Automated identification of fraudulent financial statements by analyzing data traces

Evangelos Hytis
Department of Accounting and Finance
University of Ioannina
Preveza, Greece
Email: ehytis@uoi.gr

Vasileios Nastos, Christos Gogos, Angelos Dimitsas
Department of Informatics and Telecommunications
University of Ioannina
Arta, Greece
Email: {vnastos, cgogos, a.dimitsas}@uoi.gr

2022 7th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM) | 979-8-3503-9858-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/SEEDA-CECNSM57760.2022.9932910

***Abstract*—Firms are obliged by law to publish accurate financial statements. Nevertheless, cases exist where publicly issued documents hide the actual bad financial statuses of companies and this is revealed years later. Since companies publish financial figures periodically, it is interesting to examine whether monitoring those values or ratios based on them can help in early detection of fraud. In this work, a dataset was constructed including 943 firm-years of Greek companies enlisted at the Athens Stock Exchange, for the period 2005-2018. Experiments using combinations of financial ratios and various machine learning algorithms were undertaken in an effort to construct a system able to automatically detect false and misleading financial statements, that were issued legitimately by firms. Several instantiations of a machine learning workflow were tested using various classifier algorithms including AdaBoost, Random Forests, and others. Experiments showed that companies that issue false financial statements can be spotted automatically for most of the cases, years before the problem is manifested. So, the potential of early detection of seemingly healthy, but in fact, distressed companies do exist. An automated tool can be constructed that should be useful for financial analysts, investors and the capital markets authorities.**

***Index Terms*—fraudulent financial statements, financial ratios, machine learning, workflow, Greece**

I. Introduction

Scrutinizing financial documents typically reveals a wealth of information about companies that issue them. Cash flows, operating costs and various financial indicators like EBIT (Earnings Before Interest and taxes), EBT (Earnings Before Tax) and others mirror current and most likely near future financial position of firms. Nevertheless, financial fraud is a real problem, especially in fast growing countries. Several cases of companies that faked their data in an attempt to hide their actual financial state are coming to light with alarming consistency when inevitable events expose the actual dire situation. Post-mortem analysis of such financial documents by analysts, often conclude that signs of financial malpractice existed long before the collapse.

ML (Machine Learning) has the ability of discovering previously unknown and perhaps useful patterns in data sets. The aim of this study is the detection of management fraud by applying machine learning. In order to detect fraud in financial statements a carefully constructed dataset of 124 Greek companies was examined. A number of machine learning algorithms were tested against this dataset.

The structure of this paper is the following. The two next sections present the problem's description and related work. Section IV presents the fraud risk indicators that were used. The next section describes the dataset acquisition and construction procedure. Section VI describes the machine learning workflow, that includes load, pre-process, learn and evaluation stages. Then, results of the experiments undertaken are presented in Section VII, and the paper closes with conclusions.

II. Problem description

Previous research indicates that fraud remains one of the most unsolved problems in world business and especially in countries with insufficient supervisory mechanisms and a low level of investor protection [1]. In the Oxford Dictionary, fraud is defined as "a false representation by means of a statement or conduct, in order to gain a material advantage", while the International Auditing and Assurance Standards Board (IAASB) of the International Federation of Accountants (IFAC) defines fraud as "an intentional act by one or more individuals among management, those charged with governance, employees, or third parties, involving the use of deception to obtain an unjust or illegal advantage".

There are many examples of companies that have issued fraudulent financial statements (FFS) and at a later time were exposed. Often, such situations reach the headlines of media and are considered extremely negative events since the business of stakeholders and the lives of many people are severely damaged. A prominent example of FFS is the ENRON scandal in USA at 2001. Moreover, in 2002 WorldCom, a telecommunications company got caught inflating assets by an amount of 11 billion $. At the same year, Tyco's executives stole 150 million $, giving themselves loans, and inflated the company's income. In 2009, an Indian IT services and back-office accounting firm, Satyam, boosted its revenues by 1.5 billion $. Accounting fraud scandals, like these, occurred in the last twenty years, and burned billions of dollars. In Greece, which is the country that our study focuses, the Folli-Follie jewelry company was exposed at 2018 for issuing FFS.

Financial statement fraud can take multiple forms, including overstating revenues by recording future expected sales,

979-8-3503-9858-8/22/$31.00 ©2022 IEEE

inflating an asset's net worth by knowingly failing to apply an appropriate depreciation schedule, hiding obligations and/or liabilities from a company's balance sheet, and incorrectly disclosing related-party transactions and structured finance deals. Another type of financial statement fraud involves cookie-jar accounting practices, where firms understate revenues in one accounting period and maintain them as a reserve for future periods with worse performances, in a broader effort to temper the appearance of volatility.

Several incentives for issuing FFSs can be identified [2]. Among them key motives for falsifying financial statements are the following:

- Incentives derived from the operation of the capital market such as pressure from financial analysts, raising capital from the Stock Exchange, borrowing from banks and other credit institutions, mergers and acquisitions and dividend policy.
- Incentives derived from contractual obligations of the company such as loan agreements and management fees.
- Incentives related to the behavior of management members such as retention of managerial positions, promotion in hierarchies, and others.
- Incentives related to the regulatory framework in which firms operate such as the regulatory framework of the industry to which they belong, antitrust and other regulations and tax avoidance efforts.
- Short-term incentives stemming from the firm's operational culture orientation, such as realistic budgets, action plans and others.

A crucial part of fighting fraud in financial reporting is the identification of suitable indicators (red flags) and the construction of automated systems capable of detecting false and misleading financial statements. So, stakeholders (internal or external) have much interest in methods based on statistical and computational intelligence for fraud detection [3], [4], [5].

III. Related work

Many researchers have tried to address the problem of early fraud detection based on analyzing issued financial statements. A comprehensive review about financial fraud in general can be consulted in [6]. In [7] authors used profitability, liquidity, efficiency and cash-flow variables in order to detect financial distress for 164 manufacturing companies in Greece, for the period 2001-2002. Firstly, ReliefF [8] was used in order to find the statistical significant features. Then, 10-fold cross-validation was performed trying logistic regression, decision trees (C4.5), artificial Neural Networks (ANNs) and bayesian networks. The decision trees achieved the best results.

Financial ratios from 76 Greek manufacturing firms, where used in [9]. Among them, 38 firms issued FFS. An approach similar to [7] was employed with the difference that the ANOVA statistical test was employed and that the experiments were executed using a) training and test sets and b) 10-fold cross-validation. For the first case Neural Networks achieved the best results, while for 10-fold cross-validation the Bayesian Belief Network returned the best results.

IV. Fraud risk indicators

Relevant literature suggests that financial statements in general and financial ratios computed from financial statements in particular constitute information of potentially great value, [10], [11], [12], [13]. Meticulous analysis of financial ratios is believed to be capable of identifying fraudulent behavior, within reasonable limits.

Focusing on Greece, several studies exist that use financial ratios in an effort to identify financial fraud [14], [15], [16], [17]. The majority of these studies, up to fiscal year 2014, use financial statements that were submitted according to the National (Greek) Accounting Standards (EL-EGLS). However, their philosophy and accounting principles and rules have significant differences and deviations from those of the International Financial Reporting Standards (IAS/IFRS). IAS/IFRS are mandatory in the Greek stock exchange for all listed groups from fiscal year 2015 and thereafter. This transition, affects the comparability of financial statement figures and ratios, and should be handled carefully for studying periods that reside both in pre IFRS and post IFRS eras.

In this work, a set of variables that correspond to financial figures and ratios were selected as representative of the status of each company. It should be noted that intra-group accounting transactions are excluded, due to the IFRS Consolidated Financial Statements that are used. Consequently, the financial statuses of the companies are captured more accurately. Finally, 29 variables were selected. Among them 7 are quantities that are directly included in financial statements (F1-F7) and 22 are financial ratios (R1-R22). Variables F1 to F4 refer to the degree of asset capitalization, while variables F5 to F7 measure the company's ability to generate cash-flow. On the other hand the financial ratios are separated in four groups, namely profitability ratios, capital structure ratios, liquidity ratios and activity ratios. Table IV summarizes all variables used.

V. Dataset acquisition and construction

For this work, data were collected from the certified financial database ICAP Data-Prisma, the Athens Exchange (Hellex SA) and the Hellenic Capital Market Commission (HCMC). Initially, 148 company groups were identified. All of them were listed at the Athens Exchange and corresponded to 1028 firm-years spanning from 2005 to 2018. It should be noted that the study period includes the financial crisis in Greece that started at early 2009 and lasted until late 2018.

Following an approach similar to previous researches like [7], [18], [10] and [11], 24 companies (85 firm-years) from insurance, financial services, real estate and holding sectors were removed from the sample, due to the form of their financial statements which are specialized and the different financial ratios that they use when compared to other types of industries. The resulting sample includes 124 companies (943 firm-years) that followed the consolidated International Financial Reporting Standards (IAS/IFRS) for the period under examination (2005-2018) and that were at the same time audited by certified auditors.

Variable	Name - Description
Assets (intensive)	
F1	Total Assets (Log)
F2	Net Fixed Assets / Total Assets
F3	Tangible Assets / Total Assets
F4	Intangible Assets / Total Assets
Cash Flow Activities	
F5	Inflow from Operating Activities / Net Sales
F6	Inflow from Investing Activities / Total Assets
F7	Inflow from Financing Activities / Total Assets
Profitability Ratios	
R1	Gross Profit Margin
R2	Net Profit Margin (before interest & income tax)
R3	Net Profit Margin (before income tax)
R4	EAT (Earnings After Taxes)
R5	ROA (Return On Assets)
R22	EBITDA (Net Profit Margin)
Capital Structure Ratios	
R6	Equity to Capital Employed
R7	Capital Employed / Net Fixed Assets
R8	Cash Flows From Operations / Debt
R9	Current Liabilities / Inventories
R10	Equity / Total Assets
R11	Total Liabilities / EBITDA
Liquidity Ratios	
R12	Current Ratio (Current Assets / Current Liabilities)
R13	Quick Ratio (ACID Test)
Activity Ratios	
R14	Turover of Capital Employed
R15	Equity Turnover
R16	Collection Period (Days)
R17	Payable Period (Days)
R18	Inventory Turnover (Days)
R19	Inventories / Total Assets
R20	Turnover of Total Assets
R21	Turnover of Fixed Assets

TABLE I
VARIABLES USED FOR FRAUD DETECTION

For 51 of these companies (206 firm-years), evidences suggest that Fraudulent Financial Statement (FFS) were issued. Consequently, they have been classified as such. Such evidence were taken from published surveys by supervising authorities, opinions and observations of Certified Public Accountants, decisions and announcements of the Stock Exchange and the Capital Market Commission (e.g. inclusion in surveillance, suspension of trading, deletion of companies, recommendations to the investing public, etc.), and decisions of judicial and tax authorities and creditors. Table II provides the numerical data about the excluded firms and firm-years in our data.

Excluded ICB Code	Excluded Sector	Firms	Firm-Years
85	Insurance	1	10
86	Real estate	3	32
83 & 87	Financial services	5	22
88	Holding	15	21
		24	85

TABLE II
EXCLUDED FIRMS AND FIRM-YEARS

The criteria used for classification as FFS, derive from the provisions of the International Auditing Standard (AICPA ISA), and are in agreement with those used in earlier studies including [14], [19], [12], [13] and others. Table III presents observations data by industrial sector, alongside with the numbers of firms and firm-years classified either as fraud or as non-fraud.

Industrial sector	ICB-Code	Firms	%	Non-FFS	FFS	FYs	%	Non-FFS FYs	FFS FYs
Chemicals	13	3	2.4%	1	2	40	4.2%	34	6
Basic Resources & Construction and Materials	17 & 23	17	13.7%	8	9	162	17.2%	128	34
Industrial Goods & Services	27	42	33.9%	31	11	251	26.6%	209	42
Food, Beverage & Tobacco	35	11	8.9%	6	5	101	10.7%	82	19
Personal Care, Drug & Grocery Stores	37	10	8.1%	3	7	76	8.1%	50	26
Health Care	45	6	4.8%	3	3	62	6.6%	46	16
Retail	53	13	10.5%	9	4	45	4.8%	36	9
Media	55	8	6.5%	4	4	74	7.8%	59	15
Travel and Leisure	57	8	6.5%	5	3	62	6.6%	46	16
Utilities	75	2	1.6%	2	0	22	2.3%	22	0
Technology	95	4	3.2%	1	3	48	5.1%	25	23
Total		124	100%	73	51	943	100%	737	206
								78%	22%

TABLE III
FIRMS BY INDUSTRIAL SECTOR AND FFS CLASSIFICATION (FY STANDS FOR FIRM-YEAR)

VI. MACHINE LEARNING WORKFLOW

The problem is a classic binary classification one, with data consisting of rows that represent either fraud or non fraud cases. We applied a machine learning workflow that is presented in Figure 1. K-fold cross validation was applied to the data in order to lower the bias. This procedure randomly divides the data in k folds (parts). Then a model is fitted, to the k-1 folds, that are in effect the training set, and the remaining fold is used for testing. The procedure repeats k times, so each fold has the opportunity to be used as the test set. The final generalization (out-of-sample) error is calculated by averaging error values of each run. We used 10 as the value of k, since this is the commonly used value for k, and it produced good results in our problem. Since, the dataset contained missing values, several methods for handling them were applied. Furthermore, we identified scaling issues to the data. It is well known that ML algorithms work better when feature data are on relatively similar scale. So, we normalized the data. Then, the important phase of feature selection follows. In our case we included 31 input variables. By reducing the number of input variables, we expect an improved performance. Moreover, the computational cost is reduced which is minor for our case, due to the relatively small problem size. We observed that the data were unbalanced. About 78% of the rows were labeled as non-fraud, while only the remaining 22% were labeled otherwise. So, we balanced fraud and non-fraud rows by introducing a balancing phase that either increased the size of the minority class, or decreased the size of the majority class. This concludes the pre-processing stage. Then, we systematically employed numerous classification algorithms and ensembles that resulted in the construction of several competitive machine learning models. Those models were evaluated through standard metrics (accuracy, recall, F-measure, precision and Cohen's kappa). Some models performed significantly better than others. Details of the stages presented in Figure 1 follows. Finally, results of the computation experiments are presented in Section VII.

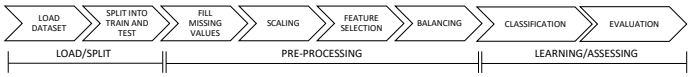

Fig. 1. ML workflow

A. Preprocessing

Preprocessing performs data manipulation that enhances the performance of the machine learning algorithms. In our problem, four steps were applied that are described in the following sub-sections.

1) Missing values: No missing values must be present in the dataset that will be fed to the ML algorithm. Several approaches for handling missing values exist. We tried four different approaches, drop, mean, median and k-NN impute. The first one simply drops rows that have any none values. In our case, since the majority of rows contained at least one missing value, the drop method was not further considered. The mean approach replaces each missing value with the mean value of the corresponding column. Likewise, median replaces missing values with the median of each column. The last approach was k-NN which is more involved that the previous ones since it replaces each missing value with the mean value of the k nearest neighbors discovered in the training set. A common used values for k that was used in our case is 10.

2) Scaling: Initially, we approached the scaling issues that existed in our data using two standard methods, normalization and standardization. In normalization, a.k.a. min-max scaling, scales values to the range 0 to 1. Likewise, in standardization, a.k.a. z-score normalization, values become centered around the mean which assumes value of 0, while the resulting distribution has a unit of standard deviation. Also, two other scaling techniques were used, maximum absolute scaling and robust scaling. The first one removes scaling issues while maintaining potential sparsity that exists in the data. It scales each feature in isolation in the range -1 to 1. On the other hand, robust scaling focuses on removal of the outliers. It computes quartile statistics and all values are squeezed to the interquartile range (IQR).

3) Feature selection: Feature selection is very important in machine learning. Ideally, a feature selection approach should keep the most relevant features and facilitate machine learning models. Several options exist and we have experimented with five different ones. Initially, we tried Variance Threshold feature elimination. This simple approach eliminates features that their variance differ less than a given threshold. Then we tried the Relief algorithm, which estimates how well a given attribute can distinguish among rows that are similar. A distinctive characteristic of the algorithm is that it can identify the most important features even in the presence of highly correlated attributes. The other feature selection approaches that we experimented with were wrapper-type approaches. This means that a machine learning algorithm (estimator) decides about the features that will prevail by optimizing a certain performance measure of the algorithm. Three such algorithms were employed, Recursive Feature Elimination (RFE), Backward Feature Elimination (BFE) and Forward Feature Selection(FFS) [20]. RFE examines recursively sets of features in a process that finally converges to a set with the desirable size. In our case the estimator used was the Random Forest algorithm. On the other hand, FFS starts with an empty set of features and in turn probes all available features for inclusion to the set. Based on a chosen evaluation metric (e.g. recall) the algorithm decides whether the feature will be selected, which is the case if the evaluation metric is improved. BFE works in the opposite direction. It starts with the full set of features and sequentially examines all of them for elimination, which occurs if the evaluation metric is improved by excluding the candidate feature.

a) Dimensionality reduction: It should be noted that another approach for identifying the most relevant features is Principal Components Analysis (PCA), which is a statistics based technique that attacks the "curse of dimensionality" problem (more features, less predictive capability). PCA is an established method but in our case, better results were obtained by the previously mentioned feature selection approaches.

4) Balancing: Machine learning algorithms tend to underperform when the input data are unbalanced. In our problem we used three balancing methods that made representation of fraud and non-fraud cases equally likely. We experimented with random undersampling, random oversampling and SMOTE [21]. Random undersampling picks and eliminates instances of the minority class in the training set. On the other hand, random oversampling constructs the training set by randomly selecting instances, while ensuring that data are balanced. Finally, the third option, SMOTE which stands for Synthetic Minority Oversampling Technique creates new instances that are interpolated based on existing data instances. The minority class is augmented with new synthetic data until its size equals the size of the majority class. It should be noted that only the training data were balanced.

B. Classification

A wealth of classification algorithms exists for binary classification problems. Due to the No Free Lunch theorem [22] we expect no single algorithm to be the better performing one for all problems and all problem inputs. So, we created a pool of well-known algorithms and experimented with our preprocessed input data. In particular the algorithms we used were Decision Trees, Support Vector Machines, k-Nearest Neighbors and Naive Bayes. Moreover, we included in our experiments two ensemble machine learning algorithms, Random Forests and AdaBoost. A brief description of the classifiers follows.

Decision tree is a fundamental machine learning algorithm [23], that based on training data gradually forms a tree by best splitting results according to the homogeneity of the produced splits. Adding more layers to the decision tree usually increases the prediction accuracy but at the risk of overfitting. A nice characteristic of the algorithm is that it can explain in human readable terms the reason why new input data are classified the way they are.

Support Vector Machine algorithm works in an n-dimensional space, where n is the number of features [24]. It detects hyperplanes of maximal margin, that split space into k classes. Samples that are nearest to each hyperplane are identified as Support Vectors and are used to categorize new samples to the proper class.

k-NN [25] is a simple learning algorithm that classifies test instances based on the class that the majority of their k nearest neighbors belongs. It is a "lazy-learner" algorithm, since it doesn't formulate a function that subsequently performs the discrimination, but instead it just memorizes the dataset itself.

Naive Bayes is based on the famous Bayes's theorem [26], focusing on the idea that features should be independent of each other. Since, in reality, features can be correlated with each other the assumption that the algorithm makes is naive in some sense. An advantage of the algorithm is that it runs in linear time instead of the higher order time that other learning algorithms run.

Random Forest [27] is an ensemble of decision trees that are typically trained using the bagging method [28] that essentially is driven by the "wisdom of crowds" concept. Several decisions trees that should have low correlations among each other contribute to the formation of a better learning agent, with lower error rates.

AdaBoost [29] is a meta-algorithm that uses other classification algorithms and through a procedure of putting emphasis in misclassified training set samples recursively produces an effective learning model. In particular, each training set sample is given a weight which is raised when misclassification occurs. In practice, AdaBoost usually returns very good results for classification problems and this was the case for our problem too.

C. Evaluation

A typical set of metrics that is used for evaluating the performance of the learning models are the confusion matrix, accuracy, recall, precision, F-measure and Cohen's kappa-coefficient. An algorithm might excel in a metric but might underperform in another. So, the user should have good understanding of the trade-offs among them.

Let TP (True Positive) be the number of positive samples that are also classified as positive and TN (True Negative) be the number of negative samples that are also classified as negative. Let FP (False Positive) be the number of negative samples mistakenly classified as positive, and FN (False Negative) be the number of positive samples mistakenly classified as negative. Then, accuracy is the percentage of correctly classified samples, given in Equation 1. Recall is the true positive rate (TPR), Equation 3, while precision is the false negative rate (FNR), Equation 2. F-measure provides a value that shows how favorite is the balance between recall and precision, and is given by Equation 4. Cohen's kappa-coefficient is a metric that uses as focal points, reliability and validity. Validity estimates the accuracy of the test, while reliability is concerned with the level that the test produces similar results under akin conditions. Cohen's kappa coefficient for binary classification systems is given by Equation 5. Finally, confusion matrix, shown in Table IV, is a numerical tableau depiction of how the model behaves. It consists of n rows and n columns, where n is the number of target classes (n is 2 for our problem) and presents a comparison between the ground truth and the predictions of the model.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F - Measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4}$$

$$\kappa = \frac{2\times(TP\times TN - FN\times FP)}{(TP+FP)\times(FP+TN)+(TP+FN)\times(FN+TN)} \tag{5}$$

	Predicted (Positive=non-fraud)	Predicted (Negative=fraud)
Actual (Positive=non-fraud)	TN	FP
Actual (Negative=fraud)	FN	TP

TABLE IV
CONFUSION MATRIX

VII. EXPERIMENTS

For our experiments we used Python, the package scikit-learn [30], and two extensions to scikit-learn, package imbalanced-learn[1] and package scikit-rebate[2]. The experiments were run in a workstation equipped with an AMD Ryzen 5700G(8C/16T) processor and 32GB of RAM, running Windows 10. A significant number of runs were executed, since we tried all combinations of preprocessing steps and classification algorithms that are described in Section VI. In particular, for 10 folds, we combined three alternatives about handling missing values, four about scaling data, five about feature selection or dimensionality reduction, three alternatives about balancing and six alternatives about classifiers, that totaled to 10800 individual runs and kept our workstation busy for about 12 hours. The metrics that we have collected for each classifier are presented in the boxplots of Figure 2.

In order, to identify the best performing classifiers the pareto front of all runs was computed, against five performance metrics (accuracy, recall, precision, f-measure and Cohen's kappa). The pareto front was computed using python's package OApackage[3] and the non-dominated solutions are presented in Table V. It can be observed that AdaBoost occupies 5 out of 6 entries of the table, while the remaining entry is contributed by the Random Forest classifier.

[1]https://imbalanced-learn.org
[2]https://epistasislab.github.io/scikit-rebate/
[3]http://www.pietereendebak.nl/oapackage/

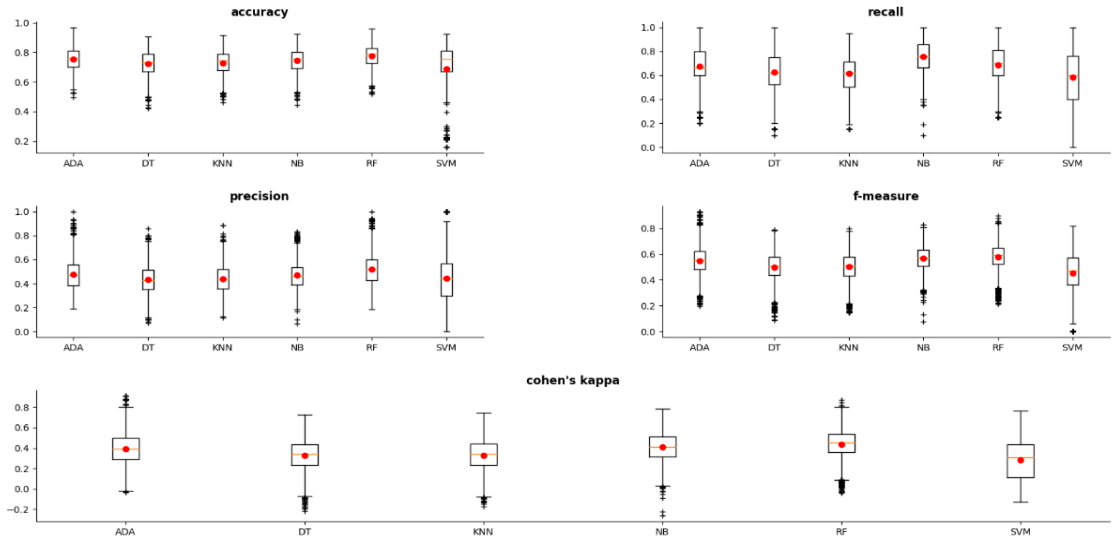

Fig. 2. Evaluation metrics of the classifiers

Missing Values	Scaling	Feature Selection	Balancing	Classifier	Accuracy	Recall	Precision	F-measure	Cohen's kappa
Mean	Min max	Relief	Undersampling	AdaBoost	0.9681	1.0000	0.8696	0.9302	0.9097
Mean	Standardization	Relief	Undersampling	AdaBoost	0.9681	1.0000	0.8696	0.9302	0.9097
Mean	Robust	Relief	Undersampling	AdaBoost	0.9681	0.9500	0.8696	0.9302	0.9097
Mean	Standardization	Relief	SMOTE	AdaBoost	0.9681	0.9500	0.9048	0.9268	0.9064
k-NN impute	Maximum absolute	FFS	Oversampling	AdaBoost	0.9574	0.8000	1.0000	0.8889	0.8630
Mean	Robust	Relief	SMOTE	Random Forest	0.9574	0.8571	0.9474	0.9000	0.8731

TABLE V
PARETO OPTIMAL SOLUTIONS OF ALL RUNS, FOR THE 5 METRICS

Confusion matrices for all six entries of Table V are presented in Figure 4. It should be noted that the high quality results presented by AdaBoost and Random Forest are not guaranteed but very likely to occur and boxplots in Figure 2 seems to support this. Finally, Figure 3 presents the number of workflow instantiations that achieved accuracy over 90%. AdaBoost, once again, has achieved the top classification scores.

VIII. CONCLUSIONS

In this paper we examined a particular case of the problem of identifying fraudulent financial statements issued by companies. We manually created a dataset of 943 firm-years of Greek companies (2005-2018). Each firm-year instance was annotated as either non-fraud or fraud by the first author of this paper after scrutinizing each financial statement, while perusing post-mortem knowledge. Then, a machine learning workflow was setup, that used six different classifiers and several choices of input data preprocessing steps. Using k-fold validation we identified algorithm AdaBoost, for our problem, to be the most performant at achieving high accuracy scores and other performance metrics. Modern tools (e.g. python, machine learning libraries, automation utilities) alongside with field knowledge can be valuable in achieving a level of problem understanding that would otherwise have been possible only under the presence of human experts. Such systems, can be used as a first line of defense against fraud in financial statements. Once a suspected fraud financial statement is identified in this way, further analysis can be ordered that might uncover financial malpractice early, before it escalates to further damages.

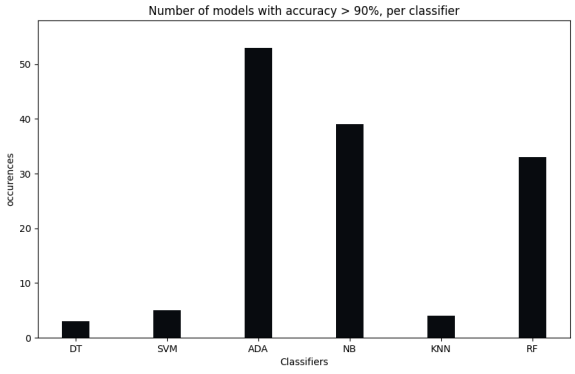

Fig. 3. Comparison of classifiers in achieving high accuracy

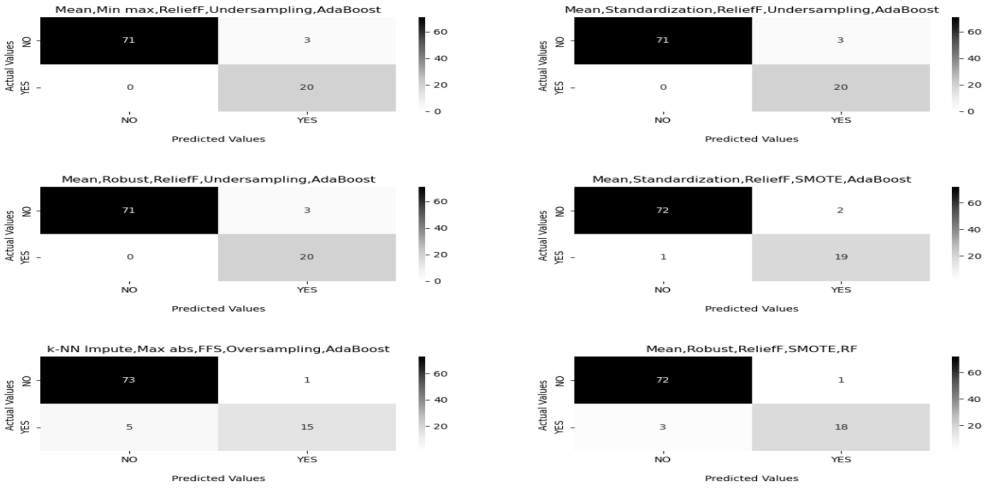

Fig. 4. Confusion matrices of pareto optimal solutions

REFERENCES

[1] Nikolaos I Karampinis and Dimosthenis L Hevas. Mandating IFRS in an unfavorable environment: The Greek experience. *The International Journal of Accounting*, 46(3):304–332, 2011.

[2] Zabihollah Rezaee. Causes, consequences, and deterence of financial statement fraud. *Critical perspectives on Accounting*, 16(3):277–298, 2005.

[3] Michail Pazarskis, George Drogalas, and Kyriaki Baltzi. Detecting false financial statements: Evidence from greece in the period of economic crisis. *Investment management and financial innovations*, (14,№ 3):102–112, 2017.

[4] Eberhard Feess and Yuriy Timofeyev. Behavioral red flags and loss sizes from asset misappropriation: Evidence from the us. In *Advances in Accounting Behavioral Research*. Emerald Publishing Limited, 2020.

[5] Michail Pazarskis, Grigorios Lazos, Andreas Koutoupis, and George Drogalas. Preventing the unpleasant: Fraudulent financial statement detection using financial ratios. *Journal of Operational Risk*, 17(1), 2021.

[6] Khaled Gubran Al-Hashedi and Pritheega Magalingam. Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019. *Computer Science Review*, 40:100402, 2021.

[7] Sotiris Kotsiantis, Euaggelos Koumanakos, Dimitris Tzelepis, and Vasilis Tampakas. Forecasting fraudulent financial statements using data mining. *International journal of computational intelligence*, 3(2):104–110, 2006.

[8] Ryan J Urbanowicz, Melissa Meeker, William La Cava, Randal S Olson, and Jason H Moore. Relief-based feature selection: Introduction and review. *Journal of biomedical informatics*, 85:189–203, 2018.

[9] Efstathios Kirkos, Charalambos Spathis, and Yannis Manolopoulos. Data mining techniques for the detection of fraudulent financial statements. *Expert systems with applications*, 32(4):995–1003, 2007.

[10] Chi-Chen Lin, An-An Chiu, Shaio Yan Huang, and David C Yen. Detecting the financial statement fraud: The analysis of the differences between data mining techniques and experts' judgments. *Knowledge-Based Systems*, 89:459–470, 2015.

[11] Suduan Chen. Detection of fraudulent financial statements using the hybrid data mining approach. *SpringerPlus*, 5(1):1–16, 2016.

[12] Petr Hajek and Roberto Henriques. Mining corporate annual reports for intelligent detection of financial statement fraud–a comparative study of machine learning methods. *Knowledge-Based Systems*, 128:139–152, 2017.

[13] Maria Jofre and Richard H Gerlach. Fighting accounting fraud through forensic data analytics. *Available at SSRN 3176288*, 2018.

[14] S Kotsiantis, D Tzelepis, E Koumanakos, and V Tampakas. Efficiency of machine learning techniques in bankruptcy prediction. In *2nd International Conference on Enterprise Systems and Accounting*, pages 39–49. Citeseer, 2005.

[15] Athanasios Tsagkanos, Antonios Georgopoulos, Dimitrios P Koumanakos, and Evangelos P Koumanakos. Corporate failure risk assessment of greek companies. *International Journal of Risk Assessment and Management*, 9(1-2):5–14, 2008.

[16] Maria Tsipouridou and Charalambos Spathis. Audit opinion and earnings management: Evidence from Greece. In *Accounting Forum*, volume 38, pages 38–54. Elsevier, 2014.

[17] Michail Pazarskis, Manthos Vogiatzoglou, Andreas Koutoupis, and George Drogalas. Corporate mergers and accounting performance during a period of economic crisis: evidence from greece. *Journal of Business Economics and Management*, 22(3):577–595, 2021.

[18] Chrysovalantis Gaganis. Classification techniques for the identification of falsified financial statements: a comparative analysis. *Intelligent Systems in Accounting, Finance & Management: International Journal*, 16(3):207–229, 2009.

[19] Christos D Katsis, Yorgos Goletsis, Paraskevi V Boufounou, George Stylios, and Evangelos Koumanakos. Using ants to detect fraudulent financial statements. *Journal of applied finance and banking*, 2(6):73, 2012.

[20] Nonso Nnamoko, Farath Arshad, David England, Jiten Vora, and James Norman. Evaluation of filter and wrapper methods for feature selection in supervised machine learning. *Age*, 21(81):33–2, 2014.

[21] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[22] David Gómez and Alfonso Rojas. An empirical overview of the no free lunch theorem and its effect on real-world machine learning classification. *Neural computation*, 28(1):216–228, 2016.

[23] Harsh H Patel and Purvi Prajapati. Study and analysis of decision tree based classification algorithms. *International Journal of Computer Sciences and Engineering*, 6(10):74–78, 2018.

[24] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28, 1998.

[25] Zhongheng Zhang. Introduction to machine learning: k-nearest neighbors. *Annals of translational medicine*, 4(11), 2016.

[26] Daniel Berrar. Bayes' theorem and naive bayes classifier. *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, 403, 2018.

[27] Nasiba Mahdi Abdulkareem and Adnan Mohsin Abdulazeez. Machine learning classification based on random forest algorithm: A review. *International Journal of Science and Business*, 5(2):128–142, 2021.

[28] Clifton D Sutton. Classification and regression trees, bagging, and boosting. *Handbook of statistics*, 24:303–329, 2005.

[29] Robert E Schapire. Explaining AdaBoost. In *Empirical inference*, pages 37–52. Springer, 2013.

[30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.