Scientific Review
UDC: 336.717.16
Paper Received: 03/11/2020
Paper Accepted: 04/12/2020

APPLICATION OF DATA MINING TECHNIQUES IN THE DETECTION OF FINANCIAL STATEMENT FRAUD

Dubravka Kopun, PhD
KOPUN Group, Zagreb, Croatia
info@kopun.hr

ABSTRACT

Financial statement fraud is one of three forms of fraud (financial statement fraud, corruption and asset misappropriation) that causes the greatest losses for companies. The increase in the number of financial statement frauds over the past 20 years (Enron, WorldCom, Parlamat, Tesco, Wirecard…) together with the development of data mining techniques, which are a prerequisite for the early detection of fraud, has led to a greater number of studies in this field.

To date, research that has been carried out on the detection of financial statement fraud has not given rise to an unambiguous model; the various studies have suggested the use of different financial analysis indicators and have used different data mining methods, which has led to various levels of success in the detection of financial statement fraud. The reason for this is the fact that these studies have not taken into account changes in the business environment, which have a direct influence on the structure of financial statements, and, by this means, also on the financial analysis indicators. This study shows that focusing on shorter time periods and on companies in identical markets can lead to a better quality model for detecting financial statement fraud.

Key words: *financial statement, fraud, detection of fraud, data mining techniques, financial analysis indicators*

1. DEFINING TERMS AND PREVIOUS RESEARCH

Financial statement fraud implies "the deliberate misrepresentation of the financial condition of an enterprise accomplished through the intentional misstatement or omission of amounts or disclosure in the financial statements to deceive financial statement users" (as defined by the Association of Certified Fraud Examiners, ACFE).

Of the three primary categories of fraud classified by the Association of Certified Fraud Examiners (financial statement fraud, corruption, and asset misappropriation), financial statement fraud caused the greatest losses for companies. [1, p. 10]. Its influence on a vast number of stakeholders is inevitable, since, in only one year, it caused a median loss of $954,000, took approximately 24 months to uncover and represented 10% of the fraud cases analysed [1]. In the long run financial statement fraud can also directly and significantly influence the economy.

In today's environment, in which great amounts of information and data are accessible, it has become possible to develop data mining methods. Data mining is the process of investigating and analysing large amounts of data with the aid of automatic or semi-automatic methods with the goal of uncovering meaningful patterns [2].

With the use of data mining methods, in addition to today's guaranteed transparency, and public availability of financial statements, the necessary preconditions for automatically detecting companies for which there are indications of financial statement fraud have been met. For this kind of detection it is essential to be familiar with the characteristics (models) of fraud and to define indicators which suggest that they exist. After defining indicators which suggest financial statement fraud exists it is possible to connect these with data mining methods that, in the end, result in the creation of a model that will detect potential financial statement fraud.

The process of data mining itself goes through separate phases, defined by the methodology, which, in the end, results in the detection of new information. The most common methodology is CRISP_DM (Cross-Industry Process for Data Mining). While carrying out data mining, according to previous research and an analysis of the CRISP-DM model, it is essential to do the following:

i) define good quality input that can detect fraud, **during the data preparation phase**. This implies defining **red flags** and their influence on the detection of fraudulent financial statements. Red flags, as a rule, pertain to the selection of key financial analysis indicators and individual positions on the balance sheet and profit and loss (income) statement that are more exposed to fraud. In a minority of the studies, red flags pertained to specific non-financial ratios; however, some research suggests that non-financial ratios are not significant in the formation of data mining models [3]. Of 112 financial and non-financial ratios analysed in fourteen (14) previous studies, there are ultimately only eight (8) financial ratios which point to financial statement fraud [3].

Table 1 – An outline of key financial ratios that can detect financial statement fraud based on previous research

Ratios:	Financial analysis indicators:
Activity ratios	Inventory to sales ratio
	Sales to total assets ratio
	Accounts receivables to sales ratio
Liquidity ratios	Working capital to total assets ratio
Solvency ratios	Total debt to total assets ratio
Profitability ratios	ROA (Profit after taxation/total assets)
	Net profit margin (Profit after taxation/sales)
Other ratios	Altman Z Score Model

ii) in the **modelling phase**, select a data mining method for the purpose of detecting financial statement fraud. As with financial ratio analyses, numerous studies apply various data mining methods without having a clear view of the most optimal method to detect financial statement fraud. An analysis of previous research carried out by Sharma and Panigrahi [4] shows that the most common data mining techniques are: logistic regression, neural networks, Naive Bayes, and classification trees.

In sum, an analysis of previously carried out research in this field suggests the following:

i) that in some studies certain financial analysis indicators are defined as significant in the financial statement fraud detection model, whereas these identical indicators in other studies are not considered significant in the model;

ii) that different data mining methods have different degrees of success in predicting financial statement fraud.

When a more detailed analysis of the previous research is carried out, it can be concluded that they encompass the following:

i) different time periods and

ii) different market environments,

which most likely leads to different results in the previously conducted studies [3, p. 85].

Consequently, it is to be expected that financial statement fraud models will change over time, primarily as a product of changing business environments, which can also be seen in the financial statements themselves.

Finally, research focussed on shorter time periods and on companies in identical markets can result in a much better model of fraud detection in financial statements. That is, it is necessary to continually monitor the characteristics of financial statement fraud and correct the existing fraud indicators and data mining methods in accordance with these changes.

2. RESEARCH

Since previous studies have defined various indicators (attributes) while defining models of fraud detection in financial statements, the hypotheses of this study are as follows:

H1: Since the market environment and perception of stakeholders has changed significantly since the beginning of the 2008 economic crisis, and taking into consideration the fact that various studies have placed importance on different indicators, the relevance and validity of existing models of fraud detection in the current business environment is brought into question.

Existing models of fraud detection need to be expanded to reflect the current business environment and to include new elements of financial statements in the data mining algorithms in order to improve the quality of the fraud detection/prediction model in financial statements.

Including elements of financial statements related to long-term intangible assets, financial assets (current and long-term), and prepaid expenses and accrued income as well as accrued expenses and deferred income in the data mining algorithm will lead to better quality detection/prediction of fraud in financial statements.

H2: Including cash flow indicators in the data mining algorithm will lead to better quality detection/prediction of fraud in financial statements.

The research itself was based on the following phases:

- **selection of companies** in which fraud was identified. Data was collected from published financial statements (for financial analysis indicators), but also from other sources (e.g. stock market analyses for information about share value, specialised journals, and daily newspapers etc.). Data on companies for whom financial statement fraud is identified in the USA is collected by analysing AAERs, that is, Accounting and Auditing Enforcement Releases, for companies listed on American stock exchanges. These releases are published by the SEC.

- **selection of attributes** which indicate financial statement fraud. This implies the selection of key indicators (primarily financial analysis indicators). By selecting key indicators, some characteristics of financial statement fraud can be identified.
- **application of data mining methods**. The algorithms have been applied to many existing programming tools; therefore, the focus has been shifted from the issue of programming the algorithm in the program code to the issue of selecting the best quality method and then mapping it to a specific problem. [5, p. 17].

2.1. SAMPLE SELECTION

This study was carried out on companies whose securities are listed on US stock markets. Data regarding fraud was obtained from the *Accounting and Auditing Enforcement Releases* (AAER) published by the *Securities and Exchange Commission* (SEC). The SEC in its AAERs identifies financial statement fraud, as well as other forms of fraud (e.g. corruption and insider trading). For that reason, even though there are a large number of AAERs, only a small number of these are suitable for the detection of financial statement fraud. In order to collect a sufficient sample, all AAERs published within a six-year period, that is, from 1 January 2011 to 31 December 2016, were the subject of this study. During this period the SEC published 619 AAERs.

Of the 619 AAERs analysed, the following companies were included in the sample:

i) those in which it was clear from the AAER that it was a case of financial statement fraud. Other forms of fraud (primarily corruption and insider trading) were not included in this study;

ii) those in which the fraud was related to a longer time period; that is, companies where fraud was detected on the basis of quarterly reports, which were corrected before preparing annual financial statements were not included;

iii) those <u>not</u> classified as falling under the financial sector, (i.e. their SIC code did not begin with the number 6);

iv) those for whom financial statements (for the current and previous year) were available;

v) those in which the fraud was <u>not</u> specific to one time period (e.g. the recognition of fictional income for the purpose of a one-off manipulation at the time of an initial public offering).

After this initial analysis, of the 619 companies analysed, 54 in which the SEC identified financial statement fraud within the given six-year period were detected for use in this study.

For the 54 companies concerned, a large number of attributes, as defined in the *Attribute selection* section, were collected from their published annual financial statements (so-called Form 10-K) for the first year in which financial statement fraud was identified.

2.2. SELECTION OF COMPARABLE COMPANIES

In order to carry out this research it was necessary to find companies in which fraud hadn't been identified but that were similar and comparable to those in which fraud had been identified. Comparable companies were considered to be the following:

i) **those that are in the same industry,** that is, companies with identical *Standard Industrial Classification Codes* (SIC codes); SIC codes are 4-digit codes assigned to companies that categorise companies according to industry/business activity. In the first step comparable companies were attempted to be found using 4 digits of the SIC code. If no comparable business entities were found at the level of 4 digits, then a comparable company at the level of three, or two digits, respectively, was attempted to be found.

ii) **those who operate in an identical time period**. In order to pair companies according to time period, comparable companies were chosen by selecting those entities whose financial period ended in the same or similar months as those in which financial statement fraud was identified. Comparable companies whose financial periods ended a maximum of nine months apart were paired together.

iii) **those with identical amounts of business activity**. While defining identical levels of business activity, selection criteria were carried out at two levels:

   i) sales were neither more nor less than 30% of the sales of the company for which financial statement fraud was identified;

   ii) total assets were neither more nor less than 30% of the total assets of the company for which financial statement fraud was identified.

The comparable company that was chosen was the one in which the absolute difference according to these two criteria was lower.

iv) there is proof that **financial statement fraud was not identified** in the selected comparable companies by testing the following:

i) the SEC AAER database from which the companies in which financial statement fraud was identified were also detected, and

ii) the Stanford Law Database on Shareholder Lawsuits.

If financial statement fraud was not identified in a comparable company, then it was included in this study as a comparable company.

In the end, the final sample included 46 pairs of companies (46 companies in which financial statement fraud was identified and 46 comparable companies).

Table 2 – Final list of companies in the sample

Company:	SIC Code (2 digits)	Year observed:	Fraud (Yes/No):
JDA SOFTWARE GROUP INC	73	2008	Yes
WIND RIVER SYSTEMS INC	73	2009	No
Saba Software	73	2008	Yes
Interactive Intelligence, Inc.	73	2007	No
MERGE HEALTHCARE INC	73	2002	Yes
Authentidate Holding Corp	73	2003	No
Imperial Petroleum	28	2011	Yes
Codexis Inc	28	2011	No
DIGI International INC	35	2005	Yes
NETOPIA INC	35	2005	No
Plastic2Oil Inc (ex. JBI, Inc)	50	2009	Yes
Rada Electronic Industries LTD	50	2009	No
China MediaExpress Holdings Inc.	73	2009	Yes
Constant Contact, Inc	73	2009	No
RINO International Corporation	35	2008	Yes
Meridian Bioscience Inc	28	2008	No
SUBAYE, Inc	73	2010	Yes
Support.com, Inc.	73	2010	No
Keyuan Petrochemicals	28	2010	Yes
Biofuel Energy Corp	28	2010	No
VOLT Information Sciences	73	2007	Yes
SPHERION CORP	73	2007	No
TheStreet, Inc	27	2008	Yes
Cavium, Inc.	36	2008	No
China North East Petroleum Holdings, Ltd	13	2009	Yes
Endeavour International Corporation	13	2009	No
CARTERS, Inc.	23	2004	Yes
HARTMARX CORP/DE	23	2004	No
Soyo Group Inc.	50	2007	Yes

Company:	SIC Code (2 digits)	Year observed:	Fraud (Yes/No):
CRAFTMADE INTERNATIONAL INC	50	2006	No
Huron Consulting Group Inc.	87	2006	Yes
ICF International, Inc.	87	2006	No
Isilon Systems, Inc.	35	2006	Yes
OVERLAND STORAGE, INC.	35	2006	No
ISLAND PACIFIC INC	73	2001	Yes
CATAPULT COMMUNICATIONS CORP	73	2001	No
KOSS Corporation	36	2005	Yes
Touchtunes Music Corporation	36	2004	No
Hansen Medical Inc	38	2008	Yes
Alpha Pro Tech, LTD.	38	2008	No
Syntax-Brillian Corporation	36	2006	Yes
LOUD TECHNOLOGIES INC.	36	2006	No
Arthrocare Corp	38	2006	Yes
ZOLL MEDICAL CORP	38	2006	No
Symmetry Medical Inc.	38	2004	Yes
ANIMAS CORP	38	2004	No
INPHONIC INC	48	2005	Yes
LODGENET INTERACTIVE CORP	48	2005	No
International Commercial Television Inc.	59	2007	Yes
ASIA GLOBAL HOLDINGS CORP.	59	2007	No
Basin Water, Inc.	49	2006	Yes
U.S. ENERGY SYSTEMS	49	2006	No
STARMEDIA NETWORK INC	73	2000	Yes
ONESOURCE INFORMATION SERVICES INC	73	2000	No
THOR INDUSTRIES INC	37	2003	Yes
OSHKOSH TRUCK CORPORATION	37	2003	No
LocatePlus Holdings Corporation	73	2005	Yes
INTERNET AMERICA, INC	73	2006	No
MICHAEL BAKER CORP	87	2006	Yes
GEVITY HR, INC.	87	2006	No
GSI GROUP INC	36	2004	Yes
ROFIN SINAR TECHNOLOGIES INC	36	2004	No
ZALE CORP	59	2004	Yes
DICK’S SPORTING GOODS, INC.	59	2005	No
Satyam Computer Services Limited	73	2005	Yes
CIBER INC	73	2004	No
DHB INDUSTRIES, INC.	38	2003	Yes

Company:	SIC Code (2 digits)	Year observed:	Fraud (Yes/No):
Wright Medical Group	38	2003	No
NUTRACEA	20	2007	Yes
TOFUTTI BRANDS INC	20	2007	No
Fischer Imaging Corp.	38	2000	Yes
American Science and Engineering, Inc.	38	2001	No
Computer Sciences Corporation (CSC)	73	2010	Yes
YAHOO! Inc.	73	2010	No
Stein Mart, Inc.	56	2011	Yes
Buckle, Inc.	56	2011	No
OCZ Technology Group, Inc (now: ZCO Liquiidating Corp)	35	2011	Yes
Fusion-IO, Inc.	35	2011	No
Diebold Nixdorf, Inc. (previously Diebold)	35	2003	Yes
NetApp, Inc. (full name: Network Appliance, Inc.)	35	2004	No
MARRONE BIO INNOVATIONS, INC.	28	2013	Yes
Flexible Solutions International, Inc.	28	2013	No
ModusLink Global Solutions, Inc.,	73	2007	Yes
Global Payments Inc.	73	2007	No
ENER1, Inc.	36	2010	Yes
AstroNova, Inc. / previously Astro-Med, Inc.	36	2011	No
IEC Electronics Corp	36	2012	Yes
PARK ELECTROCHEMICAL CORP.	36	2012	No
WEATHERFORD INTERNATIONAL LTD	13	2007	Yes
SMITH INTERNATIONAL INC	35	2007	No
HENRY JACK & ASSOCIATES INC	73	2012	Yes
MENTOR GRAPHICS CORPORATION	73	2012	No

Analysing the companies by industry, it can be seen that the majority of companies are either in manufacturing (36) or in services (30).

2.3. ATTRIBUTE SELECTION

Based on previous research a total of 122 attributes were chosen (in the form of single data points or derived indicators), while collecting data on companies in which financial statement fraud was identified. Since there were a large number of attributes involved, and in order to speed up individual algorithms and make them more efficient, it was necessary to select key attributes and to eliminate redundant and unimportant attributes from the data set. With the help of the ReliefF algorithm[1] all unimportant attributes were eliminated. The ReliefF algorithm is an attribute selection algorithm, which is used in binary classifications with an emphasis on polynomial classification. This algorithm uses random sampling of instances and locates the nearest neighbour from the same or opposite class. The ReliefF algorithm and its variations (RelieveD, A-Relief) are considered successful due to their simplicity and efficiency [6, p. 219]. Attributes whose values were greater than 0.0025 were chosen; that is, the following 18 attributes were chosen:

Table 3 – A summary of selected features (attributes) after applying the ReliefF algorithm (in order of importance of individual attributes)

Attribute name:	Attribute description:	Attribute significance:
Intangible assets (t) / Total Assets (t)	Share of long-term intangible assets in the total assets in the observed year (t)	0.01501105619795581
Long Term Financial Assets (t) + Intangible Assets (t) + Other Long Term Assets (t)/Total Assets (t)	Share of long-term intangible assets, financial assets and other long-term assets in the total assets in the observed year (t)	0.01236795117087239
Current Assets (t) / Total Assets (t)	Share of current assets in the total assets in the observed year (t)	0.008730616329139618
Intangible assets (t-1) / Total Assets (t-1)	Share of long-term intangible assets in the total assets in the year prior to the observed year (t-1)	0.0066119239661836535
Logarithm of Total Assets (t)/ Logarithm of Total Assets (t-1)	Logarithm of total assists index	0.004967516809373735
Operating lease obligation (t) / Long term debt (t)	Operating lease obligation to long term debt ratio in the observed year (t)	0.0045648386718648995
Accounts Receivables (t-1) / Sales (t-1)	Accounts receivables to sales ratio in the year prior to the observed year (t-1)	0.004349345647190639
Cash flow from financial activities (t) / Cash flow from financial activities (t-1)	Cash flow from financial activities index	0.0034483082335137775
Long-Term Liabilities (t)	Long-term liabilities in the observed year (t)	0.003061224489795918
Long-Term Liabilities (t-1)	Long-term liabilities in the year prior to the observed year (t-1)	0.003061224489795918
Financial assets (t)	Current financial assets in the observed year (t)	0.003061224489795918

[1] A large number of studies have shown that the success of the application of individual algorithms is increased after carrying out attribute selection. (Rosario & Thangadurai, 2015), (Kononenko, Robnik-Šikonja, & Pompe, 1996)

Attribute name:	Attribute description:	Attribute significance:
Profit / Loss after taxation (t-1)	Profit/loss after taxation in the year prior to the observed year (t-1)	0.003061224489795918
Cash flow from investing activities (t-1)	Cash flow from investing activities in the year prior to the observed year (t-1)	0.003061224489795918
Working Capital (t) / Total Assets (t)	Net working capital to total assets ratio in the observed year (t)	0.0030171516614922555
(Current Assets (t) / Total Assets (t))/(Current Assets (t-1) / Total Assets (t-1))	Share of current assets to total assets index	0.0028653939284571496
(Sales (t)/No of employees (t))/ Sales (t-1) / No of employees (t-1))	Sales to number of employees ratio index	0.0027046374460642886
Operating Cash flow (t-1) / Sales (t-1)	Operating cash flow to sales ratio in the year prior to the observed year (t-1)	0.002676100539391554
Net Profit (t-1) / Sales (t-1)	Net profit margin the year prior to the observed year (t-1)	0.0025641030463030255

A survey of these key indicators proves the first hypothesis of this research, that is, that there are some other indicators that can detect financial statement fraud in addition to the indicators defined in previous research, *Table 1 – An outline of key financial ratios that can detect financial statement fraud based on previous research*.

2.4. DATA MINING METHOD

Data mining of the prepared data was carried out with the help of the WEKA program package, from the University of Waikato, New Zealand (*http://www.cs.waikato.ac.nz/~ml/index.html*). Weka is set of tools for machine learning written in the Java programming language. Before testing individual data mining algorithms, data editing was based on two prior steps, as follows:

Step 1: Discretization of numerical data from financial statements

Discretization is the process of transforming numerical data into nominal ones, in such a way that the numerical values are moved to appropriate groups, which have a specific number. In this particular sample, discretization of numerical data was carried out on specific elements of the financial statements, that is on the attributes listed in *Table 3 - A summary of selected features (attributes) after applying the ReliefF algorithm (in order of importance of individual attributes).* While carrying out the discretization, the basic, pre-installed assumptions in WEKA were used, that is, the unsupervised equal width binning approach was used.

Step 2: Testing by means of algorithm

After carrying out the discretization of data, specific algorithms were tested. Since previous research does not show that there is a unique algorithm that can be used in the detection of financial statement fraud, eight (8) key data mining algorithms from the literature were selected. All the algorithms were used with their basic functions (pre-installed in the WEKA program package). The following algorithms were used:

ZeroR (Weka name: *ZeroR*) is the simplest classification algorithm, which is based on a model with only one rule. Regardless of the number of attributes, ZeroR always predicts on the basis of the most common attribute. It is precisely because of its simplicity that it is used exclusively to compare the successfulness of other algorithms with respect to this simple algorithm.

- **Naive Bayes** (Weka name: *NaiveBayes*) is a probabilistic classifier, which is based on the application of Bayes theorem with strong assumptions of independence between features. The foundation of this algorithm is a probability calculation of each individual category for the sample, after which an output variable is defined for only the category with the greatest probability. The basic settings of the algorithm assume the use of normal distribution.
- The **SVM** (Weka name: *SMO*) model is based on a graphic representation of each data item as a point in space. The data is mapped into two groups (in this case into financial statement fraud and financial statements for which financial statement fraud has not been identified), while new data depending on its position is attached to one of the two existing groups. The basic settings of the algorithm assume the use of kernel functions, that is, a function that corresponds to a scalar product in a higher dimension.
- **K nearest neighbour** (Weka name: *Lazy.IBk*) is used for classification or regression, and it is based on a calculation of Euclidean distance, that is, on finding the shortest distance between two points in space. The basic settings of the algorithm assume using Euclidean distance with one defined pre-installed number of nearest neighbours.
- **AdaBoostM1** (Weka name: *AdaBoostM1*) is a meta-algorithm, which is designed for classification. Its basic purpose is to speed up some of the existing algorithms. In the basic WEKA settings the Decision-Stump algorithm is used.
- **Bagging** (Weka name: *Bagging*), as with AdaBoostM1, is a meta algorithm which is used for classification and regression. In essence it speeds up existing algorithms by decreasing variance. In classification

tasks (as is the case in this study), the algorithm makes predictions by calculating the weighted value of probability. In its basic settings WEKA uses the REPTree algorithm.

- **C 4.5** (Weka name: *J48*) is an algorithm which generates a decision tree, and thus it is generally used for classification. In practice, it is one of the most commonly used algorithms.
- **Random forest** (Weka name: *RandomForest*) is used both for classification and regression, and this algorithm creates a large number of decision trees.

On the basis of the above, the results of testing individual algorithms were as follows:

Table 4 – Results of the successfulness of testing algorithms (shown in order of successfulness)

No.	Algorithm name:	WEKA name for the algorithm:	Test results:
1.	SVM	SMO	62.19%
2.	AdaBoost M1	AdaBoostM1	59.19%
3.	Random Forest	RandomForest	57.17%
4.	Naive Bayes	NaiveBayes	54.76%
5.	k nearest neighbour	Lazy.IBk	50.33%
6.	C 4.5	J48	50.09%
7.	Bootstrap / Bagging	Bagging	48.56%
8.	ZeroR	ZeroR	45.56%

The ZeroR algorithm, as the simplest algorithm, is least able to detect fraud, which brings us to the conclusion that an increase in the complexity of the algorithm leads to the improvement of its successfulness. The most successful algorithm is SVM, which in 62.19% of the cases correctly detected financial statement fraud.

In order to prove hypotheses 1 and 2, that is, to prove that including other financial analysis indicators, especially those related to long-term intangible assets, financial assets (current assets) and cash flow indicators, can result in detecting financial statement fraud more successfully, a comparable test was also carried out; however, only indicators which were identified as key indicators in previous research *Table 1 – An outline of key financial ratios that can detect financial statement fraud based on previous research* were used. The Altman Z score model was not included, while the other seven indicators were included in the respective tests. Using these financial analysis indicators the following results of the successfulness of individual algorithms were obtained:

Table 5 – Results of the successfulness of comparable testing of algorithms (in order of successfulness)

No.	Algorithm name:	WEKA name for the algorithm:	Test result:
1.	Naive Bayes	NaiveBayes	53.23%
2.	Bootstrap / Bagging	Bagging	48.56%
3.	C 4.5	J48	47.02%
4.	Random Forest	RandomForest	45.72%
5.	ZeroR	ZeroR	45.56%
6.	AdaBoost M1	AdaBoostM1	45.21%
7.	k nearest neighbour	Lazy.IBk	41.07%
8.	SVM	SMO	40.71%

The results of comparative testing show that using only the frequently used indicators from previous research led to a lower success rate of the algorithms in fraud detection than when including indicators related to intangible assets, financial assets and cash flow.

The hypotheses of this study have been proven, that is:

H1: Since the market environment and perception of stakeholders has changed significantly since the beginning of the global economic crisis in 2008, and taking into consideration the fact that various studies have placed importance on different indicators, the relevance and validity of existing models of fraud detection in the current business environment is brought into question.

Existing fraud detection models require certain additions to take into account changes to the current business environment, and including new elements of financial statements in the data mining algorithms improves the quality of fraud detection/prediction models in financial statements.

Including elements of the financial statements related to long-term intangible assets, financial assets (current and long-term) and prepaid expenses and accrued income as well as accrued expenses and deferred income leads to better quality detection/prediction of fraud in financial statements.

H2: Including cash flow indicators in the data mining algorithm leads to better detection/prediction of fraud in financial statements.

A summary of the indicators (attributes), which were used in individual algorithms in order to detect financial statement fraud, and which prove Hypotheses 1 and 2 are as follows:

Table 6 – Summary of the indicators (attributes) used to prove Hypotheses 1 and 2 of the study

Attribute name:	Attribute description:	Segment:
Intangible assets (t) / Total Assets (t)	Share of long-term intangible assets to total assets in the observed year (t)	Proves hypothesis 1
Long Term Financial Assets (t) + Intangible Assets (t) + Other Long Term Assets (t)/Total Assets (t)	Share of long-term intangible assets, financial assets and other long-term assets to total assets in the observed year (t)	Proves hypothesis 1
Intangible assets (t-1) / Total Assets (t-1)	Share of long-term intangible assets to total assets in the year prior to the observed year (t-1)	Proves hypothesis 1
Financial assets (t)	Current financial assets in the observed year (t)	Proves hypothesis 1
Operating lease obligation (t) / Long term debt (t)	Operating lease obligation to long term debt ratio in the observed year (t)	Proves hypothesis 1
Cash flow from financial activities (t) / Cash flow from financial activities (t-1)	Cash flow from financial activities index	Proves hypothesis 2
Cash flow from investing activities (t-1)	Cash flow from investing activities in the year prior to the observed year (t-1)	Proves hypothesis 2
Operating Cash flow (t-1) / Sales (t-1)	Operating cash flow to sales ratio for the year prior to the observed year (t-1)	Proves hypothesis 2

3. CONCLUSION

Silversone et al. [7, pp. 25-26]. question whether the size and complexity of fraud such as was the case with Enron, Lehman Brothers, AIG, and Bernie Madoff, mean that we are living in a new period – a period of fraud, or are these frauds a product of the new media age – an age in which the fast reactions of the media to fraud detections prohibit experts from analysing in detail and explaining the reasons which led to fraud.

Regardless of the reasons and explanations for fraud, nowadays, the technical conditions to detect financial statement fraud in real time are available. Being able to receive documentation electronically, and to receive and process documents in XBRL, are the basic preconditions that are necessary to quickly and promptly identify potential financial statement fraud.

In the past few years a certain number of studies have been published in the field of fraud detection in financial statements using data mining. However, each study placed importance on certain attributes (typically financial analysis indicators), and used various algorithms for the detection of financial statement fraud. From these studies, however, it wasn't possible to define a unique model for the early detection of financial statement fraud.

The vast number of studies selected basic financial analysis indicators without taking into account changes in the economic environment, which are reflected in the financial statements of particular companies. For example, many mergers and acquisitions leave their mark on financial assets (in individual financial statements), and on intangible assets – goodwill (in consolidated financial reporting). Increasingly complex demands for recognising financial assets, combined with the global economic crisis, require users of financial statements to monitor financial assets better. On the other hand, the emphasis on technology and the fourth industrial revolution (i.e. digital revolution) has led to an increasing number of companies who develop intangible products, for which part of the expenses related to development are recorded in intangible assets.

Research that has been carried out up to now hasn't taken into account these two, increasingly more frequent positions in the financial statements of a great number of companies – that is, intangible assets and financial assets.

This study has shown that including these indicators in the data mining algorithms leads to more successful detection of financial statement fraud than not including them (62.19% compared to 53.23%).

The second hypothesis of the study is related to including cash flow indicators in the algorithms used to detect financial statement fraud. The main reason for including these indicators is related to the fact that in general it isn't possible to separate the success of the company from its cash flow. Due to changes in the financial reporting standards, which to a great extent today are based on defining "fair values" on each day of financial reporting, it is possible for there to be discrepancies between cash flow and success indicators. In part, these changes to the financial reporting standards are connected to the ever more common abandonment of accounting conservatism under the pretext that in such a way "good" investments are undervalued in the sense that conservative ways of reporting classify them as "bad" investments. It is often forgotten, however, that, even though it isn't explicitly mentioned in the accounting standards, better quality financial statements should be conservative in order to protect the company's shareholders [8].

The model in this study also includes three cash flow indicators, which proves the second hypothesis.

The model that was set out in this study has its limitations. Its basic limitation is the fact that financial statement fraud models change from period to period, as does the structure of financial reporting in relation to market conditions. Visible increases in financial assets and intangible assets in the balance sheets of companies, in addition to the fact that these positions in the balance sheet are not yet fully understood by users of financial statements, have left open the possibility of fraud precisely in these positions.

As a re1, and to protect the financial system, it is our duty to implement this kind of system as soon as possible.

LITERATURE

1. **Association of Certified Fraud Examiners.** *Report to the Nation - 2020 Global Study on occupational Fraud and Abuse.* Austin, USA : Association of Certified Fraud Examiners, 2020.
2. **Pejić Bach, Mirjana.** Rudarenje podataka u bankarstvu. *Zbornik Ekonomskog fakulteta u Zagrebu.* 3, 2005, Vol. 1, pp. 181-193.
3. **Kopun, Dubravka.** A Review of the Research on Data Mining Techniques in the Detection of Fraud in Financial Statements. *Journal of Accounting and Management.* VIII, 2018, Vol. I, pp. 1-16.
4. **Sharma, A. & Panigrahi, K.P.** A Review of Financial Accounting Fraud Detection based on Data Mining Techniques. *International Journal of Computer Application.* 39, 2012, Vol. 1, pp. 37-47.
5. **Šušnjar, Marko.** Razvoj prediktivne metode određivanja pogonske čvrstoće materijala (doktorska disertacija). Split : Sveučilište u Splitu - Fakultet elektrotehnike, strojarstva i brodogradnje, 2012. doktorska disertacija.
6. **Rosario, Francisca S. & Thangadurai, K.** RELIEF: Feature Selection Approach. *International Journal of Science and Research.* 10 2015, Vol. 4, 11, pp. 218-224.
7. **Silverstone, Howard, & dr.** *Forensic Accounting and Fraud Investigation for Non-experts (third edition).* New Jersey : John Wiley & Sons, Inc., 2012.
8. **Solomons, David.** *Making Accounting Policy: The Quest for Credibility in Financial Reporting.* s.l. : Oxford University Press, 1986.

PRIMJENA TEHNIKA RUDARENJA PODATAKA U OTKRIVANJU FINANCIJSKIH PRIJEVARA

SAŽETAK RADA

Financijska prijevara jedna je od tri vrste prijevara (financijska prijevara, korupcija i pronevjera imovine) koja kompanijama uzrokuje najveće gubitke. Povećanje broja financijskih prijevara u proteklih 20 godina (Enron, WorldCom, Parlamat, Tesco, Wirecard…) zajedno s razvojem tehnika rudarenja podataka koje su preduvjet za rano ortkrivanje prijevare dovele su do velikog broja radova u ovom području.

Do danas, provedena istraživanja nisu dovela do jednoznačnog modela; različite studije ukazuju na korištenje različitih financijskih indikatora i korštenje različitih tehnika rudarenja podataka što je dovelo do različitih razina uspjeha u otkrivanju financijskih prijevara. Razlog je taj što studije nisu u analizu uključile i promjenu u poslovnom okruženju koje imaju izravan utjecaj na strukturu financijskih izvještaja te posredno i na indikatore koji se koriste pri financijskoj analizi. Ovaj rad pokazuje da fokusiranje na kraće vremensko razdoblje i na kompanije koje djeluju na istim tržištima može dovesti do kvalitetnijeg modela za otkrivanje financijskih prijevara.

Ključne riječi: *financijska izvješća, prijevare; otkrivanje prijevara, rudarenje podataka, indikatori financijske analize*