against imputed datasets where seven imputation methods are used. Thereafter, the predictive performance of imputed datasets is compared to the original data classification using three cost-sensitive classifiers: Support Vector Machines, Naïve Bayes and Random Forest.*
Keywords: Financial statement fraud, Missing values, Imputation, Distance metrics, Cost-sensitive classification

1. **Introduction.** Financial statement fraud (also known as management fraud) is a deliberate and wrongful act carried out by public or private companies using materially misleading financial statements that may cause monetary damage to investors, creditors and the economy. Financial statements contain information about the financial position, performance and cash flows of a company [1]. The statements also inform the reader about related party transactions. The requirement of *public* companies to issue financial statements is to allow for a *standardized* comparability between companies. In practice, financial statement fraud (FSF) might involve [2]:

1. the manipulation of financial records;
2. intentional omission of events, transactions, accounts, or other significant information from which financial statements are prepared; or
3. misapplication of accounting principles, policies, and procedures used to measure, recognize, report, and disclose business transactions.

The fall of companies such as Enron Broadband and Worldcom sparked interest in the research field of FSF detection. The collapse of Enron alone caused a $70 billion market capitalization loss which hurt investors, employees and pensioners (not to mention market sentiment/investor confidence). The Worldcom scandal, caused by alleged FSF, is the biggest bankruptcy in United States history [3].

Auditors are given the difficult task of detecting companies which intentionally issue fraudulent financial statements. According to the International Auditing and Assurance Standards Board (IAASB)[1], the auditor shall perform risk assessment procedures to provide a basis for the identification and assessment of risk of material misstatement at the financial statement and assertion levels (ISA 315) [4]. Part of the risk assessment procedures include 'Analytical Procedures' (APs) as a component. ISA 520 defines analytical procedures as evaluations of financial information through analysis of plausible relationships among both financial and non-financial data. This ISA standard deals with the auditor's use of APs as substantive analytical procedures. It requires auditors to perform APs as part of audit planning with an objective of identifying the existence of unusual events, amounts, ratios and trends that might indicate matters relating to financial statement and audit planning implications. An investigation of using financial ratios to detect fraudulent financial reporting was performed by Kaminski et al. [5]. This study used COMPUSTAT data and the paired t-test to find significant financial ratios between fraudulent and non-fraud firms.

In general, two approaches can be used to assist flagging potential cases of FSF: statistical and data mining. The data mining approach has received more attention than statistical methods in this regard. According to Kirkos et al. [6], data mining maintains a theoretical advantage over statistical methods. This is because data mining methods do not impose arbitrary assumptions over the data space. In some cases, the data in the FSF detection domain may contain missing values. For example, a data provider may not collect all the information with respect to certain aspect of a given company. Some data mining methods are not immune to the missing data problem (i.e., Support Vector Machines and k-Nearest Neighbors). Missing values in data mining can be handled in three different ways [7]:

- discard instances with missing values in their attributes, i.e., deleting attributes with elevated levels of missing values;
- the use of maximum likelihood procedures, where the parameters of a model for the complete data are estimated, and later used for imputation by means of sampling;
- estimate missing values using imputation procedures.

The focus of this study will be on using imputation techniques to estimated missing values in FSF detection. Generally, imputation is preferred when missing values in the dataset are greater than 5% of the total data amount. A fundamental advantage of the imputation approach is that the missing value treatment is independent of learning algorithm that will be used during classification. There are a wide variety of imputation techniques that have been presented in literature. These can be broken down into two main categories: statistical (such as 'mean', 'regression' and 'multiple' imputation) and machine learning-based imputation (i.e., k-Nearest Neighbor algorithm (kNN), multi-layer perceptron and self-organizing maps) [8]. One approach to compare the effectiveness of imputation schemes is to benchmark them against known values ('ground truth'). This may be achieved by measuring the similarity between imputed and known values.

Distance/Similarity measures are functions that output a non-negative number between two points. These measures are crucial to determine the closeness or similarity of one object to another. Distance measures have been used in many applications such as biometrics

[1]The International Auditing and Assurance Standards Board (IAASB) is an independent standard-setting body that serves the public interest by setting high-quality international standards for auditing, quality control, review, other assurance, and related services, and by facilitating the convergence of international and national standards. The IAASB's efforts are focused on development, adoption and implementation of International Standards (ISAs) addressing audit, quality control, review, other assurance, and related services engagements.

[9], network intrusion detection [10] and finance [11]. A classic example of where distance measures are used is in the kNN algorithm. The performance of kNN is highly dependent on the choice of distance measure. A large number of distance metrics can be found in the literature. In fact, a dictionary is available for most existing similarity metrics [12]. The main task is to select an appropriate metric that will enable optimal performance. This is, however, not a perfect science but driven by the data and domain application. This case is shown by Bharkad and Kokare [13] where the Sorgel distance metric was found to be superior to the traditional Euclidean distance with respect to the genuine acceptance rate of fingerprints.

Paragraph A12 of ISA 520 states that the *reliability* of the data is influenced by its source and nature, and is dependent on the circumstance under which it was obtained. What has not been investigated in previous literature is cases where there exists missing or corrupted data. In application domains such as DNA microarray gene expression [14], the robustness of imputation techniques (with respect to classification accuracy) has been shown when there exists poor quality or missing data. The ability of handling missing data has become a fundamental requirement for pattern classification, because inappropriate treatment of missing data may cause large errors or false results on classification. In FSF detection, should the data be missing or incorrect the results of APs can lead to incorrect conclusions about the overall audit decision. Should an auditor be placed in a situation where certain data is missing or is faced with data quality issues, the use of imputation could be a valid option. The tendency in FSF detection is to remove instances with missing values. In this study, the objective is to investigate the impact of imputation using *authentic* financial statement fraud data which contains missing or poor quality data. The paper will present a comprehensive study of the effectiveness of imputation (to estimate missing values) on classification accuracy of several cost-sensitive classifiers and for varying amounts of missing data. Thus, finding 'suitable' values to impute is at the core of this investigation. The impact of missing value imputation will be measured in two ways. Firstly, imputed values will be evaluated via distance metric scores to attain a measure of similarity to known values ('ground truth'). Furthermore, classification performance of imputed data sets will be compared against the benchmark data. This will assess the impact of imputation to assist in detecting FSF. The findings in this paper will be of use to both researchers and practitioners in the field of financial statement fraud detection who are faced with missing or unreliable data. The remainder of this paper is structured as follows. The related work is given in Section 2. Section 3 provides a short introduction into missing values and patterns of missingness, various imputation techniques and distance metrics. An analysis of the sample data and experimental setup are presented in Section 4. Section 5 presents the results of the experiment. The final section concludes the paper.

2. **Related Work.** This section provides a brief and yet comprehensive survey in the field of financial statement fraud detection. Specifically, the review focuses its attention on data driven approaches that have been utilized in this domain. Table 1 gives a quick overview of some of the literature in the past decade in the field. The first column gives the author and year of publication and the second column states the data mining technique used in the evaluation. The aim of each article is stated in the 'Main Objective' column. 'Class Distribution' states the percentage of fraudulent companies in the dataset used. Since the problem is a binary classification problem, the non-fraud class percentage is 1 - fraud. The 'Missing Data Treatment' column is meant to give an idea of how authors reported what action was taken when (or if) the data contained any incomplete instances.

TABLE 1. Summary of some literature in the field since 2005

Author	Data Mining Techniques	Main Objective	Class Distribution	Missing Data Treatment
Doumpos et al. [15] 2005	SVM	investigate the development of models that combine publicly available financial information with credit-risk indicator to explain qualifications in audit reports	Fraud (14%)	NA
Öğüt et al. [1] 2009	SVM, LR LDA, PNN	predict financial information manipulation using SVM and PNN	Fraud (50%)	NA
Pai et al. [30] 2011	SVM, MLP C4.5, LR RBF, LDA	propose a support vector machine-based fraud warning (SVMFW) model to reduce risk	Fraud (25%)	NA
Ravisankar et al. [27] 2011	MLP, SVM PNN, LR GP, GMDH	use data mining techniques in order to identify companies which resort to financial statement fraud	Fraud (50%)	NA
Persons [24] 2011	Step-wise LR	develop parsimonious models to identify factors associated with fraudulent financial reporting	Fraud (50%)	Removal of missing values
Li and Ying [17] 2010	SVM	use SVM linear and RBF kernels to detect regulating profits financial statement fraud	Fraud (50%)	NA
Gupta and Gill [31] 2012	DT, NB GP	implement data mining methodology for preventing fraudulent financial reporting	Fraud (25%)	NA
Ata and Seyrek [21] 2009	DT and NN	use data mining techniques to assist auditors detect financial statement fraud	Fraud (50%)	NA
Kotsiantis [26] 2006	SVM, C4.5 kNN, LR RIPPER, BN	explore the effectiveness of machine learning in detecting firms which issue fraudulent financial statements	Fraud (25%)	NA
Perols [29] 2011	LR, SVM ANN, C4.5 bag, stack	compare the performance of six popular statistical and machine learning models in detecting FSF under different cost assumptions	Fraud (0.3%)	Removal of missing values
Hoogs et al. [32] 2007	Genetic algorithm	present a genetic algorithm approach to detect patterns in publicly available data	Fraud (14%)	Removal of missing values
Deng [16] 2009	SVM	design a fraudulent financial statement detection model based on support vector machines	Fraud (50%)	NA
Amara et al. [25] 2013	LR	test the impact of the fraud triangle on the detection of fraud in financial statements	Fraud (50%)	NA
Gaganis [28] 2009	SVM, DA, LR ANN, PNN, KNN UTADIS, MHDIS	develop classification models for the detection of fraudulent financial statements	Fraud (50%)	NA
Roxas [33] 2011	Probit Benford's Law	compare the effectiveness of two analytical procedures in detecting earnings management through revenue manipulation	Fraud (33%)	Removal of missing values
Lou and Wang [23] 2011	LR	develop and test a logistic regression model for evaluation in the likelihood of fraudulent reporting	Fraud (16%)	Removal of missing values
Katsis et al. [34] 2012	LR, NB, QDA NN, C4.5 Ant Miner	investigate the use of a swarm intelligence technique for fraudulent financial statement detection	Fraud (18%)	NA
Lin et al. [20] 2015	LR, NN CART	examine all aspects of the fraud triangle using public data and discuss whether the results of data mining techniques agree with expert opinion	Fraud (22%)	NA
Cecchini et al. [18] 2010	SVM Financial Kernel	develop a financial kernel which constructs features that are helpful in detecting management fraud	Fraud (3%)	Removal of missing values

Doumpous et al. [15] observed the robustness of SVM as a tool to explain qualifications in audit reports. The study used data involving 1754 companies in the U.K. during the period 1998-2003. The authors concluded that non-linear SVM models did not provide improved results compared to the linear model. It was also shown that the SVM model is robust since its performance did not deteriorate significantly when tested on future data. The authors, however, did not attempt to use other classifiers in the study. Other studies which only utilized SVM as the classifiers to detect FSF were presented by Deng [16], Li and Ying [17]. Both these papers used fairly balanced datasets from Chinese listed companies.

Cecchini et al. [18] presented a financial kernel (FK) in order to be used in conjunction with SVMs. The aim of the study was to create a model with the best overall prediction while controlling the Type I error. Data from the US was used in the experiment where the ratio of fraud to non-fraud companies was 1:31. The SVM-FK model produced superior results compared to previous studies using Probit, NN and LR. However the authors in this study chose to remove attributes which contained missing values of 25% (or greater).

Another article which showed SVMs superiority to detect FSF was presented by Öğüt et al. [1] using Turkish financial statement data. The eight financial ratios used in this study were suggested by Beneish [19]. In the experiment, SVM, Logistic Regression (LR), Linear Discriminant Analysis and Probabilistic Neural Networks were used in order to predict fraudulent companies. The balanced dataset consisted of 75 (of 150) companies who were known to have committed financial statement fraud. The results showed that SVM outperformed all other classifiers using a holdout set. The authors did not mention if there were missing values in the dataset. Lin et al. [20] used three data mining techniques (LR, NN and DT) in order to detect FSF. Data from Taiwan was used in this experiment with a fraud rate of approximately 22%. Again in this study there was no mention of the treatment of missing data (or incomplete). A paper which considered only Turkish manufacturing firms to detect FSF is presented by ATA and SEYREK [21]. The data used was taken from the Istanbul Stock Exchange and period of the investigation was the year 2005. The class distribution between fraudulent firms and non-fraudulent firms was 50%. Twenty-four variables were used in the study and the t-test showed 15 variables to be statistically significant. The classification results showed NN superior to DT with an accuracy of 77.36%. The study did not mention missing data.

Pai and Hsu [22] proposed an SVM-based algorithm to minimise audit related risks by classifying FSF and presenting the auditor with comprehensible decision rules. The 75 listed companies used in the experiment were taken from the Taiwan Stock Exchange (TSE). Features used in the experiment were suggested by previous research. SVM was shown to outperform Multi-Layer Perceptron (MLP), C4.5 and LR. Although the study showed some innovation in providing the auditor with some rules, only 75 companies were used in the experiment. There was no mention of missing values in the dataset. Lou and Wang [23] used data from the TSE to develop an LR model for the evaluation in the likelihood of fraudulent reporting. The fraud instance made up 16% of the total instances in the experiment. Companies with incomplete information were not included in the dataset. Logistic stepwise models were used by [24] using US data whereby the class distribution used to train the model even. Firms from financial services industry were excluded from the data since certain financial statement variables were not available for such companies. Amara et al. [25] use LR in order to show that performance issue exerted on managers is a factor of pressure leading to commit fraud in financial statements. The data utilized in the study consists of French companies of which half were shown to be fraudulent. The authors did not mention how missing data was treated.

The successful use of supervised machine learning algorithms to detect fraudulent financial statements is presented by Kotsiantis et al. [26]. The Altman z-score and twenty four (24) financial ratios each covering profitability, leverage, liquidity, efficiency and cash flow were used as model inputs. Data consisting of only manufacturing firms from the Athens Stock Exchange was used in the experiment. Forty-one (41) out of the 164 companies in the experiment were found to have submitted fraudulent financial statements. Bayesian Networks, Decision Trees (DT), SVM, Neural Networks (NN), LR, k-Nearest Neighbors (kNN) and RIPPER were fit on the data. A proposed stacking variant methodology was shown to achieve increased performance compared to any of the examined simple and ensemble methods. The study only included companies who were in the manufacturing sector. Companies in other sectors could have been considered. Kirkos [6] utilized Greek manufacturing firms and data mining techniques to detect FSF. Using a 38 non-fraud and 38 fraud firms, the authors showed that BN achieved a superior performance.

Ravisankar et al. [27] tested the ability of six classifiers in order to detect FSF. The dataset involved 202 from listed Chinese companies (101 were found to have issued fraudulent financial statements). Thirty-five ratios were used in the experiment and the t-test was used to filter the relevant variables. The authors did not state how they dealt with missing or incomplete instances. A comparative study using ten classifiers was performed by Gaganis [28] in order to identify falsified financial statements. The matching principle was used in order to select the sample; therefore, the class distribution was balanced. The author also included some additional information (which are not in the form of financial ratios) such as the type of auditor and whether the company had a litigation against them. Unfortunately there was no mention of missing data.

A more recent study undertaken by Perols [29] compares the performance of machine learning algorithms using data from 1998 through to 2005. Using American listed companies, logistic regression (LR), artificial neural networks, bagging, stacking, C4.5 and SVM were used in the experiment. The data set consisted of 15934 non-fraud and 272 fraud observations. Out of 272 fraudulent firms, 221 were discarded from the experiment due to some form of missing data. The results showed that LR and SVM outperform the other methods under this high class-imbalance. The non-removal of many fraudulent companies may have lead to better classification results.

Thus far, the above literature review has shown the progress made both in detecting financial statement fraud using data mining techniques and the use of imputation in the finance domain. No one algorithm has shown to be superior in detecting FSF using data from different countries. However, datasets which contain less than 10% have not been extensively investigated in this domain to identify companies who issue fraudulent financial statements. Only two publications in Table 1 investigate this scenario. In terms of imputation in the FSF domain, to the best of our knowledge, no effort has been made to keep instances with missing data. In Table 1, six papers removed instance which contain missing values and the rest do not mention whether the data contained any missing values. This study intends to add to the body of work in the field of financial statement fraud detection by evaluating the performance of *imputation* using an authentic (real world) dataset containing a *high class-imbalance*. The paper intends to show that imputation can be seen as a valid option when encountered with missing data in this domain. The similarity of imputed datasets with respect to the ground truth will be assessed using distance metrics. Since the main objective is to ultimately train models to classify instance into fraud and non-fraud, the imputed datasets tested against the benchmark classification performance. The study will be useful to practitioners in this field (or possibly in other finance-related domains) in the scenario where there are data quality issues or data is missing.

3. **Background Review.** This section provides brief description of missing data mechanisms, imputation techniques and distance metrics.

3.1. **Missing data mechanisms.** Generally, given data with missing values, the following options are available to practitioners and researchers:

- case-wise deletion; or
- imputation of missing values.

Case-wise deletion is the removal of any instance in a dataset which contains a missing value. This approach may be feasible if the data contains a very small amount of missing values. Otherwise, cases which may contain valuable information will be removed and experiments could lead to biased/misinformed results.

The imputation of missing values can be broadly split into two categories: statistical and machine learning based imputation. Statistical imputation includes methods such as mean, hot-deck and multiple imputation methods based on regression and the expectation maximization (EM) algorithm [8]. Machine learning approaches for imputing missing values create a predictive model to estimate missing values. These methods model missing data estimation based on available information in the data. For example, if the observed dataset contains some useful information for predicting missing values, the imputation procedure can utilize this information and maintain a high precision.

Before any researcher or practitioner decides to use imputation, he/she needs to ask an important question: 'Why is the data missing?'. Identifying the nature of missingness is important since it assists in the understanding of the data and justifies the choice of imputation used. Identifying patterns of missingness helps to determine whether values are:

- missing completely at random (MCAR);
- missing at random (MAR); or
- missing not at random (MNAR).

Missing completely at random (MCAR) occurs when the probability that a variable is missing is independent of the variable itself and any other external influences. This implies that the available variables contain all information to make inferences. For example, a financial data provider's systems failed to capture stock market data for a specific company because of a technical fault. Missing at random (MAR) is a mechanism where the missingness is independent of the missing variables but the pattern of data missinginess is traceable or predictable from other variables in the dataset. Given a set of financial statement ratios, if 'Dividend per Share' is missing for a given company instance then we would expect 'Dividend Cover' and 'Dividend Yield' to be missing. This scenario implies that no dividends were issued for that company instance. Therefore, in that case, the missingness is predictable and hence MAR. Finally, missing not at random (MNAR) refers to the pattern of data missingness which depends on the missing variable. In this situation, the missing variables cannot be predicted only from the available variables in the data. If market data for a specific listed company is unavailable due to market regulators halting trading for that stock, since it has reached a particular level, then the data is considered MNAR.

When data are MCAR and MAR, the missing data mechanism is termed ignorable. Ignorable mechanisms are important, because when they occur, a researcher can ignore the reasons for missing data in the data analysis, and thus simplify the methods used for missing data analysis [8]. For a more comprehensive explanation of missing data mechanisms, the reader is referred to [35].

The performance can be generally measured in two ways. In the first case, once the imputed values have been filled in, the classification error rate is measured over the imputed dataset [36]. The second way to compare imputation methods is to measure the similarity between imputed data and the ground truth. Olivas et al. [37] suggest two ways of doing this predictive accuracy (PAC) and distributional accuracy (DAC). The authors state that PAC can be given by the Pearson correlation between imputed and ground truth values. Correlation values closer to 1 will imply good imputation. DAC involves finding the distance between distribution function for both the imputed and the ground truth values. The Kolmogorov-Smirnov (KS) distance is used to determine the distance between the two distribution functions. A good imputation method will give values closer to 0 (i.e., a smaller distance value).

3.2. **Imputation techniques.** A standard data mining approach, when encountering missing values, involves imputing estimates for values where the ground truth is unknown. In this subsection we describe the imputation techniques which will be used in the experimental section.

Mean imputation is one of the simplest methods to estimate missing values. Consider a matrix X containing a full data set. Suppose that the value x_{ij} belongs to the kth class C_k and it is missing. Mean imputation replaces x_{ij} with $\bar{x}_{ij} = \sum_{i:x_{ij}\in C_k} \frac{x_{ij}}{n_k}$, where n_k represents the number of non-missing values in the jth feature of the kth class.

In kNN imputation [38], missing cases are imputed using values calculated from corresponding k-nearest neighbors. The nearest neighbor of an arbitrary missing value is calculated by minimizing a distance function. The most commonly used distance function is the Euclidean distance between two instances y and z as $d(y,z) = \sqrt{\sum_{i\in D}(x_{yi} - x_{zi})^2}$, where D is a subset of the matrix X containing all instances without any missing values. Once k-nearest neighbors are computed, the mean (or mode) of the neighbors is imputed to replace the missing value.

Principal Component Analysis (PCA) imputation involves replacing missing values with estimates based on a PCA model. Suppose that the columns of matrix X are denoted by d-dimensional vectors $y_1, y_2, \cdots, y_n$. PCA imputation assumes that these vectors can be modeled as $y_j \approx Wz_j + m$, where W is a $d \times c$ matrix, z_j are the c-dimensional vectors of principal components and m is a bias vector. This imputation method iterates and converges to a threshold by minimizing the error $C = \sum_{j=1}^{n} \|y_j - Wz_j - m\|^2$.

The Expectation-Maximization (EM) is an iterative procedure that computes the Maximum Likelihood Estimator (MLE) when only a subset of the data is available. Let $X = (X_1, X_2, \cdots, X_n)$ be a sample with conditional density $f_{x|\Theta}(x|\theta)$ given $\Theta = \theta$. Assume that X has missing variables $Z_1, Z_2, \cdots, Z_{n-k}$ and observed variables $Y_1, Y_2, \cdots, Y_k$. The log-likelihood of the observed data Y is

$$l_{obs}(\theta; Y) = \log \int f_{X|\Theta}(Y, z|\theta) v_z(dz). \tag{1}$$

To maximize l_{obs} with respect to θ, the E-step and M-step routines are used. The E-step finds the conditional expectation of the missing values given observed values and current estimates of parameters. The second step, the M step, consists of finding maximum likelihood parameters as though the missing values were filled in [39]. The procedure iterates until convergence.

Singular Value Thresholding (SVT) [40] is a technique that has been used in Exact Matrix Completion (MC). MC enables the recovery of a low-rank matrix or approximately low-rank matrix $M \in \mathbb{R}^{n_1 \times n_2}$ from at least $O(nr\nu \ln^2 n)$ entries selected uniformly at random (with ν corresponding to the so-called incoherence), where $n = \max\{n_1, n_2\}$ and

$r = rank(M)$. The original matrix can be recovered from the partially observed matrix by solving the convex optimization problem

$$\begin{aligned} &\min_X \|X\|_* \\ &\text{s.t.} \quad X_{ij} = M_{ij}, \quad (i,j) \in \mathcal{I} \subset \{1, \cdots, n_1\} \times \{1, \cdots, n_2\}, \end{aligned} \tag{2}$$

where $|\mathcal{I}| \geq Cnr \ln^2 n$ denotes the number of observed entries (C is a positive constant), $X \in \mathbb{R}^{n_1 \times n_2}$ is the decision variable and the nuclear norm is defined as $\|X\|_* = \sum_{q=1}^{\min\{n_1,n_2\}} \sigma_q$ with $\sigma_1, \cdots, \sigma_{\min\{n_1,n_2\}} \geq 0$ corresponding to the singular values of X. The SVT algorithm solves the following problem:

$$\begin{aligned} &\min_{X \in C} \tau \|X\|_* + \frac{1}{2}\|X\|_F^2, \\ &\text{s.t.} \quad \mathcal{A}_{\mathcal{I}}(X) = \mathcal{A}_{\mathcal{I}}(M), \end{aligned} \tag{3}$$

where $\tau \geq 0$ and the first and second norms are the nuclear and Frobenius norms respectively. In the above equation, $\mathcal{A}$ is the standard matrix completion linear map where $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \leftarrow \mathbb{R}^k$. SVT is comprised of following two iterative steps:

$$\begin{cases} X_t = \mathcal{D}_\tau \left(A^*_{\mathcal{I}}(y_{t-1}) \right) \\ y_t = y_{t-1} - \delta \left(A_{\mathcal{I}}(X_t) - b \right). \end{cases} \tag{4}$$

In the above equation, the shrinkage operator $\mathcal{D}_\tau$, also known as the soft-thresholding operator, is denoted as $\mathcal{D}_\tau = U \Sigma_\tau V^T$ where U and V are matrices with orthonormal columns and $\Sigma_\tau = diag(\max\{\sigma_i - \tau, 0\})$ with $\{\sigma_i\}_{i=1}^{\min\{n_1,n_2\}}$ corresponding to the singular values of the decomposed matrix. The step size of the iterative algorithmic process is given by δ.

Random Forests (RF), introduced by Breiman [41], is an extension of a machine learning technique named bagging which uses Classification and Regression Trees (CART) to classify data samples. RF extends the idea of bagging by allowing random selection of both the number of instances (rows in X) and predictors (columns of X) at each splitting step. Imputation via RF begins by imputing predictor means in place of the missing values. An RF is subsequently built on the data using roughly imputed values (numeric missing values are re-imputed as the weighted average of the non-missing values in that column). This process is repeated several times and the average of the re-imputed values is selected as the final imputation.

A brief explanation of Singular Value Decomposition (SVD) imputation follows. Consider the SVD of a matrix $X \in \mathbb{R}^{n_1 \times n_2}$ of rank r. In this instance, $X = U \Sigma V$, U and V are $n_1 \times r$ and $n_2 \times r$ orthogonal matrices respectively and $\Sigma = diag\left(\{\sigma_i\}_{1 \leq i \leq r}\right)$. The σ_is are known as the positive singular values. SVD imputation begins by replacing all missing values with some suited value (mean or random). The SVD is computed and missing values replaced with their prediction according to SVD decomposition. The process is repeated until the imputed missing data fall below some threshold.

3.3. **Distance/similarity metrics.** In a formal sense, distance can be defined as follows. A distance is a function d with non-negative real values, defined on the Cartesian product $X \times X$ of a set X. It is termed a metric on X if $\forall\, x, y, z \in X$ it has the following properties:

1. $d(x, y) = 0 \iff x = y$;
2. $d(x, y) + d(y, z) \geq d(x, z)$; and
3. $d(x, y) = d(y, x)$.

Property 1 asserts that if two points x and y have a zero distance then they must be identical. The second property is the well known triangle inequality and states, given three distinct points x, y and z, the sum of two sides xy and yz will always be greater than or equal to side xz. The last property states that the distance measure is symmetrical.

According to Deza and Deza [12], distances that are suitable to this study are classified under 'Distances and Similarities in Data Analysis'. The data in this category can take the following form:

- numerical (including continuous and binary numbers);
- ordinal (numbers expressing rank only); or
- nominal (not ordered).

The experimental data that is presented in this work is numerical; hence only distance metrics which are suitable to this type of data will be considered. Furthermore, Cha [42, 43] provides a taxonomy of numerical distance measure. These measures are divided into the following families: L_p Minkowsky, L_1, Intersection, Inner Product, Fidelity or Squared-chord, Squared L_2 or χ^2, Shannon's entropy and combination. A description of one metric in each relevant family follows.

The Lorentzian distance [12] is represented by the natural log of the absolute difference between two vectors,

$$d(x, y) = \sum_{i=1}^{n} \ln(1 + |x_i - y_i|), \tag{5}$$

where $x = (x_1, x_2, \cdots, x_n)$ and $y = (y_1, y_2, \cdots, y_n)$. To ensure that the non-negativity condition is adhered to, one is added. This distance metric is sensitive to small changes since the log scale expands the lower range and compresses the higher range.

A commonly used metric is shown in the following equation

$$d(x, y) = \sqrt[p]{\sum_{i=1}^{n} |x_i - y_i|^p}. \tag{6}$$

This distance metric is known as the Minkowski distance [44]. For $p = 1$, it is known as the Manhattan distance. The Euclidean distance is a special case where $p = 2$. When $p = \infty$, it is known as the Chebyshev distance.

The dice distance, from the intersection family of distance metrics, is defined by

$$d(x, y) = \frac{\sum_{i=1}^{n} (x_i - y_i)^2}{\sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} y_i^2}, \tag{7}$$

where $x = (x_1, x_2, \cdots, x_n)$ and $y = (y_1, y_2, \cdots, y_n)$. This metric can be sensitive to values near zero. The dice distance is commonly used in information retrieval in documents and biological taxonomy [12].

A distance measure which forms the basis for the χ^2 family of distance metrics is the Squared Euclidean,

$$d(x, y) = \sum_{i=1}^{n} (x_i - y_i)^2. \tag{8}$$

If $d(x, y)$ has a small value, it indicates that the vector x is close to y. This metric is the same as the Euclidean distance without the square root.

The Motyka similarity [12] is a measure that is in the intersection family of distance metrics. It is defined by

$$d(x, y) = \frac{\sum_{i=1}^{n} \max(x_i, y_i)}{\sum_{i=1}^{n} (x_i + y_i)}. \tag{9}$$

The numerator is normalized by the sum of the elements at each point. This distance metric is equivalent to half of the Czekanowski metric in the same family. In this metric, the smaller the value, the closer the similarity.

The above similarity metrics each represent five of the eight types of similarity families. The data that will be used is not suitable for the Squared-chord, Shannon's entropy and combination families since it contains negative real values (see Table 2). The following section outlines some of the literature in the field.

4. **Empirical Studies.**

4.1. **Data description.** The dataset used in this experiment was obtained from INET BFA, one of the leading providers of financial data in South Africa. The data comprises publicly listed companies on the Johannesburg Stock Exchange (JSE) between years 2003 and 2013. The different sectors for the listed companies on the JSE are: Basic Materials; Consumer Goods; Consumer Services; Financial; Health Care; Industrial; Oil and Gas; Technology; Telecommunications and Utilities. In the dataset, 123 (out of 3043) instances were known to have received a qualified financial report by an accredited auditor.

TABLE 2. Summary statistics of the selected variables in the data

Variable	Min	1st Quartile	Median	Mean	3rd Quartile	Max
Assets-to-Capital Employed	−26.120	1.060	1.240	1.612	1.60	224.30
Book Value per Share	−966810	79	418	22279	1610	49596137
Cash flow per Share	−7198	8	84	3177	396	5159700
Current Ratio	0.010	1.00	1.410	3.283	2.251	726.130
Debt to Assets	0.0	0.270	0.4800	0.8804	0.680	1103.00
Debt to Equity	−182.370	0.350	0.820	2.516	1.730	760.940
Earnings per Share	−460338.5	2.3	44.5	391.6	218.2	825937.7
Inflation adjusted Profit per Share	−9232	1	43	2858	239	4898104
Inflation adjusted Return on Equity	−87472.97	3.11	13.45	−55.60	23.37	17063.16
Net Asset Value per Share	−373040	60	405	29438	1817	66658914
Quick Ratio	0.01	0.730	1.050	2.949	1.720	726.130
Retention Rate	−7204.35	57.67	89.39	70.97	100.00	5214.29
Return on Equity	−13600.00	4.045	14.830	−3.549	25.420	17063.160
Return on Capital Employed	−13600.00	1.500	8.700	−0.551	17.415	6767.330

The complete dataset contains 48 financial ratios (features). The ratios capture different aspects of company performance such as profitability, solvency, liquidity, leverage and valuation. For the purpose of this experiment, 14 ratios were used. The featured ratios, along with some summary statistics, are presented by Table 2. The choice of these ratios is deliberate. An explanation follows.

It was observed that all the features in the dataset contained at least one missing value. Therefore, to retain all the minority class instances, features in this class which did not contain any missing values were kept and all others discarded. Of the 48 features, only 14 (see Table 2) contained no missing values. To complete the selection, the 14 features from the minority class were used as a criterion to select the majority class instances. Companies (in the majority class) which did not contain missing values for the 14 features were included for the experiment. Thirty-two (32) majority class company instances were ignored since they did not meet the criterion. Therefore, the final data set used in the experimental setup contained a dataset with 14 features and 3011 instances (123 fraud). A brief description of the financial ratios in Table 2 is given below.

The 'Assets-to-Capital Employed' ratio is a measure of the total assets per capital employed (owners equity). A high value for this ratio indicates that a company has greater current liabilities. 'Book Value per Share' is a common metric for the valuation of a company. A ratio which measures a company's liquidity per given share is termed

'Cash flow per Share'. A higher ratio value could be due to a low amount of ordinary shares issued or a greater amount of disposable cash. 'Current Ratio' is a metric that is used to ascertain whether a company can meet its short-term obligations with short-term assets[2]. 'Debt to Assets' ratio shows the proportion of a company's assets which are financed through debt. If the ratio is less than 1, most of the company's assets are financed through equity. The 'Debt to Equity' metric is a ratio of total liabilities to shareholders equity. It is categorized as a leverage ratio and measures the degree to which the assets of a business are financed by the debt and shareholders. The term 'Earnings per Share' (EPS) represents the portion of company earnings, net of taxes and preferred stock dividends, that is allocated to each share of common stock. 'Inflation-adjusted Profit per Share' measures the amount of profit for the number of ordinary shares issued. An indicator of a company's short-term liquidity is measured by the 'Quick ratio'. The 'Quick Ratio' measures the ability to meet short-term obligations with liquid assets. The 'Retention Rate', sometimes called the plow-back ratio, is a financial ratio that measures the amount of earnings or profits that are added to retained earnings at the end of the year. 'Return on Equity' (ROE) measures an organization's profitability by revealing how much profit is generated with money shareholders have invested. 'Return on Capital Employed' (ROCE) is financial ratio that measures the profitability and efficiency with which capital is employed. A higher ROCE indicates a more efficient use of capital. ROCE should be higher than capital cost; otherwise it indicates that a company is not employing its capital effectively and not generating shareholder value.

4.2. **Experimental setup.** The experimental setup is a crucial task that needs to be addressed correctly in order to meet the objectives of the study. Two key areas that need to be assessed are:

- the similarity of imputed datasets to the ground truth data; and
- classification performance of imputed data.

The first item will be addressed using Monte Carlo simulation. Random missingness will be artificially created for each trail and seven imputation schemes will be used to fill the missing values. Then the seven similarity metrics (See Section 3.3) will be used to check similarity between the ground and imputed datasets. Once the Monte Carlo simulation is complete then the median distance and standard deviation will be presented for each distance metric. This is done in order to show the accuracy of imputation with reduced bias. The imputation schemes have parameters that needed to be selected. The parameter settings were guided by a previous study [45] and are presented in Table 3.

Once the Monte Carlo simulation is complete, ten randomly selected trails are used for classification. Three classifiers will evaluate the accuracy of the imputed data: class-weighted Support Vector Machines (CW SVM), cost-sensitive Random Forests (CS RF) and cost-sensitive Naïve Bayes (CS NB). Cost-sensitive/weighted learners provide an alternative when encountered with class- imbalanced data. Using a cost matrix, a cost-sensitive learner forms a generalization such that the average cost on previously unobserved instances is minimized (instead of the average misclassification rate). For more about imbalanced learning and cost-sensitive classification we refer the reader to [46, 47, 48]. The choice of classifiers, for this specific data, was guided by a previous study [49]. Also, cost-sensitive learners were previously used by [29] in order to detect FSF using US data. Ten-fold cross-validation (CV) was used in order to avoid over-fitting. Parameter tuning was performed using the grid search technique. Receiver Operating Curve (ROC) will be

[2]Short-term in a finance is considered to be a period no more than 12 months. In the case of short-term obligations (such as loans), the current ratio measures the ability to pay back loans with assets that are expected to be around for less than a year.

TABLE 3. Imputation method parameters

Method	Parameters
EM	$tol = 1e-04$ empirical prior = 1%
kNN	$k = 3$ dist func = weighted mean
Mean	NA
PCA	max iterations = 1000 principal components = 2 threshold = $1e-06$
RF	ntree = 300 iter = 5 mtry = 4
SVD	rank approximation = 3 max iterations = 1000
SVT	threshold = $1e-03$ max iterations = 10

used as a measure of ability of a classifier to separate the fraud from the non-fraud case. The results from the ten trails are taken and one ROC curve is drawn for each classifier. Parameter tuning for the SVM and RF are as follows. For class-weighted SVM RBF kernels, the cost parameter grid $C = 2^{-6}, 2^{-5}, \cdots, 2^{8}, 2^{9}, 2^{10}$ was chosen. The values for hyper-parameter sigma (σ), in the RBF, were varied using $\sigma = 2^{-6}, 2^{-5}, \cdots, 2^{8}, 2^{9}, 2^{10}$. The specific grid methodology for the SVM was suggested by Hsu et al. [50]. Two parameters in the cost-weighted RF were varied. Then number of trees was varied using the grid $ntree = \{50, 100, 150, 200\}$ and the number of randomly selected features $nfeat = \{2, 3, 4, 5, 6\}$.

The experiments for this paper were conducted on an Intel(R) Core (TM) i5-3337U CPU @ 1.80 GHz with 6 GB memory. The implementation for algorithms are performed using the following R packages: 'imputation'[3] 'Amelia' [51], 'randomForest' [52], 'yaImpute' [53], 'CORElearn' [54] and 'caret' [55].

1. Create 6 levels of missingness randomly using 1%, 2%, 5%, 10%, 15% and 20% as missing proportion[4];
2. Impute missing values on each missingness level using SVD, kNN, PCA, SVT, Mean, EM and RF imputation;
3. Compute the distance between the imputed and the corresponding 'ground truth' values using the 5 distance/similarity measures (given in Section 3.3).
4. Classify 10 randomly selected Monte Carlo trails using class-weighted SVM, cost-weighted RF and cost-weighted NB.

The justification for using the imputation methods is that the pattern of missingness in the data is missing at random (MAR). Since the missingness was generated in a random fashion (see Step 1 above), the pattern of missingness can be termed MAR. This implies the missingness is independent of the missing variables.

[3]This package has been archived in CRAN repository.

[4]According to Acuna and Rodriguez [56], rates of less than 1% missing data are generally considered trivial and 1-5% rates are manageable. However, missingness rates that are 5-15% require sophisticated models to handle, and more than 15% may severely impact any kind of interpretation.

5. **Results.** In this section, the results of the Monte Carlo simulation are given. An analysis of the performance of the imputed data relative to the ground truth will be undertaken. In addition, the classification results are presented.

5.1. **Similarity metric results.** The figures in this subsection are box plots for the similarity measures averaged over 100 Monte Carlo simulation trails. For each plot, the horizontal line within the box indicates the median distance, the bottom and top edges represent the 25 and 75 percentiles respectively, the whiskers extend to the most extreme points which are not considered outliers, and the points marked in red are outliers.

The first two similarity metrics results that will be presented are the predictive accuracy (PAC) and distributional accuracy (DAC) measures for the seven imputation schemes. PAC and DAC are suggested by [7] and [37] for measuring the quality of missing data estimation using different missing data percentages and different combination of attributes. Figure 1 gives the results for PAC over the 100 Monte Carlo simulations. The values for this metric range from -1 to 1 as a result of using Pearson correlation [57]. Values closer to one indicate that the imputed values are closer to the ground truth. At all levels of missingness, PCA imputation outperforms all other imputation schemes with respect to the median value. At higher levels of missingness, the standard deviation of PCA imputation seems to be greater than other methods. kNN imputation at missingness greater than 5% seems to produce zero median distances with very little standard deviation.

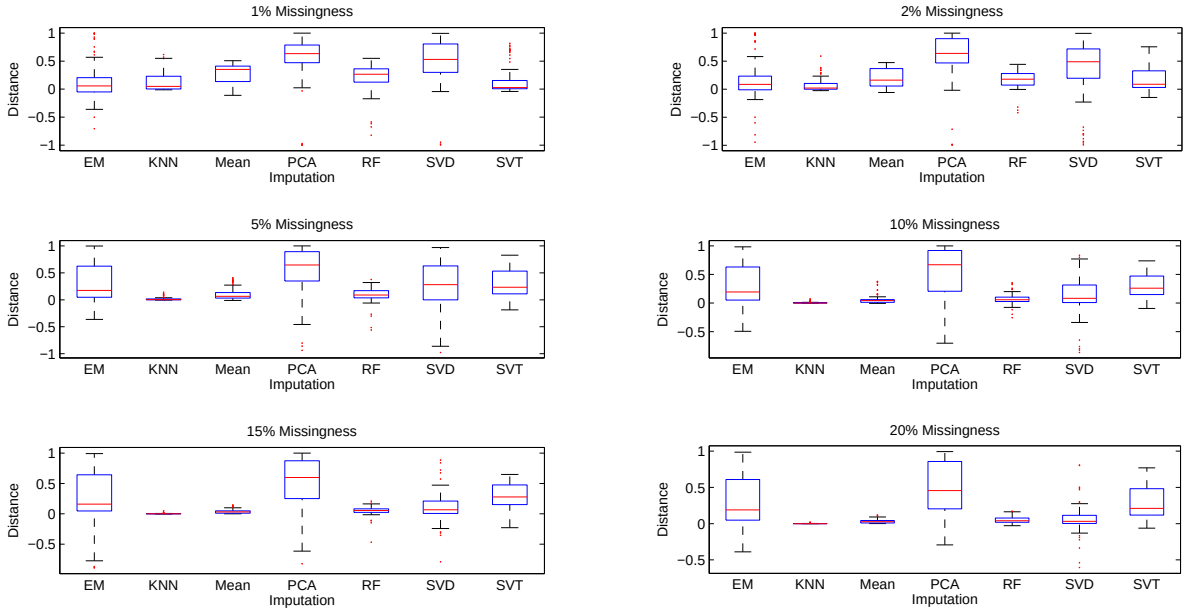


FIGURE 1. Predictive accuracy (PAC) for the seven imputation schemes

The next set of results represents the DAC of the seven imputation schemes. Figure 2 represents the box plots using the Kolmogorov-Smirnov (KS) [58] distance. The preservation of the ground truth distribution by the imputed datasets is represented by a score between zero and one. Values closer to zero represent good preservation. The first thing to note is that the standard deviation of each method is smaller, using this metric, than the standard deviation in Figure 1. SVT imputation has the worst performance using the KS distance for all levels of missingness. The median scores for this imputation scheme

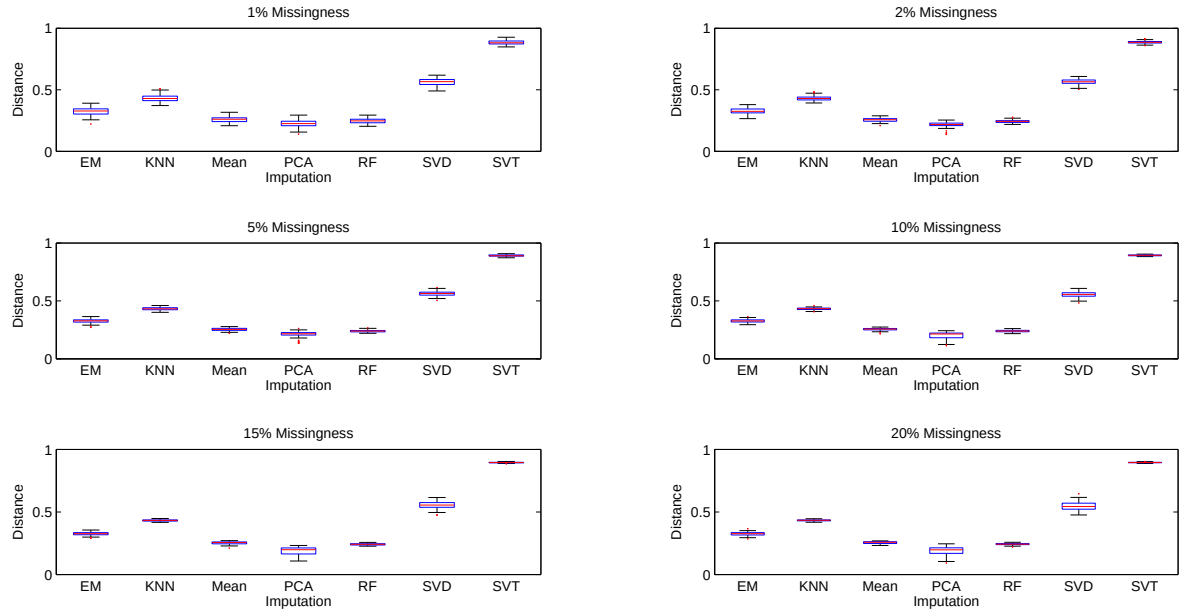


FIGURE 2. Distributional accuracy (DAC) for the seven imputation schemes

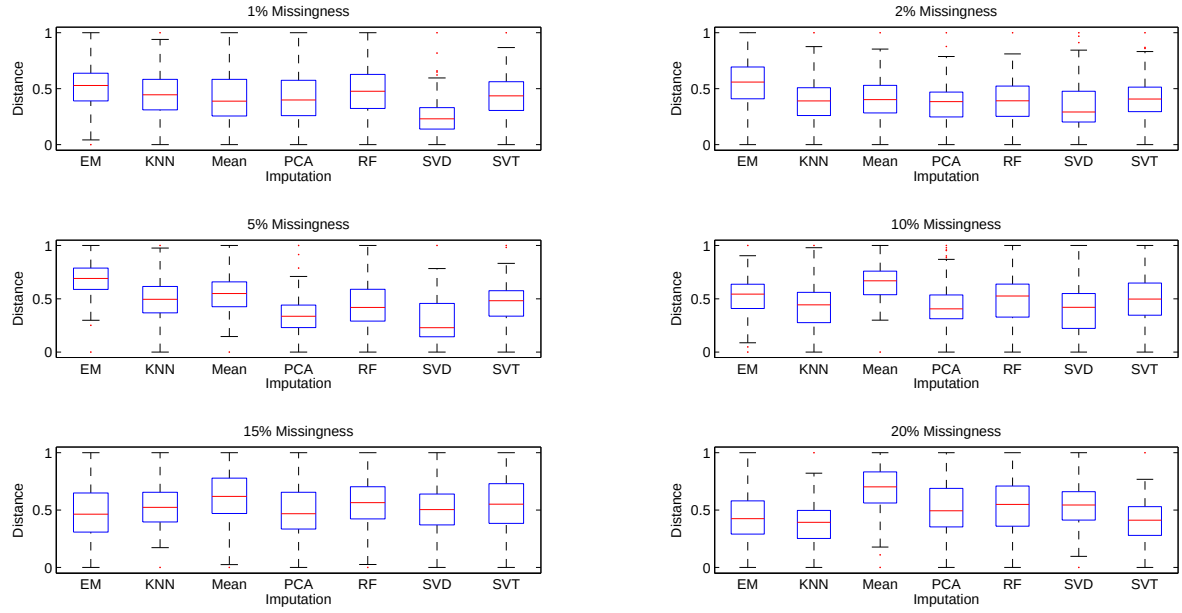


FIGURE 3. Lorentzian distances for the seven imputation schemes

are closer to 1. PCA imputation produced the lowest median distance using the KS distance over the 100 Monte Carlo simulations. It is interesting to note that the percentage missingness seems to make little difference to the result using this metric. The following results present the similarity measures chosen in this experiment.

The results in Figure 3 show the Lorentzian distance does not favor one particular type of imputation scheme for all levels of missingness. For 1% missingness, SVD imputation

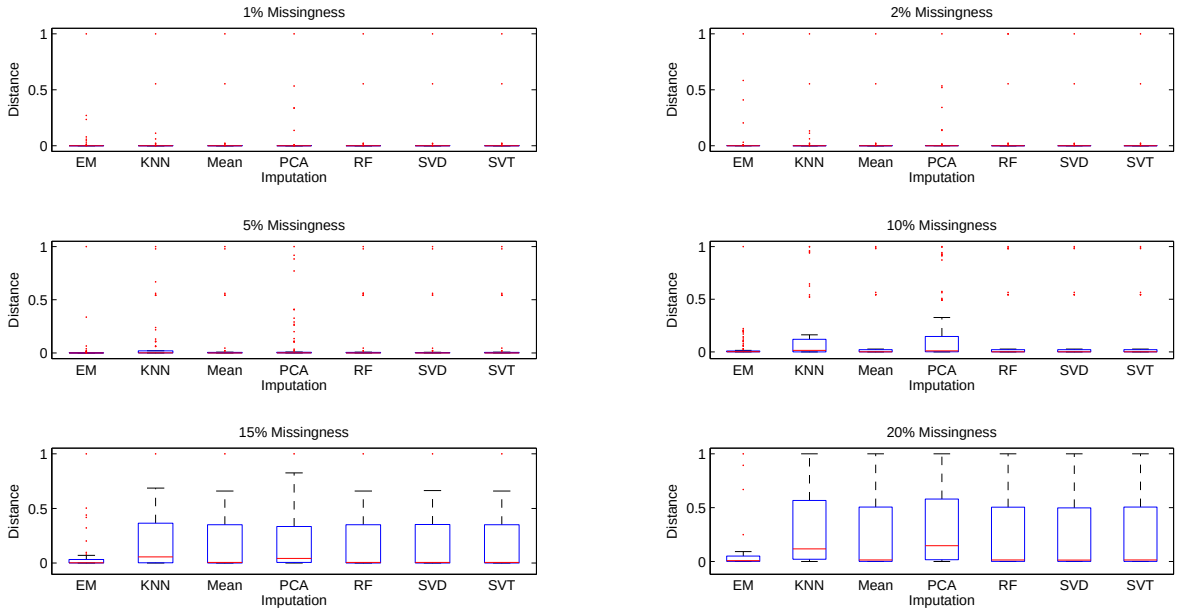


FIGURE 4. Squared Euclidean distances for the seven imputation schemes

attains the least median value and a smaller variance compared to the rest. At the 20% level of missingness, SVT and kNN imputation produce favorable results obtaining lower median values and standard deviations.

Figure 4 presents the results for the Squared Euclidean distance metric. It can be seen that for the lowest 3 levels of missingness, the results are comparable for all imputation methods. However, for missingness greater than 5%, EM outperforms all other schemes with very low medians and standard deviations. As the missingness increases from 10%-20%, EM maintains a similar median and standard deviation while other imputation methods obtain increased standard deviations.

Using the Dice distance, Figure 5 shows that RF achieves the lowest median values for 1%, 5% and 15% missingness. RF also achieves low standard deviations for these missingness levels. For the remaining levels, PCA and SVD outperform other imputation methods. kNN imputation generally produces the highest median distance for this metric.

The Manhattan distance presents the most intuitive (expected) results for different missingness levels. In Figure 6, the median values and standard deviations grow as the level of missingness increases. This shows similar behavior given by the Squared Euclidean distance metric. For missingness levels 1%-10%, mean imputation shows superior performance. EM imputation achieves satisfactory results for missingness greater than 10%, i.e., lower median scores and tighter standard deviation bands.

Figure 7 shows the peculiar behavior of the Motyka distance metric. For missingness levels 1%, 5% and 20%, EM achieves fairly low median and standard deviation scores. However, for all other levels, EM score medians above 0.5 with low standard deviation. Using the Motyka distance shows that, in general, as missingness increases the standard deviation increases. Comparing levels 15% and 20% to 5% and 10% illustrates this point.

In summary, both the PAC and DAC measures favor PCA imputed datasets with respect to median scores for all levels of missingness. SVT imputation shows consistently inferior performance with respect to DAC. Using the Lorentzian distance does not give

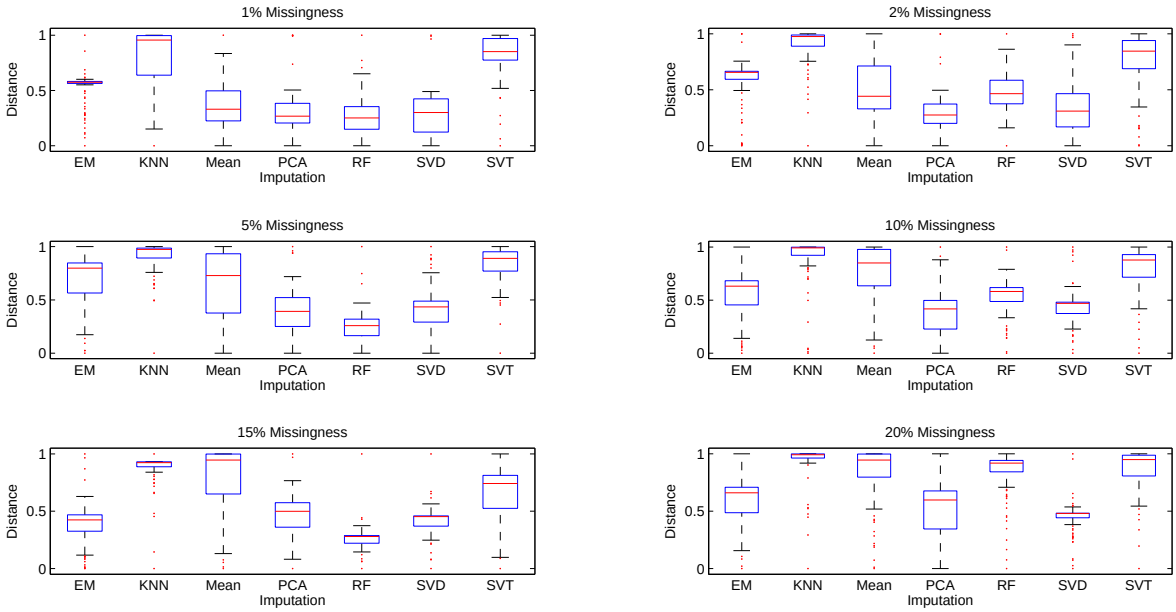


FIGURE 5. Dice distances for the seven imputation schemes

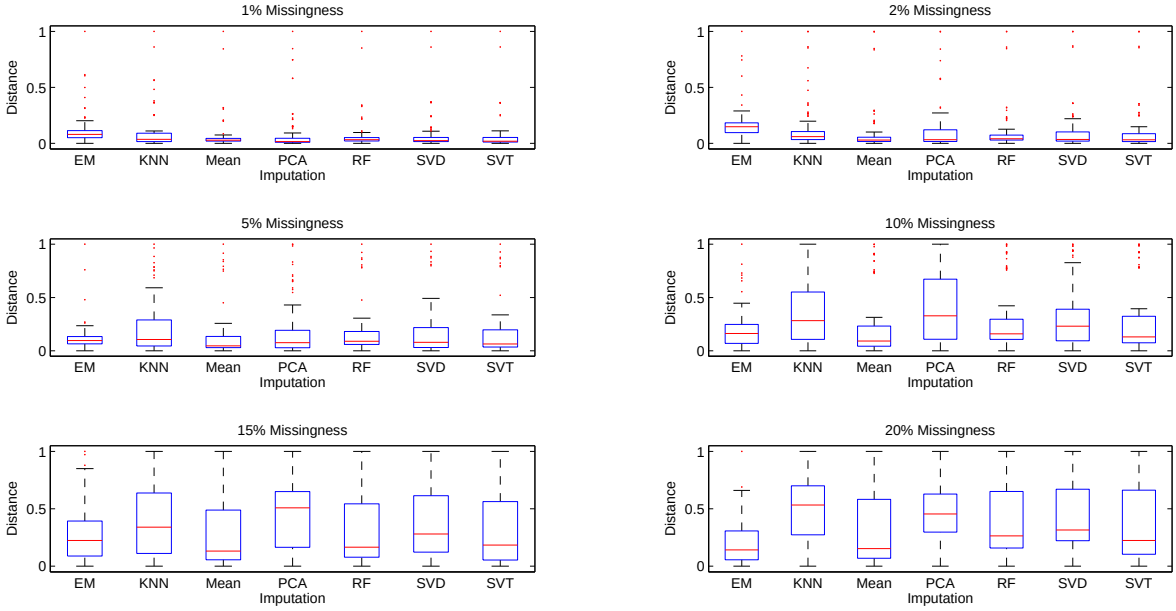


FIGURE 6. Manhattan distances for the seven imputation schemes

a clear indication of which imputed dataset is similar to ground truth for most or all missingness levels. The Squared Euclidean metric shows very small standard deviation and median values for all imputation schemes using missingness levels of 1%-5%. Generally, this metric seems to favor EM imputation for missingness greater than 5%. The Dice distance, in general, seems to favor (with respect to median and standard deviation) RF imputation for most levels of missingness. The median values are generally higher

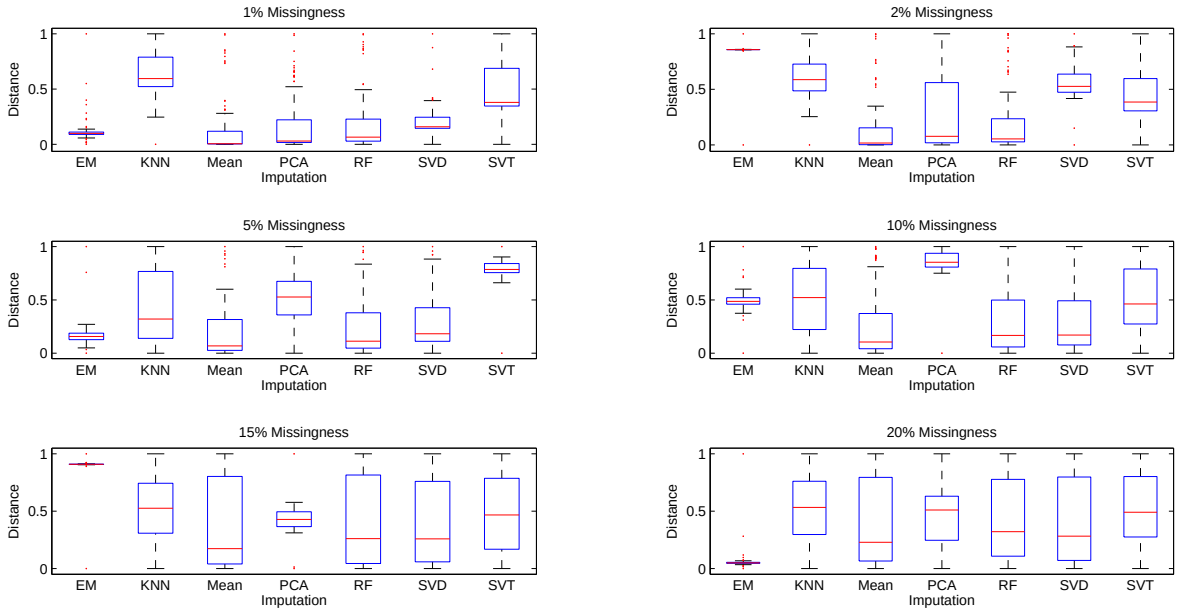


FIGURE 7. Motyka distances for the seven imputation schemes

than 0.5. The results presented by Figure 6 (Manhattan distance) highlights the fact that EM and mean imputation produce the lowest median as well as the lowest variation. As the missingness increases, the standard deviation rises. The Motyka metric shows that EM imputation has the lowest variation at all levels of missingness while in some cases the median values can be close to 1. Generally, mean imputation produces lower median scores.

5.2. **Classification results.** The classification results are presented in this section. The analysis will begin with the ground truth classification. This entails using the original dataset *without* comparison to any imputed datasets.

The ROC curve [59] plots the percentage of true positive (TP) against the false positive (FP) rate. The closer the curves are to the upper left-hand corner, the higher the model accuracy. Previous studies [18, 28] chose ROC curves to compare classifier performance.

ROC curves enables researchers to not only cope with skewed data but to also visualize the performance of classifiers [60]. As a part of Analytical Procedure (AP) in the audit process, the objective is to select a classifier to detect fraudulent companies with a high TP rate without having an increased FP rate. Since the data contains a high class-imbalance (approximately 5% of instances are fraudulent out of a total of 3011) it is key to maintain low false positives. This will enable a classifier that has been deployed into a system to flag fraudulent companies without requiring many non-fraudulent cases to be re-audited. Therefore, the ideal/target false positive rate is less than 20%.

Figure 8 presents the results for the ground truth dataset using three classification methods: CW SVM, CS NB and CS RF. In this figure, the ROC curves for the three classifiers are presented. For FP rate less than 20%, the CS RF curve lies above both CS NB and CW SVM. This shows superior performance in this region of the ROC space. At 80% TP, RF produces a lower FP rate (16.2%) while CS NB and CW SVM score 17.1% and 22.3% respectively. Therefore, using the ground truth dataset it can be concluded

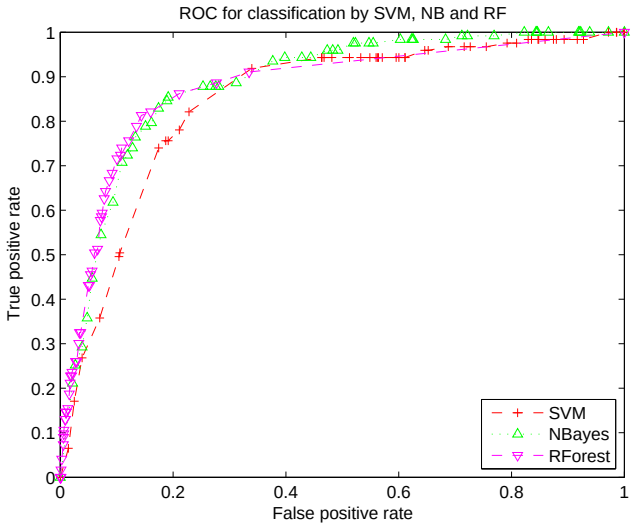
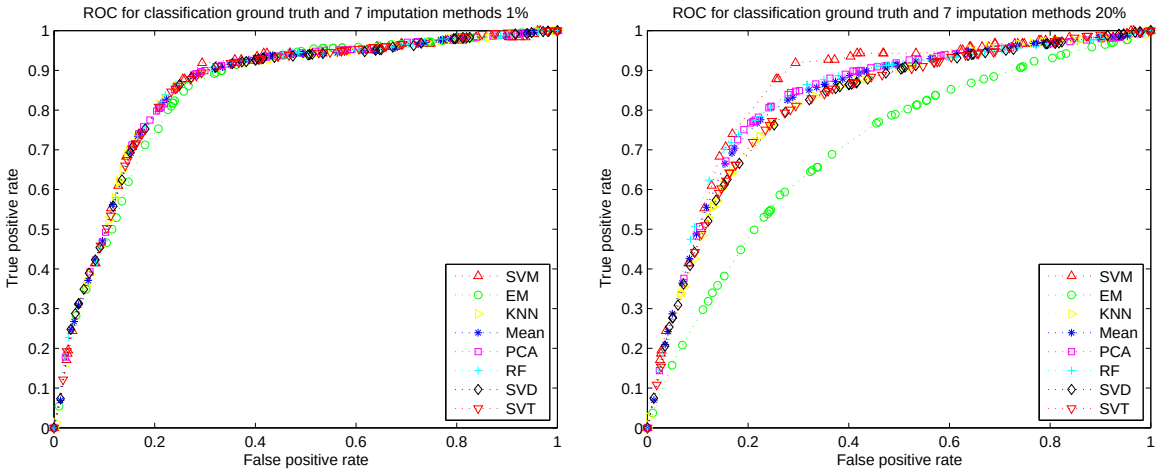


FIGURE 8. Ground truth classification results using CW Support Vector Machines, CS Naïve Bayes and CS Random Forest

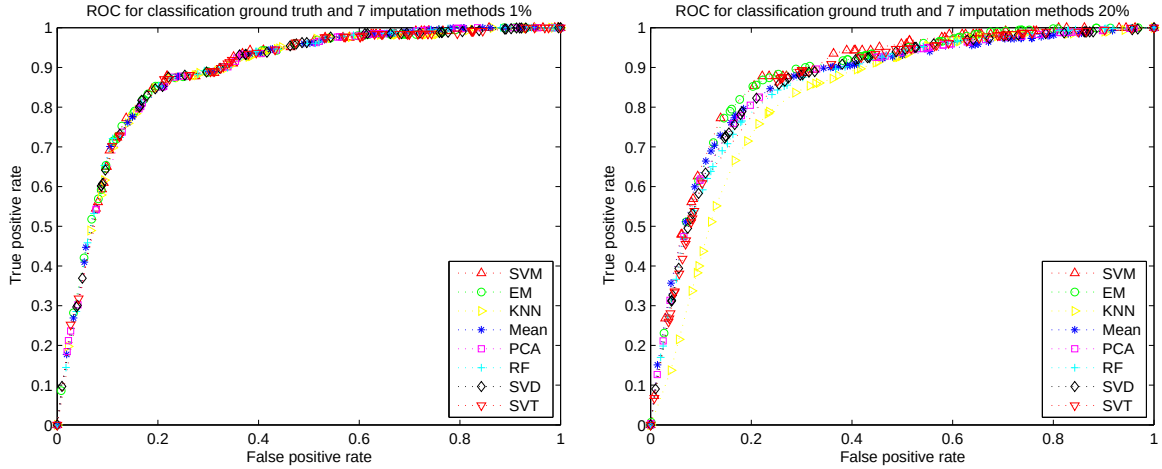


(a) 1% missing (b) 20% missing

FIGURE 9. Classification using CW Support Vector Machines and 7 imputation schemes at 1 and 20% missingness

that RF outperforms the other classifiers (marginally when comparing to CS NB). The analysis of the imputed dataset, with respect to the classifier performance follows.

The results containing the comparison between the ground truth CW SVM along with the 7 imputed datasets are presented by Figure 9. This figure gives the results for 1% and 20% missingness levels (extreme cases of missingness). Figure 9(a) shows that at 1% level of missingness, there is very little difference between the ground truth and the imputed datasets. The imputed sets closely match the ground truth with possibly the exception being EM imputation which is slightly below the others (from 9% to 20% FP rate). The similarity of the results of imputed data compared to the ground truth is to be (intuitively) expected since there is only a small amount of data that was estimated. Also, with respect to the similarity results in the previous subsection, using most of the distance metrics (with the exception of possibly the DAC, Dice and Motyka measures) suggests



(a) 1% missing (b) 20% missing

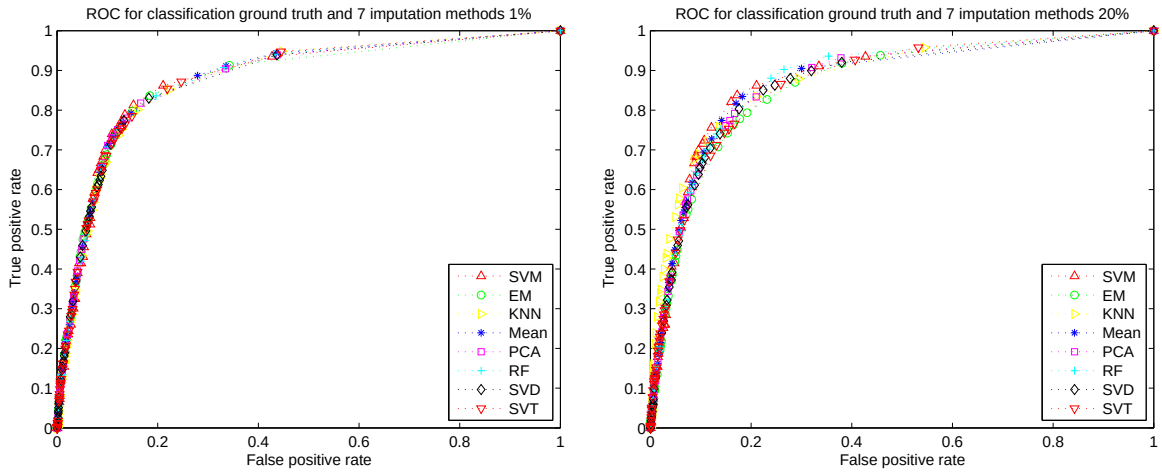
FIGURE 10. Classification using CS Naïve Bayes and 7 imputation schemes at 1 and 20% missingness

that the classification results should not differ drastically. Using 20% missingness, Figure 9(b) shows more deviation from the ground truth. In certain regions between 0 and 20% FP rate, the curve for RF imputation lies above ground truth but when considering a TP is greater than 70%, in the same region, the two curves converge. All other curves show slightly weaker performance with the worst being EM with a 43% TP rate at 20% FP. This, however, cannot be necessarily be justified by the results of similarity measures.

The CS Naïve Bayes classification comparison is presented in Figure 10. The results show that at 1% missingness, all the imputed data ROC curves in Figure 10(a) are overlapping the ground truth NB ROC curve. The similarity to the ground truth of the imputed datasets, using this classifier, is greater than that of CW SVM (see Figure 9(a)). The case where 20% missingness is considered, the imputed curves deviate slightly in the region of interest (FP $\in [0, 20]$). EM tracks the ground truth curve closely when TP rate is above 70% in the corresponding region. The other curves are below EM and the worst performing scheme is kNN imputation. A point to note is that at the 20% level, the worst performing imputation scheme shows greater accuracy than using some imputed datasets along with the CW SVM classifier (see Figure 9(b)). Out of the seven similarity measures, the Squared Euclidean distance (Figure 4) median scores could possibly explain the results in the above figure.

The analysis using the CS Random Forest classifier is outlined by Figure 11. Similar behavior is seen in Figure 11(a) for 1% missingness as with the CS NB and CW SVM. The ROC curves for the imputed datasets overlap the ground truth data. Comparing the ROC curves in the 20% missingness case, it can be seen that (in the region of interest $[0, 20\%]$ FP rate) the ground truth is above all the curves when TP is above 70%. SVT and EM lie below all other ROC curves in the region but achieve superior performance as compared to kNN and EM in Figures 10(b) and 9(b) respectively.

A summary of classification results follows. The imputed datasets at a 1% level of missingness shows very little deviation from the ground truth ROC curves. Figures 9(a), 10(a) and 11(a) highlight this assertion. This behavior is expected since the data missing is almost negligible. The similarity measures which capture this expected behavior are the Squared Euclidean and Manhattan distance metrics. The ROC curves of the imputed data using CS NB and CS RF are identical to the ground truth. This also shows that



(a) 1% missing (b) 20% missing

FIGURE 11. Classification using CS Random Forest and 7 imputation schemes at 1 and 20% missingness

one imputation scheme cannot be preferred to another. The same cannot be said when analyzing Figures 9(b), 10(b) and 11(b). At 20% level of missingness, EM imputation suffers greatly when using CW SVM. However, there is very little evidence from the similarity measure to explain the performance. Using the CS NB classifier, the kNN imputation produces the least satisfactory performance. Even so, it achieves greater accuracy than EM imputation using the CS SVM classifier. The CS RF classification shows the least amount of deviation among the ROC curves while maintaining satisfactory results. From the DAC measure, the expectation is that the SVT imputed data would result in markedly different classification performance. Instead the SVT ROC curve (in Figure 11(b)) fared slightly better than EM. It can therefore be concluded that, to some extent, the quality of imputation at all levels of missingness does not significantly affect classifier performance using CS NB and RF. Therefore, imputation is justifiable in the domain of FSF detection and should not be overlooked to include instances which contain missing values.

6. **Conclusion.** The objective of the research was to investigate the impact/role of imputation using authentic financial statement fraud data. Two approaches were considered in order to measure the effect of imputation. Firstly, seven imputation techniques were utilized to measure the quality of the imputed data with respect to the ground truth dataset. This was performed through the use of seven similarity measures. The second approach comprised of using three cost-sensitive classification techniques on imputed datasets to test predictive performance. With respect to the quality of missing value estimation, EM imputation generally produced the least amount of variation using the seven similarity measures. PAC and DAC showed that PCA imputation achieved the lowest median values. The Lorentzian distance was the least informative metric with respect to which imputation outperformed others using median and standard deviation scores. The Squared Euclidean and Manhattan distance generally favor EM and Mean imputation with respect to median and variation. The Dice distance results show that RF imputation achieves lower scores and standard deviation for three missingness levels. Mean imputation is closest to ground truth with respect to the Motyka distance. The predictive ability of the imputed data was measured using class-weighted Support Vector Machines,

cost-sensitive Naïve Bayes and cost-sensitive Random Forests. For all classifiers, at 1% missingness, the imputed datasets ROC curves mirror (closely) that of the ground truth dataset. This is more obvious when using CS NB and CS RF. For the extreme case, 20% missingness, the imputed datasets' ROC curves deviate from the ground truth, especially with respect to the CW SVM classifier. The CS RF classifier exhibits the least amount of variation at the largest missingness level which shows stability as missingness is increased.

The results have shown that imputation has a potential to play a pivotal role when predictive accuracy is of utmost importance in the field of FSF detection. Instead of removing instances with missing data altogether, imputation can be used. This is especially critical when auditors or practitioners encounter with instances (especially of the minority class) with incomplete data. Imputation (at different levels of missingness) along with CS RF can be a valid solution to deploy a classifier as part of the Analytical Procedures auditing requirement.

This study is seen as initial research when encountering datasets with missing values in the field of financial statement fraud detection. There were some limitations which could be addressed by future work. Parameter values of imputation schemes may need to be varied instead of using default values. Also, the case where missingness is not entirely random needs to be investigated, i.e., where certain features contain more missing values then others. Other possible extensions of the work can include using variables (financial ratios) for which the ground truth is unknown and analyzing predictive accuracy. The impact of imputation on variable importance may also be investigated. This may shed some more light with respect to how much 'noise' or bias the imputation techniques introduce into the data.

Acknowledgment. The current work is being supported by the Department of Science and Technology (DST) and Council for Scientific and Industrial Research (CSIR). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation of this paper.

REFERENCES

[1] H. Öğüt, R. Aktaş, A. Alp and M. M. Doğanay, Prediction of financial information manipulation by using support vector machine and probabilistic neural network, *Expert Systems with Applications*, vol.36, no.3, pp.5419-5423, 2009.

[2] W. Zhou and G. Kapoor, Detecting evolutionary financial statement fraud, *Dezhoucision Support Systems*, vol.50, no.3, pp.570-575, 2011.

[3] Z. Rezaee, Causes, consequences, and deterence of financial statement fraud, *Critical Perspectives on Accounting*, vol.16, no.3, pp.277-298, 2005.

[4] International Federation of Accountants, *Handbook of International Quality Control, Auditing, Review, Other Assurance, and Related Services Pronouncements*, 2010.

[5] K. A. Kaminski, T. S. Wetzel and L. Guan, Can financial ratios detect fraudulent financial reporting? *Managerial Auditing Journal*, vol.19, no.1, pp.15-28, 2004.

[6] E. Kirkos, C. Spathis and Y. Manolopoulos, Data mining techniques for the detection of fraudulent financial statements, *Expert Systems with Applications*, vol.32, no.4, pp.995-1003, 2007.

[7] J. Luengo, S. García and F. Herrera, On the choice of the best imputation methods for missing values considering three groups of classification methods, *Knowledge and Information Systems*, vol.32, no.1, pp.77-108, 2012.

[8] P. J. García-Laencina, J.-L. Sancho-Gómez and A. R. Figueiras-Vidal, Pattern classification with missing data: A review, *Neural Computing and Applications*, vol.19, no.2, pp.263-282, 2010.

[9] K. I. Chang, K. W. Bowyer and P. J. Flynn, Multimodal 2D and 3D biometrics for face recognition, *IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, pp.187-194, 2003.

[10] A. Lazarevic, L. Ertöz, V. Kumar, A. Ozgur and J. Srivastava, A comparative study of anomaly detection schemes in network intrusion detection, *SDM*, pp.25-36, 2003.

[11] N. Powell, S. Y. Foo and M. Weatherspoon, Supervised and unsupervised methods for stock trend forecasting, *The 40th Southeastern Symposium on System Theory*, pp.203-205, 2008.
[12] M.-M. Deza and E. Deza, *Dictionary of Distances*, Elsevier, 2006.
[13] S. D. Bharkad and M. Kokare, Performance evaluation of distance metrics: Application to fingerprint recognition, *International Journal of Pattern Recognition and Artificial Intelligence*, vol.25, no.6, pp.777-806, 2011.
[14] Y. Sun, U. Braga-Neto and E. R. Dougherty, Impact of missing value imputation on classification for DNA microarray gene expression data: A model-based study, *EURASIP Journal on Bioinformatics and Systems Biology*, 2009.
[15] M. Doumpos, C. Gaganis and F. Pasiouras, Explaining qualifications in audit reports using a support vector machine methodology, *Intelligent Systems in Accounting, Finance and Management*, vol.13, no.4, pp.197-215, 2005.
[16] Q. Deng, Application of support vector machine in the detection of fraudulent financial statements, *The 4th International Conference on Computer Science & Education*, pp.1056-1059, 2009.
[17] X. Li and S. Ying, Lib-svms detection model of regulating-profits financial statement fraud using data of Chinese listed companies, *International Conference on E-Product E-Service and E-Entertainment (ICEEE)*, pp.1-4, 2010.
[18] M. Cecchini, H. Aytug, G. J. Koehler and P. Pathak, Detecting management fraud in public companies, *Management Science*, vol.56, no.7, pp.1146-1160, 2010.
[19] M. D. Beneish, The detection of earnings manipulation, *Financial Analysts Journal*, vol.55, no.5, pp.24-36, 1999.
[20] C.-C. Lin, A.-A. Chiu, S. Y. Huang and D. C. Yen, Detecting the financial statement fraud: The analysis of the differences between data mining techniques and experts' judgments, *Knowledge-Based Systems*, 2015.
[21] H. A. Ata and İ. H. Seyrek, The use of data mining techniques in detecting fraudulent financial statements: An application on manufacturing firms, *The Journal of Faculty of Economics and Administrative Sciences*, vol.14, no.2, pp.157-170, 2009.
[22] P.-F. Pai and M.-F. Hsu, An enhanced support vector machines model for classification and rule generation, *Computational Optimization, Methods and Algorithms*, pp.241-258, 2011.
[23] Y.-I Lou and M.-L. Wang, Fraud risk factor of the fraud triangle assessing the likelihood of fraudulent financial reporting, *Journal of Business & Economics Research (JBER)*, vol.7, no.2, 2011.
[24] O. S. Persons, Using financial statement data to identify factors associated with fraudulent financial reporting, *Journal of Applied Business Research (JABR)*, vol.11, no.3, pp.38-46, 2011.
[25] I. Amara, A. B. Amar, A. Jarboui et al., Detection of fraud in financial statements: French companies as a case study, *International Journal of Academic Research in Accounting, Finance and Management Sciences*, vol.3, no.3, pp.40-51, 2013.
[26] S. Kotsiantis, E. Koumanakos, D. Tzelepis and V. Tampakas, Forecasting fraudulent financial statements using data mining, *International Journal of Computational Intelligence*, vol.3, no.2, pp.104-110, 2006.
[27] P. Ravisankar, V. Ravi, G. R. Rao and I. Bose, Detection of financial statement fraud and feature selection using data mining techniques, *Decision Support Systems*, vol.50, no.2, pp.491-500, 2011.
[28] C. Gaganis, Classification techniques for the identification of falsified financial statements: A comparative analysis, *Intelligent Systems in Accounting, Finance and Management*, vol.16, no.3, pp.207-229, 2009.
[29] J. Perols, Financial statement fraud detection: An analysis of statistical and machine learning algorithms, *Auditing: A Journal of Practice & Theory*, vol.30, no.2, pp.19-50, 2011.
[30] P.-F. Pai, M.-F. Hsu and M.-C. Wang, A support vector machine-based model for detecting top management fraud, *Knowledge-Based Systems*, vol.24, no.2, pp.314-321, 2011.
[31] R. Gupta and N. S. Gill, Prevention and detection of financial statement fraud – An implementation of data mining framework, *Editorial Preface*, vol.3, no.8, 2012.
[32] B. Hoogs, T. Kiehl, C. Lacomb and D. Senturk, A genetic algorithm approach to detecting temporal patterns indicative of financial statement fraud, *Intelligent Systems in Accounting, Finance and Management*, vol.15, nos.1-2, pp.41-56, 2007.
[33] M. L. Roxas, Financial statement fraud detection using ratio and digital analysis, *Journal of Leadership, Accountability, and Ethics*, vol.8, no.4, pp.56-66, 2011.
[34] C. D. Katsis, Y. Goletsis, P. V. Boufounou, G. Stylios and E. Koumanakos, Using ants to detect fraudulent financial statements, *Journal of Applied Finance and Banking*, vol.2, no.6, pp.73-81, 2012.

[35] J. L. Schafer and J. W. Graham, Missing data: Our view of the state of the art, *Psychological Methods*, vol.7, no.2, p.147, 2002.
[36] R. O. Duda, P. E. Hart and D. G. Stork, *Pattern Classification*, John Wiley & Sons, 2012.
[37] E. Olivas, J. Guerrero, M. Sober, J. Benedito and A. Lopez, Handbook of research on machine learning applications and trends: Algorithms, *Methods and Techniques*, 2009.
[38] P. Jonsson and C. Wohlin, An evaluation of k-nearest neighbour imputation using Likert data, *Proc. of the 10th International Symposium on Software Metrics*, pp.108-118, 2004.
[39] W. C. Regoeczi and M. Riedel, The application of missing data estimation models to the problem of unknown victim/offender relationships in homicide cases, *Journal of Quantitative Criminology*, vol.19, no.2, pp.155-183, 2003.
[40] J.-F. Cai, E. J. Candès and Z. Shen, A singular value thresholding algorithm for matrix completion, *SIAM Journal on Optimization*, vol.20, no.4, pp.1956-1982, 2010.
[41] L. Breiman, Random forests, *Machine Learning*, vol.45, no.1, pp.5-32, 2001.
[42] S.-H. Cha, Comprehensive survey on distance/similarity measures between probability density functions, *City*, vol.1, no.2, p.1, 2007.
[43] S.-H. Cha, Taxonomy of nominal type histogram distance measures, *City*, vol.1, no.2, p.1, 2008.
[44] P. J. F. Groenen and K. Jajuga, Fuzzy clustering with squared Minkowski distances, *Fuzzy Sets and Systems*, vol.120, no.2, pp.227-237, 2001.
[45] S. O. Moepya, S. S. Akhoury, F. V. Nelwamondo and B. Twala, Measuring the impact of imputation in financial fraud, *Computational Collective Intelligence*, pp.533-543, 2015.
[46] H. He, E. Garcia et al., Learning from imbalanced data, *IEEE Trans. Knowledge and Data Engineering*, vol.21, no.9, pp.1263-1284, 2009.
[47] B. Zadrozny and C. Elkan, Learning and making decisions when costs and probabilities are both unknown, *Proc. of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.204-213, 2001.
[48] C. Elkan, The foundations of cost-sensitive learning, *International Joint Conference on Artificial Intelligence*, vol.17, pp.973-978, 2001.
[49] S. O. Moepya, S. S. Akhoury and F. V. Nelwamondo, Applying cost-sensitive classification for financial fraud detection under high class-imbalance, *IEEE International Conference on Data Mining Workshop (ICDMW)*, pp.183-192, 2014.
[50] C.-W. Hsu, C.-C. Chang, C.-J. Lin et al., *A Practical Guide to Support Vector Classification*, 2003.
[51] J. Honaker, G. King, M. Blackwell et al., Amelia II: A program for missing data, *Journal of Statistical Software*, vol.45, no.7, pp.1-47, 2011.
[52] L. Breiman, Randomforest: Breiman and Cutler's random forests for classification and regression, *R News*, 2006.
[53] N. L. Crookston, A. O. Finley et al., Yaimpute: An R package for kNN imputation, *Journal of Statistical Software*, vol.23, no.10, pp.1-16, 2008.
[54] M. Robnik-Sikonja, P. Savicky and M. M. Robnik-Sikonja, *Package CORElearn*, 2013.
[55] M. Kuhn, *The Caret Package*, 2012.
[56] E. Acuna and C. Rodriguez, The treatment of missing values and its effect on classifier accuracy, *Classification, Clustering, and Data Mining Applications*, pp.639-647, 2004.
[57] J. Benesty, J. Chen, Y. Huang and I. Cohen, Pearson correlation coefficient, *Noise Reduction in Speech Processing*, pp.1-4, 2009.
[58] H. W. Lilliefors, On the Kolmogorov-Smirnov test for normality with mean and variance unknown, *Journal of the American Statistical Association*, vol.62, no.318, pp.399-402, 1967.
[59] T. Fawcett, An introduction to ROC analysis, *Pattern Recognition Letters*, vol.27, no.8, pp.861-874, 2006.
[60] M. A. Maloof, Learning when data sets are imbalanced and when costs are unequal and unknown, *ICML-2003 Workshop on Learning from Imbalanced Data Sets II*, vol.2, 2003.