Technische Universität Ilmenau
Fakultät für Wirtschaftswissenschaften und Medien
Institut für Betriebswirtschaftslehre
Fachgebiet Rechnungswesen und Controlling

THE DETECTION OF FRAUDULENT FINANCIAL STATEMENTS USING TEXTUAL AND FINANCIAL DATA

Dissertation zur Erlangung des akademischen Grades eines

Doktors der Wirtschafts- und Sozialwissenschaften (Dr. rer. pol.)

vorgelegt von:
Tobias Christian Gleichmann

Erstgutachter: Prof. Dr. Michael Grüning
Zweitgutachter: Prof. Dr. Jörg. R. Werner

Tag der wissenschaftlichen Aussprache: 01.09.2020

urn:nbn:de:gbv:ilm1-2020000457

Table of Content

List of Tables

List of Figures

List of Abbreviations

ABC	Apple-Bushel-Crop (Theory)
abs.	Absolute
AAER	Accounting and Auditing Enforcement Releases
AM	Asset misappropriation
ACFE	Association of Certified Fraud Examiners
AGID	Automatically Generated Inflection Database
AICPA	American Institute of Certified Public Accountants
ANN	Artificial neural network
AS	Auditing standard
ASAF	Accounting Standards Advisory Forum
ASB	Auditing Standards Board
ASC	Accounting Standards Codification
ASU	Accounting Standards Update
AUC	Area under the curve
CAPM	Capital asset pricing model
CEO	Chief Executive Officer
CFE	Certified Forensic Examiner
CIK	Central index key
COSO	Committee of Sponsoring Organisations of the Treadway Commission
CPA	Certified Public Accountant
CSR	Corporate social responsibility
DM	Deutsche Mark (German Currency)
DSMR	Design science research methodology
ECMH	Efficient capital market hypothesis
EDGAR	Electronic Data Gathering, Analysis and Retrieval System
FA	Forensic accountant

FASB	Financial Accounting Standards Board
FERF	Financial Executives Research Foundation
FN	False negative
FOTM	Fraud on the market
FP	False positive
FPR	False-positive rate
FSF	Financial statement fraud
GAAP	Generally Accepted Accounting Principles
GAAS	Generally Accepted Auditing Standards
GAO	(United States) Government Accountability Office
IAF	Internal Audit Function
IAS	International Accounting Standards
IASB	International Accounting Standards Board
IASC	International Accounting Standards Committee
IDW	Institut der Wirtschaftsprüfer (Institute of Public Auditors in Germany)
IFRS	International Financial Reporting Standards
IG	Information gain
IGR	Information gain ratio
IIA	Institute of Internal Auditors
IT	Information technology
IRH	Incomplete revelation hypothesis
KNN	k-nearest neighbour
LDA	Latent Dirichlet allocation
MD&A	Management Discussion & Analysis of Financial Condition and Results of Operations
MICE	Money-Ideology-Coercion-Entitlement (Model)
MoU	Memorandum of Understanding
NB	Naïve Bayes

NYSE	New York Stock Exchange
PCAOB	Public Company Accountant Oversight Board
perc.	Percentage
qual.	Qualitative (features)
quant.	Quantitative (features)
ROC	Receiver operating characteristic
SAP	Statement on Auditing Procedure
SAS	Statement on Auditing Standards
SEC	(United States) Securities and Exchange Commission
SIAS	Statement on Internal Auditing Standards
SIC	Standard industry classification
SOX	Sarbanes-Oxley Act
stdev.	Standard deviation
SVM	Support vector machine
TN	True negative
TP	True positive
TPR	True-positive rate
VarCon	Variant Conversion Info
XBRL	eXtensible Business Reporting Language

List of Symbols

α	Hyperplane location parameter
β	Hyperplane location parameter
γ	Complexity parameter (radial basis function)
λ	Fraction of the error
μ	Mean
ρ	Pearson correlation
σ	Variance
τ	Single feature
cov	Covariance
err	Classification error
a	Input vector
c	Class label
C	Penalty value
d	Distance
E	Entropy
e	Euler’s number (constant)
F	Number of features/number of input neurons
I	Information value
$IntI$	Intrinsic information value
K	Number of classes
k	Number of neighbours
L	Length of computation (time)
n	Number of observations
P	Point in an n-dimensional space
p	Portion of observations belonging to a certain class
Q	Point in an n-dimensional space

S	Data set
T	Threshold
w	Weighting term
x^*	New observation
x_i	Input from i-th neuron
X	Set of features
Y	Classification outcome/activation
Y^*	True outcome/label

1 Introduction

Financial statements represent a major source of information for various market participants and other stakeholders, hence their validity and reliability are of utmost importance. In its Report to the Nations of 2018, the Association of Certified Fraud Examiners (ACFE) estimated that the losses of occupational fraud exceeded $ 7 billion for the respective period.[1] Although occupational fraud covers corruption and asset misappropriation in addition to financial statement fraud, the latter overwhelmingly causes the most damage per case. Accounting fraud and financial statements that are not fully in line with generally accepted accounting principles (GAAP) severely compromise market efficiency. Accordingly, regulators, policy-makers and auditors all strive to deter, prevent, and detect financial statement fraud to preserve market participants' trust in corporate disclosures, especially in audited statements like annual reports. Although various recent measures aimed at fraud (e.g. the Public Company Accounting Reform and Investor Protection Act of 2002,also known as the Sarbanes-Oxley Act – SOX) have probably reduced the occurrence of fraudulent activities, risks remain. In its summary of performance and financial information for the fiscal year 2016, the United States Securities and Exchange Commission (SEC) stated that "rooting out financial and disclosure fraud must be a priority for Enforcement".[2] In addition to the in-depth analysis of individual transactions, a range of predictors from publicly available data indicating an increased likelihood of accounting errors for a particular firm may improve the efficiency and effectiveness of accounting fraud identification.

The existing literature documents the evolution of financial statement fraud detection. Initially, financial information was mostly gathered from financial statements and in conjunction with capital market data was used to predict fraudulent cases. It is an obvious consequence of manipulation that financial statements do not correctly reflect the economic situation of a firm. Managers must make considerable effort to conceal their fraudulent actions in order to hide manipulations from auditors and the recipients of the financial statements. Although accounting fraud originates in an initial manipulation, further fraudulent alterations in subsequent periods are often required to obscure and hide the initial fraud scheme. For example, a firm may illegitimately shift earnings from future to current periods to hide financial distress or boost performance. However, in subsequent periods, the

[1] ACFE (2018), Report to the Nation. Retrieved from: http://www.acfe.com/report-to-the-nations/2018/.

[2] SEC (2016), Summary of Performance and Financial Information. Retrieved from: https://www.sec.gov/files/2017-03/sec-summary-of-performance-and-financial-info-fy2016.pdf.

firm may feel compelled to mask prior fraud in order to continuously meet market expectations. Accordingly, manipulations may last over several periods and it might prove particularly challenging to distinguish between fraudulent and truthful reports.[3] Previous literature suggests that approaches that are limited to baseline financial ratios that rely on annual data have low predictive power, whereas a time series analysis of financial statements or the provision of additional context to the raw company financials might improve fraud detection considerably (e.g. Abbasi, Albrecht, Vance, & Hansen, 2012).

Recent research has broadened the focus from raw financial information to the additional consideration of the textual analysis of corporate narratives, which might reveal further information with predictive power (e.g. Cecchini, Aytug, Koehler, & Pathak, 2010a; Purda & Skillicorn, 2015). Textual analysis in accounting and finance is applied to examine how management creates narratives and facilitates examination of the interaction of attributes of corporate disclosure and the underlying management and firm characteristics. To predict fraud, this perspective on the management's decision-making process is of utmost importance. It is plausible that it is managers who commit or direct fraudulent accounting activities. Therefore, the narratives created by these managers are supposed to reveal potential clues of manipulated financial reports.

Extensive research has been conducted on the topic and has linked various characteristics of corporate narratives to firm performance and management characteristics. From a wider perspective, the results suggest that the textual components of corporate disclosure are able to reveal subtle details of a firm and its management. Moreover, managers can and probably do influence recipients' absorption of the given information by manipulating the textual components. For financial statement fraud detection, where subtle details beyond the raw financials and from the inside of the firm and the management may be important, utilizing textual analysis offers a great opportunity to identify additional clues. Therefore, in this study, the quantitative and qualitative (textual) data of 10-K filings are utilized to identify fraudulent accounting practices. The quantitative data comprise company financials and other corporate characteristics, while the qualitative data include narratives from textual sections.

A combination of both types of predictors, tested in an environment that reflects real-world applicability, while ensuring the reliability and robustness of the results, has yet to be

[3] In this study, the terms truthful and non-fraudulent will be used interchangeably, for further substantiations see section 4.1.1.

conducted. Therefore, this study's research goal is the development of a detection model for future financial statement fraud, relying on qualitative and quantitative information from annual reports.

1.1 Accounting Research and Nature of the Study

This study in financial accounting focuses on the use of publicly available data from annual reports to detect fraudulently altered financial statements. Financial accounting is essentially considered as the measurement, summarization and communication of the economic activities of an organization to outside recipients (Sutton, 2006, p. 2). Although the type of organization is not restricted to corporate activities, financial accounting has gained particular practical importance from the need to hold businesses accountable to their creditors, owners and other stakeholders. Accounting practices have been developed by accountants in their everyday business as well as through the involvement of governments, which have increasingly recognized the need to engage in rule setting to protect shareholders' and stakeholder' interests (Schroeder, Clark, & Cathey, 2019, pp. 27–29). Initially, academic accounting research was scarcely deemed relevant or present and did not play a decisive role in the development of the profession (Ryan, Scapens, & Theobald, 2002, pp. 94–113). However, over time, accounting scholars have had a greater impact on the development of financial accounting, especially with the formation of numerous government agencies and professional bodies in which academic exchange with the research community is possible. This development was driven by a multitude of factors, especially of an environmental nature, such as economic growth and the social and political situation and thus varying considerably across countries.

Of particular interest in this regard was the shift towards positive accounting research in the United States of America (USA) in the 1980s (Ryan et al., 2002, pp. 106–109).[4] Positivism is a derivative of empiricism and states that knowledge regarding a subject is derived from the observation of natural phenomena, their appearances and properties as well as their relations (Ryan et al., 2002, p. 17). Consequently, positivists and empiricists argue that true knowledge can only be derived from perceptions within a reality that is value-free and independent, with meaningful statements verified by observation.

[4] Nevertheless, positive accounting research began to flourish in the 1960s. See the review of Watts and Zimmerman (1990) for a detailed discussion.

Within empirical research in accounting, two major streams can be identified: behavioural accounting research and market-based accounting research (Ryan et al., 2002, p. 103). The former deals with the generation and absorption of financial information (Hofstedt, 1976, pp. 44–45), whereas the latter draws upon the impact of financial accounting information on capital markets (Lev & Ohlson, 1982, pp. 251–252). In the context of empirical accounting research, the research question is typically answered by a set of statistical tests. In this way, empirical research relies on variables that depict the properties of an event or the phenomena of an observed object (Ryan et al., 2002, p. 118). The measurement of the variables can be used to distinguish the research design in the social sciences between qualitative and quantitative research (Bazeley, 2004, p. 142). For qualitative research, the variables are constructed from qualitative data like textual data, whereas quantitative research focuses on data in numerical form. However, there also exist mixed-model or mixed-method approaches, whereby qualitative and quantitative data are used in conjunction.

In a perfect world, the data used to test a hypothesis would be derived from controlled experiments (Ryan et al., 2002, pp. 122–131). When conducting empirical research, scholars in finance and accounting mostly rely on quasi-experimental research designs, rather than actual experiments, as opposed to the common design in the natural sciences. This is due to the fact that the researcher is usually unable to directly manipulate the variables under study. In empirical accounting research, an *ex post facto* design is often the only possible solution, as the event of interest has already occurred and the variables must be chosen afterwards, without the direct control of the researcher during the occurrence of the event of interest.

When studying financial statement fraud, it is essential to identify the drivers behind fraudulent actions (what brings a person to fraudulently alter financial statements?) and the clues that can help to detect fraudulent behaviour (how can a person/company that has altered the statements be identified?) In addition to the general interest in identifying and statistically testing the potential determinants of fraud, utilizing these findings to build a sound detection model represents a second step. Although both steps build upon each other, the research design and therefore the evaluation of the outcome can substantially differ. Indeed, whereas the former primarily focuses on explanatory statistical techniques to test a hypothesis and to report (for example) on the significance of specific determinants, the latter is interested in performance evaluations like the fraction of detected cases. Creating a detection model is neither a novel nor fundamentally innovative idea. However, the concept behind the development of the model and the potential to assess the performance of different models in conjunction may generate a decisive contribution.

In information science, the so-called design-science paradigm has been developed to support the creation and innovation of new technology or artefacts like models, methods, constructs, and instantiations (Hevner, March, Park, & Ram, 2004). Design-science originated in the engineering discipline and at its core, constitutes a problem-solving paradigm. An artefact, for example in the form of a detection model, is developed and applied to an environmental need, relying on a rigorous knowledge base that is constituted of a theoretical and methodological foundation.[5] The theoretical foundation rests in the domain knowledge about a particular subject of interest, whereas methodological foundations are often associated with the skills required for empirical work, like statistics, programming languages or machine learning fundamentals. Design-science has developed its own research methodology guidelines and frameworks to support researchers in conducting high-quality research (e.g. Hevner et al., 2004; Peffers, Tuunanen, Rothenberger, & Chatterjee, 2014). These generally outline the importance of rigorous methods for the construction and evaluation of the design artefact and the need to ensure a precise and verifiable contribution.[6] Peffers et al. (2014) state that the frameworks and guidelines should not be followed blindly but rather emphasize good practices to help researchers in this often interdisciplinary and unstructured field to conduct good research.

Within the business sphere, design-science seeks to offer the desired solutions for readily defined problems using domain and design-science knowledge (van Aken, 2005, pp. 20–22). In accounting research, design-science approaches have often been established in areas where the research object interferes with information technology (IT) systems, but they remain rather scarce.[7] In the fraud detection context, for example, Abbasi et al. (2012) used a design-science approach to develop a MetaFraud model that combines a detection process-model for quantitative variables. In this way, it can answer the hypothesis about the usefulness of data from quarterly reports for detection purposes by evaluating the detection performance of models with different data sources. Especially in the fraud or bankruptcy detection (prediction) literature, a large number of studies rely on a research approach that has been influenced by design-science and that often depends on machine learning approaches. This study builds upon the guidelines for design science to ensure the quality of

[5] See the information systems research framework in Hevner et al. (2004, p. 80).

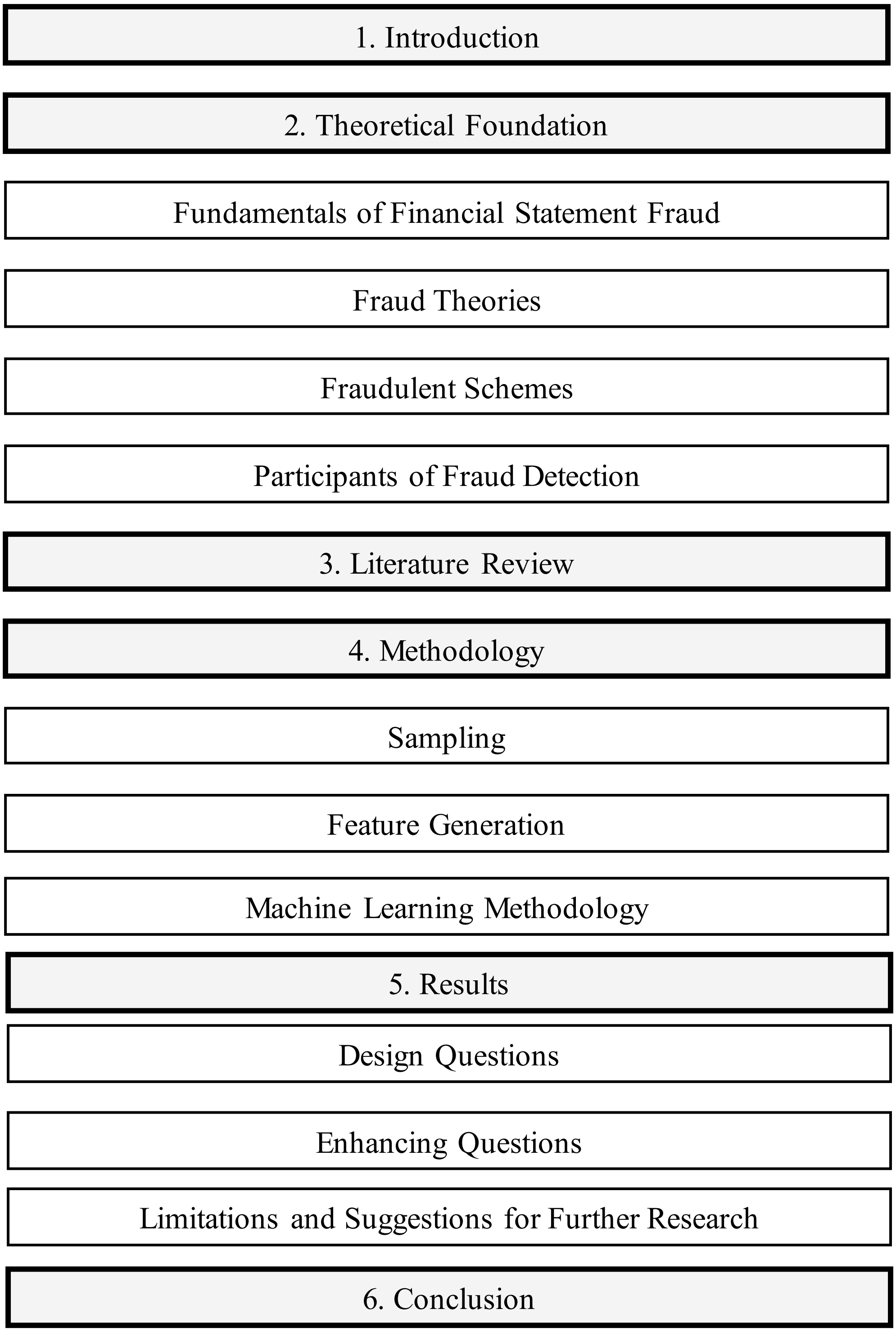
[6] See Hevner et al. (2004, p. 83) for the seven guidelines of design science or Peffers et al. (2014, p. 54) for the DSRM framework.

[7] For further details see Geerts (2011) on accounting information systems.

the results and to enable comparison with the outcomes of similar studies, an essential aspect of contributing to the fraud detection literature.

This study mainly falls in the behavioural accounting literature stream as it deals with the potential to find clues for fraudulent manipulations in annual reports. Access to SEC enforcement actions and annual reports from the EDGAR system provides the basis for the study's *ex post facto* design. Through a mixed-model approach, qualitative and quantitative data are utilized both solely and in conjunction to create sound and comprehensive detection models. The construction and validation of the detection models are carried out under the guidelines of design-science and common machine learning practices. In the following section, the general structure of the study will be outlined.

1.2 Structure of the Study



1. Introduction

2. Theoretical Foundation

Fundamentals of Financial Statement Fraud

Fraud Theories

Fraudulent Schemes

Participants of Fraud Detection

3. Literature Review

4. Methodology

Sampling

Feature Generation

Machine Learning Methodology

5. Results

Design Questions

Enhancing Questions

Limitations and Suggestions for Further Research

6. Conclusion

Figure 1 – Framework of the study

The study is divided into six parts, as depicted in Figure 1. After the introduction in the first chapter, the theoretical foundation is outlined, with chapter 2 comprising four subchapters. First, definitions of fraud in general and financial statement fraud in particular are discussed and the economic implications of fraud are presented. Thereafter, an overview of influential and elaborated fraud theories is provided, with a special focus on potential fraud factors that can be incorporated into the fraud detection model. In the following subchapter, typical schemes regarding the fraudulent manipulation of financial statements are presented. As with the previous chapter, potential factors that may help in identifying fraud are discussed and identified for the future goal of developing a comprehensive fraud detection model. In the final subchapter of this part of the study, the participants involved in fraud detection and their need for reliable fraud detection models are outlined.

In the third part of the study, a broad literature review of qualitative empirical accounting research regarding the examination of narratives from annual reports for fraud detection purposes is presented. In conjunction with chapter 2, the research questions and hypotheses are defined.

The fourth chapter begins by describing the sampling process, before explaining the generation of the qualitative and quantitative features of the fraud detection models. The following subchapter describes the machine learning approach, first by highlighting the validation procedure and the learning methods on which this study relies, before explaining the four classifier approaches. For each classifier, the results for the hyperparameter tuning process are highlighted and later utilized to answer the research questions.

The results in chapter 5 are presented following the questions and hypotheses derived from the main research goal in the third chapter. A distinction between design and enhancing questions is made: whereas the former deal with the general set-up of the fraud detection model, the latter are designed to validate the results and increase accessibility, comparability and relatability. Following the presentation of the results, the limitations are discussed and further research possibilities are outlined.

The study closes with a summary and a conclusion, outlining the usefulness of the results.

2 Theoretical Foundations

Financial statement fraud detection, as suggested by the term, comprises three parts that stand for the different disciplines primarily involved in the topic, namely accounting (as regards financial statements), fraud theory and its borrowing from sociology, psychology, criminology and computer science (through the automated detection approach relying on a machine learning foundation). In the second chapter, the theoretical basis behind the accounting and fraud background will be laid out to ensure the thorough development of a reliable financial statement fraud detection approach in the second half of this study.

2.1 Fraud and Financial Statement Misrepresentation

In its 2018 Report to the Nations, the Association of Certified Fraud Examiners (ACFE)[8] assumed that companies lose 5% of their annual revenue to all kinds of fraud that occur in the corporate sphere.[9] This amount represents the mean loss of over 2,000 estimations from fraud experts around the world. When talking about fraud, this study focuses on the manipulation of financial statements. However, fraud is diverse and comes in different forms. In the following sections, the different types of fraud will be identified and separated, with particular attention paid to financial statement fraud.

2.1.1 Categories of Fraud

Fraud occurs in different forms. By identifying similarities and differences between fraudulent actions, a structure can be established to help elaborate on particular occurrences of fraud (such as financial statement fraud) within the greater picture. A commonly adopted classification of fraudulent schemes is the fraud tree released by the ACFE (e.g. Singleton & Singleton, 2010; Zack, 2013).[10] The original version is discussed in the following paragraphs and depicted in Figure 2, although it has been adjusted and modified to encompass a number of special cases and developments in the fraud literature, the basic structure remains unchanged (e.g. Mackevičius & Kkazlauskienė, 2009). Instead, additional

[8] The ACFE is a non-profit organization founded in 1988 in Austin, Texas. It focuses on the detection and prevention of fraud and white-collar crime and offers educational training in the field.

[9] ACFE (2018), Report to the Nations – Global Study on Occupational Fraud and Abuse. Retrieved from https://www.acfe.com/report-to-the-nations/2018.

[10] For the latest version, see ACFE (2018), Report to the Nations – Global Study on Occupational Fraud and Abuse. Retrieved from https://www.acfe.com/report-to-the-nations/2018.

categories are included or existing categories are structured in different ways (Sabau, 2012, pp. 110–112). Relatedly, other fraud taxonomies tend to have a high degree of similarity with the original fraud tree (e.g. Albrecht, Albrecht, Albrecht, & Zimbelman, 2016, pp. 9–13; Gottschalk, 2018, p. 4).[11] Common differences from other typologies are attributable to the scope of fraud with which the ACFE is mainly concerned, that is, occupational fraud (Albrecht et al., 2016, p. 10). This covers the fraudulent actions of employees or owners that result in direct or indirect damage to the organization; in contrast, fraud committed on a customer level like insurance or credit card fraud as well as fraudulent actions outside of the business sphere such as charity fraud are not covered.

The fraud tree, which is regarded as the most comprehensive blueprint of occupational fraud, divides fraud into three main categories: corruption, asset misappropriation, and financial statement fraud. The categories are further structured into subcategories, spanning across 58 fraud schemes. A scheme is a plan or an arrangement used to attain a particular object, in this context to benefit the perpetrator through the fraudulent action (Gao & Srivastava, 2007, p. 3).

Corruption covers schemes in which an employee abuses his or her influence or capability in a business-related transaction in a way that contravenes his or her duties as an employee to obtain a direct or indirect benefit (Albrecht et al., 2016, pp. 522–524). Such schemes include conflicts of interest, bribery, illegal gratuities, or economic extortion. A basic example of corruption under conflicts of interest would be a purchase scheme in which a vendor overbills a company in business-related transactions in which an employee of the company has an undisclosed interest. Corruption schemes are often based on related-party transactions in which the relationship is rarely known (Singleton & Singleton, 2010, pp. 83–84).

Under asset misappropriation, schemes are categorized in which employees steal, abuse or misuse the resources of their organization (Albrecht et al., 2016, pp. 512–515). Asset misappropriation usually affects resources within the personal sphere of influence of the respective employee, which are typically those they are entrusted to manage. The schemes can be separated in terms of the resources of interest to the perpetrator, mostly cash and assets, for example from inventories. Schemes comprise larceny, skimming and fraudulent disbursements like check tampering.

[11] See Singleton and Singleton (2010, pp. 54–68) for a comprehensive overview of different fraud taxonomies.

Financial statement fraud captures schemes in which employees intentionally misstate or omit information, leading to materially altered financial information of the organization.[12] The involved employees are often in management positions, which led to the development of the term “management fraud” that is often used analogously for financial statement fraud (Albrecht et al., 2016, p. 10). Typical schemes can be categorized in terms of net worth/net income over- and understatements. Overstating revenue or understating expenses are common schemes, resulting in misleading, overly positive financial statements. Financial statement fraud is usually committed by executives of organizations out of personal motivation such as bonuses or shareholder pressure (e.g. Rezaee & Riley, 2009, pp. 4–7).

[12] A detailed discussion of definitions of financial statement fraud will be given in the subsequent section 2.1.2.

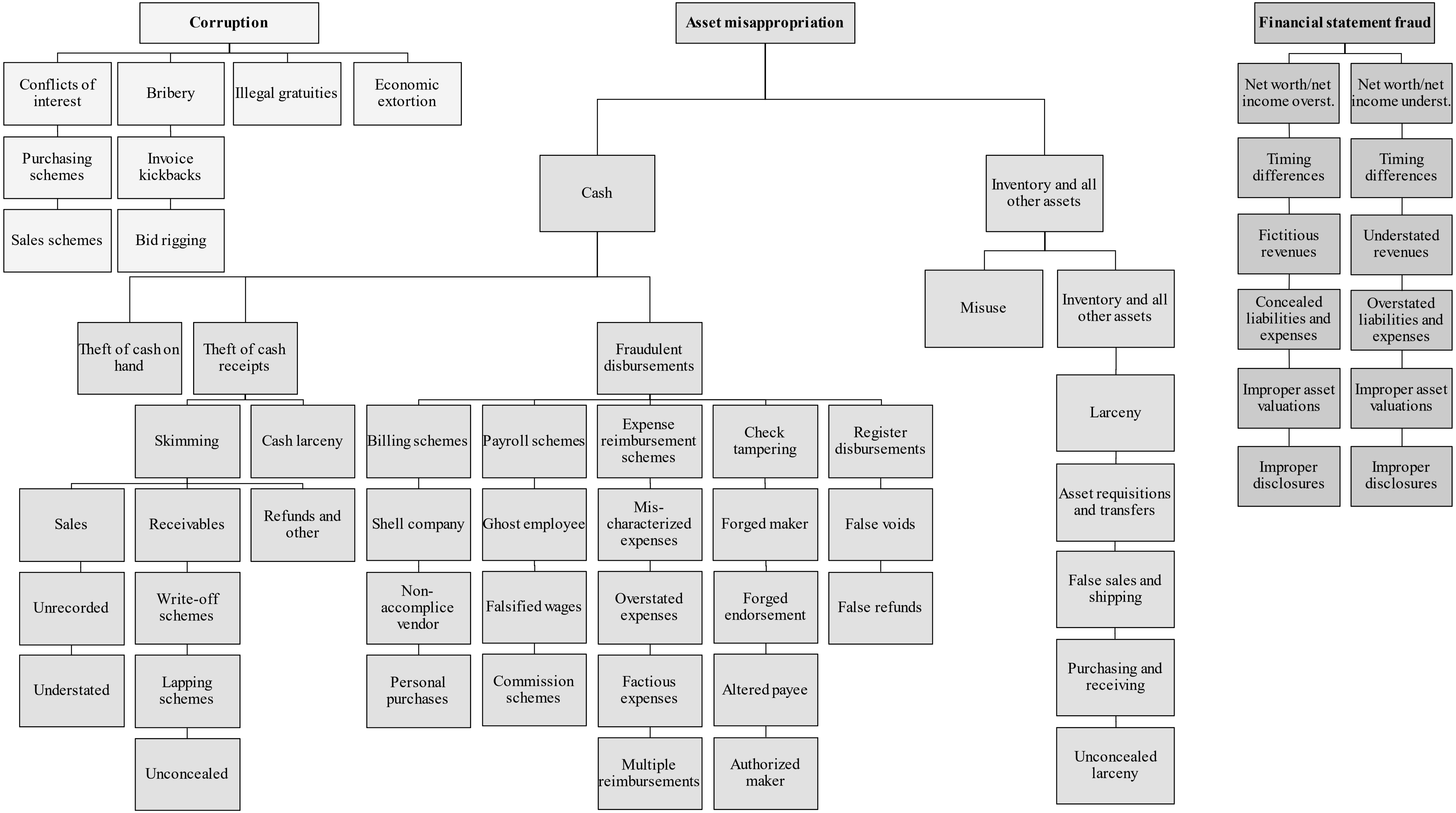


Figure 2 – Fraud tree

The three categories overlap to a certain extent, as cases of occupational fraud examined by the ACFE show. Figure 3 presents the distribution of the 2,690 cases reported from 125 countries across the three categories for the report of 2018 as well as their overlaps. Most commonly, schemes are related to asset misappropriation and corruption, whereas financial statement fraud occurs least frequently. Where cases fall into multiple categories, asset misappropriation combined with corruption constitutes the most frequent overlap, followed by a combination of schemes from all three categories. Those broader schemes can be explained by the general process of committing fraudulent actions, which often includes cover-up techniques, resulting in additional violations (Zack, 2009, pp. 7–8). In particular, asset misappropriation schemes are predestined for concealment and therefore potentially lead to financial statement fraud.

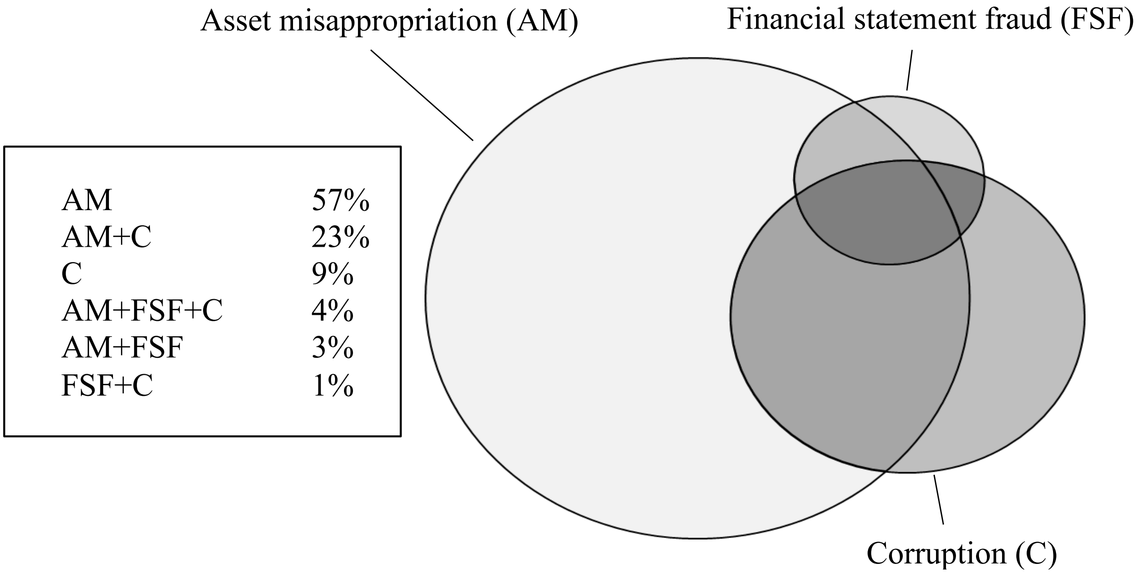


Figure 3 – Overlap of fraud categories

By combining the most recent descriptive statistics of occupational fraud by the ACFE with the examination of fraudulent actions by Singleton and Singleton (2010), Zack (2013) and Gottschalk (2018), Table 1 provides a broad overview of the categories and their characteristics. The cases are not exclusive to one category and may overlap. Relative to older reports, the overall figures show temporal stability, thus only the most recent figures are reported in Table 1.[13]

[13] Median losses represent the only category where the values change considerably over time, especially for financial statement fraud. For further details and the assessment of cost-sensitive results, see section 5.2.4.

Descriptors	Corruption	Asset misappropriation	Financial statement fraud
Usual fraudsters	Insiders and outside accomplices	Mixed	Multiple insiders
Median loss	$250,000	$114,000	$800,000
Frequency*	~14%	~87%	~9%
Benefactors	Fraudster	Fraudster (against company)	Company and fraudster
Industries with highest relative frequencies of cases per category**	Energy (1), manufacturing (2)	Services (1), arts, entertainment and recreation (2)	Construction (1), technology (2)
Most likely overlap of categories	Asset misappropriation	Corruption	Asset misappropriation
Most common concealment method	Creating/altering fraudulent physical documents	Creating/altering fraudulent physical documents	Creating/altering fraudulent physical documents
Median duration	~22 months	~18 months	~24 months
Most likely internal control weaknesses**	Lack of internal controls (1), overriding of existing controls (2)	Lack of internal controls (1), lack of management review (2)	Lack of internal controls (1), poor tone at the top (2)
Source of detection**	Employees (1), internal control (2)	Employees (1), internal control (2)	Employees (1), internal control (2)

* Sum can be larger than 100% due to overlap of categories.
** (1) and (2) indicating the first and second most likely control weaknesses, source of detection or first and second most common industry.

Table 1 – Characteristics of fraud categories

The overview in Table 1 suggests that although financial statement fraud occurs the least often, the median loss per case is considerably higher than that for the reaming categories. Moreover, financial statement fraud schemes last the longest, with an average length of about two years before being unveiled. The deviation in the relative occurrence of schemes from particular categories across industries suggests that different schemes are predestined or perhaps easier to carry out in certain industries (Gottschalk, 2018, pp. 19–24). The percentage of cases of financial statement fraud to total cases in the industry is lowest for energy with 3% (in comparison to 16% of total cases for construction and technology), while energy reports the highest percentage of corruption, with 53% of total cases. This overview suggests the existence of industry-specific characteristics that influence the occurrence of particular schemes.

Regarding internal company characteristics, the methods of concealment, the control weaknesses, and the detection sources can further be used to characterize the three different types of fraud. The internal factor that influences the occurrence of fraud the most according to the studies is the configuration of the internal control system. Internal control weaknesses in particular create opportunities to commit fraud of all types. The possibility of overriding internal controls is common to corruption schemes, while a lack of management review is more concerned with asset misappropriation. For financial statement fraud, poor tone at the top, which may be translatable into corporate culture, is the internal control factor that fosters fraudulent behaviour the second most often.[14]

The primary method of concealment and the sources of detection are rather similar for different types of fraud. The primary method for each category is the alteration of documents, which is deemed necessary to avoid obvious salience. The primary source of (initial) fraud detection according to the ACFE and Gottschalk (2018, pp. 31–40) comprises tips, often from employees.[15] Considerably less common in second place are internal control systems.[16] Unfortunately, both studies only provide fundamental insights into the different sources, rendering it almost impossible to make accurate statements as to the specific categories. For this study, financial statement fraud is the primary concern, although the preceding overview adds to the general understanding given the interconnectivity of different fraudulent actions. The following sections will take a deeper dive into financial statement fraud and facilitate an in-depth understanding of the subject, before discussing fraud theory and identifying ways to detect this undesirable behaviour.

2.1.2 Defining Financial Statement Fraud

Examining the reasons behind and finding ways to detect financial statement fraud has been the objective of regulators, governmental institutions, auditing and accounting associations

[14] Fraud theories related to corporate culture are further discussed in section 2.2.

[15] Gottschalk (2018) utilizes a different taxonomy for fraudulent schemes and examines cases of white-collar crime that encompass similar but not identical categories to the ACFE’s occupational fraud report.

[16] For additional detailed information about the participants of financial statement fraud detection, see section 2.4.

and, scholars alike (Vanasco, 1998).[17] Defining fraud properly is essential to reliably and clearly distinguish relevant cases from plain errors or honest mistakes. The most influential definitions are offered by standard setters concerned with financial statement fraud as well as scholars studying the subject.

External auditors are provided with guidelines to help identify fraudulent reports. In the USA, the American Institute of Certified Public Accountants (AICPA)[18] has been publishing auditing standards since the 1940s.[19] Within the AICPA, the Auditing Standards Board (ASB) is responsible for standard setting. In the aftermath of the major accounting scandals of the 2000s and the emerging lack of trust in external audits, the Sarbanes Oxley Act was enacted in 2002, resulting in the creation of the Public Company Accountant Oversight Board (PCAOB). The PCAOB oversees the audits of public accountants and should restore trust in audited financial statements. The PCAOB's authority (among other areas) lies in the issuing of applicable auditing standards. However, through SOX §103, the PCAOB can also adopt modified or unmodified standards from other institutions such as those issued by the ASB. Having two standard-setting bodies has led to a certain degree of divergence between the standards issued by the ASB and the PCAOB (Cullinan, Earley, & Roush, 2013, 7-8). Particularly applicable in the financial statement fraud context are the AU-C Section 240 – Consideration of Fraud in a Financial Statement Audit[20] (Supersedes AU Section 316) issued by the ASB and the AS 2401 – Consideration of Fraud in a Financial Statement Audit[21] issued by the PCAOB.

The definition of fraud is similar to that of the AU-C 240 and the AS 2401. A misstatement is deemed fraudulent if the action leading to the misstatement was made intentionally, as opposed to errors, which occur unintentionally.[22] According to both guidelines, the two relevant types of misstatements arise either from fraudulent financial reporting or from the misappropriation of assets.[23] For both types, an intention to misstate financial statements is the core characteristic of fraud. According to AS 2401.06,

[17] See Vanasco (1998) for an extensive literature review of definitions and the role of professional associations, governmental agencies and international accounting and auditing bodies in promulgating standards for fraud prevention.

[18] Similar institutions exist in other countries, such as the Institute of Public Auditors (IDW: Institut der Wirtschaftsprüfer - Institute of Public Accountants) in Germany, which releases the IDW's pronouncements like the IDW Auditing Standards or IDW Accounting Principles.

[19] The organization was named Institute of Public Accountants until 1957.

[20] Retrieved from https://www.aicpa.org/Research/Standards/AuditAttest/DownloadableDocuments/AU-C-00240.pdf.

[21] Retrieved from https://pcaobus.org/Standards/Auditing/Pages/AS2401.aspx.

[22] Compare AU-C 240.03/AUC-C 240.11 and AS 2401.05.

[23] Compare AU-C 240.03 and AS 2401.06.

intentionality can be assumed when "(m)isstatements arising from fraudulent financial reporting are intentional misstatements or omissions of amounts or disclosures in financial statements designed to deceive financial statement users where the effect causes the financial statements not to be presented, in all material respects, in conformity with generally accepted accounting principles (GAAP)". Before discussing intent, AU-C 240.A1/A2 hints at the three factors of the fraud triangle (opportunity, rationalization and incentive/pressure) that when combined instigate fraudulent actions, before describing the potential circumstances surrounding fraudulent behaviour.[24] AU-C 240.A4 states that determination of intent falls beyond the scope of an audit, whose "objective is to obtain reasonable assurance about whether the financial statements as a whole are free from material misstatement, whether due to fraud or error". The standard further describes how fraudulent manipulations may be accomplished as well as their characteristics in paragraph A5 – A7.

The definitions that scholars have developed over time tend to be rather similar, typically differing only in wording or small nuances to emphasize specific details. Elliott and Willingham (1980, p. 4) define financial statement fraud as "the deliberate fraud committed by management that injures investors and creditors through materially misleading financial statements". This definition has not significantly changed over time and still contains the relevant constituent characteristics of financial statement fraud. Goel, Gangolly, Faerman, and Uzuner (2010, p. 27) have created a similar definition, simply broadening the scope regarding the aggrieved party to view financial statement fraud as the "illegitimate act, committed by management, which injures other parties through misleading financial statements". Moreover, Wallace (1995), Flesher (1996) and Arens, Loebbecke, Elder, and Beasley (2000) have all integrated deception and concealment in their definitions as an additional factor of financial statement fraud.[25]

Both streams of definitions highlight the intentionality of the perpetrators behind the actions and the deceptive nature of the manipulations, transcending mere misreporting in annual reports. In this regard, not every misstatement is financial statement fraud. Empirically studying fraud is accompanied by the problem of identifying fraudulent cases. Therefore, it is usually necessary to rely on external sources that have identified fraud or that are eligible to serve as proxies for fraud. Potential proxies of fraud and typical ways in which fraud is identified in studies of financial statement fraud will be discussed in section 4.1.1.

[24] The fraud triangle is discussed in section 2.2.2.

[25] See Vanasco (1998, pp. 4–6) for a detailed discussion of fraud definitions.

2.1.3 The Evolution of Financial Statement Fraud

To understand and identify fraudulent actions and thereby derive prevention, detection, and deterrence mechanisms, it may be worth exploring the evolution of financial statement fraud, particularly attending to cases and developments that have helped to facilitate the development of fraud detection-related disciplines. Occupational fraud has most likely existed since the beginning of commerce (Dorminey, Fleming, Kranacher, & Riley, 2012, p. 556). Historically, identifying trustworthy market participants was laborious and largely relied on rudimentary biometrics (Woodward, Orlans, & Higgins, 2003, pp. 25–26). Detecting and deterring fraudsters has been relevant ever since. The shortcomings of modern corporations through abuse and fraud was recognized by Adam Smith (1776, p. 130) with the first major cases of occupational fraud probably dating back to the late 17th and early 18th centuries to the English East India Company (Robins, 2007, p. 37). The Company, which the English East India Company was also referred to, started as a marginal importer of spices and developed into one of history's biggest multinational corporations, pairing a trading monopoly with military power (Robins, 2007, pp. 31–34). Its eventual fall can most likely be traced back to management malpractices driven by shareholder pressure to yield immediate and excessive returns and lacking regulatory supervision and internal control (Robins, 2012, pp. 85–88). With regard to financial statement fraud, the first major cases may also be dating back to the same period of excessive trade. With the South Sea Bubble emerging in the early 18th century, the South Sea Company, which was founded in 1711 and was funded through government bonds, claimed exclusive trading rights with Spanish South America (Singleton & Singleton, 2010, p. 3). Given that its profits were lower than expected, the South Sea Company started to require additional funds. By circulating false reports about trade success, the company stock climbed, reaching unrealistic heights. As soon as the directors involved in the scheme came to sell their stock, confidence in the company eroded and eventually, the stock crashed. The government began to investigate the books of the company, revealing a massive accumulation of fraud and corruption with both company and government officials involved. Although this was considered one of the first major accounting and fraud scandals, it also prompted the advent of chartered accountants in Great Britain and the certified public accountant (CPA) profession (Singleton & Singleton, 2010, p. 4).

Another historical scheme worth mentioning is the case of Kreuger & Toll of 1932 (Lindgren, 1982). The international conglomerate covered a multitude of businesses but was most renowned for its near-monopoly in the match industry. Its securities were widely held in Europe and the USA. Its size and consequent operational complexity could hardly be captured in its financial statements because feasible accounting principles did not exist at that time. After Ivar Kreuger died in 1932, his manipulations of the financial statements were revealed and resulted in immense losses for investors. The scheme led to increased demand for audited financial statements and thus brought about the importance of the auditing profession. It has even been said that the Securities Act of 1933 and the Securities Exchange Act of 1934 were influenced not only by the stock market crash of 1929 but also directly by the Kreuger & Toll schemes (Singleton & Singleton, 2010, p. 5).[26]

The significance of fraud has barely changed over time. However, the number and the severity of known cases increased sharply around the 2000s (Dechow, Ge, Larson, & Sloan, 2011, pp. 26–31).[27] Accounting fraud scandals like Waste Management (1998), Enron (2001), WolrdCom (2002), Tyco (2002), Conseco (2002) and HealthSouth (2003), just to mention some of the most prevalent ones, have shocked professionals in auditing and accounting as well as capital market participants (Rezaee & Riley, 2009, pp. 3–4). The magnitude of these cases is especially apparent when considering the largest bankruptcies in US history, with Enron ($65.5 billion), WorldCom ($103.9 billion) and Conseco ($61.5 billion) representing three examples of financial statement fraud that had severe consequences (total assets at the time of the discovery of fraud) (Abbasi et al., 2012, pp. 1293–1294). Measured in terms of market capitalization, the collapse of Enron resulted in a loss of $70 billion, and when combined with WorldCom, Qwest, Tyco and Global Crossing, its losses have been estimated as amounting to up to approximately $460 billion (Rezaee & Riley, 2009, p. 14). To understand fraud and develop reliable and effective detection systems, it is essential to discuss the potential evolution of fraudulent behaviour and cases that may have resulted in its exacerbation.

In recent literature, scholars and practitioners have identified a number of (albeit changing) factors that in the last 20 years have fostered fraudulent behaviour, with particular focus on manipulations of company financials (Zack, 2013, XIV). Before presenting an explanation for the accumulation of fraudulent cases, it should be mentioned that the number of cases detected greatly depends on the engagement of regulators and capital market supervision and it is therefore highly contingent on the release of fraud detection or

[26] The acts required financial statement audits for listed companies.

[27] See also Figure 12 and section 4.1.1 for additional information on the distribution of cases in this study.

prevention acts, for example the Dodd-Frank Whistleblower Program, which led to an increase in pending cases for the years around 2010 (Jackson, 2015, pp. 27–30). Changes in detection rates are additionally attributable to the general underlying circumstances, like social, environmental, or political factors. However, despite all efforts, the mere existence of fraud or different fraudulent schemes in general did not change. Rezaee and Riley (2009, pp. 3–4) have examined various types of fraudulent schemes over a considerable timeframe between the 1980s and the early 2000s and have argued that history appears to be repeating itself, even despite political reforms and the efforts of professional institutions to improve fraud prevention, detection and deterrence. Although the severity of schemes varies, the authors have found cases at both ends of the timeframe for each type of scheme, with a clustering of cases shortly after the millennium. The factors that seem to have affected this trend can be summarized into three categories, the first two dealing with the social and economic environment at a private or a business level and the third factor with the opportunity to commit fraud.

The first reason revolves around changes in trading behaviour as institutional investors began to focus more on raw financials than did private investors, whose investment decisions were more likely based on a belief in the products and the vision of the company, resulting in increased pressure on executives to present compelling numbers and thereby satisfy their investors (Jackson, 2015, p. 31). The second reason pertains to a potential change in the personalities of executives. Achieving and maintaining affluence and its associated social status may lead to unethical and perhaps greedy behaviour (Tunley, 2011, p. 314). Both reasons are combined when management compensation is tied to financial performance for example through stock options, which became increasingly popular around the 2000s (Jackson, 2015, p. 31). A third factor concerns the auditing profession and its role in determining financial statement fraud. Until the mid-1980s, the profession scarcely perceived itself as having a responsibility to detect fraudulently altered statements, leading to a lack of required skills and tools, rendering it ill-equipped and insufficiently interested in fraud detection at the beginning of the millennium (Wells, in Zack, 2013, XIV).[28] Jackson (2015, p. 32) further argues that the auditing companies became increasingly dependent on additional consulting services, which developed to become the major source of revenue for the branch. Combined with the lack of auditor rotation, satisfying their customers in an unethical way by compromising audits potentially contributed to the surge of cases.

[28] The so-called expectation gap and the role of auditors will be highlighted in section 2.4.6.

Additionally, weak and ineffective internal controls are supposed to have created opportunities for fraudulent actions (Jackson, 2015, p. 34).

Fraud is not static. Rather, it evolves with the environment and adaptive detection models are necessary to account for this situation (Zhou & Kapoor, 2011). Singleton and Singleton (2010, p. 7) depict the fraud environment as a pendulum, swinging from one extreme to another, barely resting and in constant movement. They suggest that human nature, as well as business and legislative cycles combine to influence the fraud-bearing environment and thus the occurrence of fraudulent behaviour. Lee, Ingram, and Howard (1999, p. 783) suggest that it will never be possible to detect discreet fraudulent actions with ease, thus qualifying the mixed and sometimes seemingly poor results of fraud detection approaches and outliers in detection performance.[29] To ensure the reliable detection of fraudulent cases and to consider the pervasive changes made to combat fraud, like SOX in 2002, it is necessary to test fraud detection models over longer timeframes and in real-world environments.

2.1.4 Market Efficiency and the Fraud-on-the-Market Doctrine

Having discussed the evolution of financial statement fraud, the ramifications for capital markets need to be ascertained. According to the efficient capital market hypothesis (ECMH), all stocks are perfectly priced according to their inherent investment properties, the knowledge of which all market participants equally possess (Fama, 1970). The ECMH comes in three forms (weak, semi-strong and strong) dependent on the nature of the information incorporated in the prices.

The weak form states that only trading information regarding the instruments (e.g. shares or bonds) is already incorporated in prices. This type of information refers to historical information such as prices and volume. Thus, Fama (1970, pp. 386–387) suggests that prices follow a random-walk model, entailing the inexistence of patterns in prices and implies that future price movements are solely determined by information not contained in the price series.

The semi-strong form then deals with the absorption of publicly available information such as announcements regarding the instrument, for example, the issuance of additional shares or financial disclosure like annual reports. The speed of the absorption of new

[29] A comparison of the results of relevant studies in financial statement fraud detection can be found in table 30 in section 5.2.5.

information is so high that no excess returns can be generated by trading with reliance on new public information. Moreover, the information absorbed is unbiased and therefore homogeneous across the market participants. In this way, an unexposed misstated financial report will influence prices based on manipulated information (Korsmo, 2014, pp. 13–14).

The strong form postulates that all information, both publicly available and private, is incorporated in prices. In this case, an information monopoly does not exist and no one can earn excess returns. Fama (1970, p. 415) notes that the strong form is to be regarded as a benchmark of market efficiency against which deviations can be judged. Deviations at this point have already been brought forward, especially in regard to monopolistic information of specialized traders or company insiders (e.g. Niederhoffer & Osborne, 1966; Scholes, 1970). With regard to fraud, perpetrators (e.g. managers) would enjoy a monopolistic information advantage if the fraudulent action were not uncovered or carried to outsiders. However, it may be possible due to managers' trading behaviour, like exercising stock options (for example during unexpected times) that the information monopoly is weakened and hints of fraud are carried outside (Grove, Cook, Streeper, & Throckmorton, 2010, pp. 284–285).

The ECMH has proved effective in addressing capital market inefficiencies. However, a degree of dissatisfaction has developed in the academic community. Bloomfield (2002) has formulated the incomplete revelation hypothesis (IRH) to address the underreaction of prices to the release of accounting information. The incomplete revelation hypothesis states that information that is harder and therefore more costly to extract from financial reports is less completely revealed in prices. An important implication for the fraud detection literature is presented by managers' decisions to make it more difficult to extract certain types of information from financial reports (for example bad news) and therefore they alter disclosure in such a way as to obfuscate its true nature (Courtis, 1998, pp. 461–462). The IRH, according to Bloomfield (2002), extends upon the ECMH and seeks to explain anomalies in price reactions as well as the disclosure decisions of companies.

The impact of manipulated information on market prices is not only a problem from an economic but also from a legal perspective. The deception of investors who rely on the accuracy and efficiency of market prices has been debated extensively in the legal sciences (e.g. Macey & Miller, 1990). A number of corporate and securities laws are based upon the ECMH, including insider trading rules and more importantly for this study the fraud-on-the-market doctrine (FOTM) (Korsmo, 2014, p. 5). The FOTM has been created by US courts and postulates that causality between untruthful information and resulting losses does not require any proof; moreover, the release of untruthful information to the public is identical to the release of untruthful information to a certain party (e.g. Fischel, 1989, p. 908). Even if no actual party directly suffers from the manipulated information, fraud on the market is perceived as if there was an aggrieved party. Even if the purchase of stock or security does not directly rely on the information (but rather the information is processed into the market price), the company can be made responsible for the damage.[30] In the case of *Basic, Inc. v. Levinson*, investors pleaded that the allegedly misrepresented information was publicly known, that the misstatement was of material nature, that the stock was traded in an efficient market and that the plaintiff traded the stock in the relevant period. The court furthermore stated that the link between the misrepresentation and the price the plaintiff paid must not be severed; otherwise, the FOTM presumption does not apply.

Prior to FOTM, it was almost impossible to attain class certifications in fraud cases (Korsmo, 2014, p. 10). The courts found that first, most investors would not have read or heard the fraudulent financial report. Second, it was virtually impossible to prove that every individual investor had read the fraudulently manipulated report. (Nevertheless, in the 1960s, even before the postulation of the ECMH and its adoption into the legal academy, opinions existed about investors having to prove that they had actually read the misstated disclosure but relied on market prices to take investment decisions (Korsmo, 2014, pp. 11–12).

Considerable criticism has been expressed towards the seemingly arbitrary adoption of the ECMH in the FOTM assumption. The courts did not postulate the form of the ECMH upon which they actually relied, especially problematic given the limited applicability of the theoretical underpinnings of each of the forms to financial statement fraud; they also failed to sufficiently emphasize the empirical verifications and instead focused on the theoretical concept of the ECMH (e.g. Carney, 1989; Macey & Miller, 1990; Jovanovic, Andreadakis, & Schinckus, 2016). The criticism did not only concern the FOTM assumption but regulators, too, as they relied inconsistently on the ECMH and its empirical evidence (e.g. Saari, 1977). Nevertheless, the criticism barely compromised the general idea of the conjunction of financial theory and the legal discipline, instead being more concerned with the manner of adoption.

The FOTM assumption illustrates the reliance of the market on the truthfulness of corporate disclosure and incorporates economic theory behind the absorption of corporate

[30] Relevant cases: Basic, Inc. v. Levinson (485 U.S. 224, 1988) and Erica P. John Fund, Inc. v. Halliburton Co. et al. (563 U.S. 804, 2011).

information into corporate law. The intersection between accounting, finance, economics, and the legal discipline emphasizes the interdisciplinary nature of fraud and again demonstrates the importance of comprehensive fraud deterrence, prevention, and detection mechanisms to ensure functioning capital markets. The following section will highlight fraud theories to enhance understanding of fraudulent behaviour and identify aspects that might help to develop a sound detection model.

2.2 Fraud Theory

Fraud theories are created to explain the circumstances and motivation behind fraudulent actions, facilitating an understanding of their origins and providing a theoretical framework that can be utilized to develop tasks and tools to counter this unwanted behaviour (Elsayed, 2017, pp. 2–4). Therefore, fraud theory is characterized by interdisciplinarity, covering topics from sociology, psychology and criminalistics (Ramamoorti, Pope, Morrison, & Koletar, 2013, p. 6). The objects of interest in most theories are usually the perpetrator, the action, and the environment. A holistic approach seems to be necessary to catch fraudulent activities in their entirety, although to date this has rarely been attempted (Dorminey et al., 2012, pp. 570–576).

The relevance of fraud theories for financial statement fraud detection has also been recognized by regulators and standard-setting bodies. For example, the American Institute of Public Accountants has adopted the fraud triangle in its Statement on Auditing Standards No. 99 – Consideration of Fraud in a Financial Statement Audit. The fraud triangle from Cressey (1950) is one of the most influential fraud theories and with its three basic factors of fraudulent behaviour (rationalization, opportunity and motivation) has laid the foundations for a number of theories that rely upon the original structure. Besides the fundamental theories, a number of new streams have been developed to capture all of fraud's peculiarities.

Maragno and Borba (2017) provide an extensive literature review of theoretical and empirical work on fraud and fraud theory. They highlight the evolution of fraud theories and depict the relationships between different approaches. The overview presented in Figure 4 will be used as a basis for the discussion of influential fraud theories and will be expanded upon to reflect recent developments as well as additional models. The chapter will close with a summarization of the models and an indication of factors that might help in developing fraud detection models that rely on publicly available data.

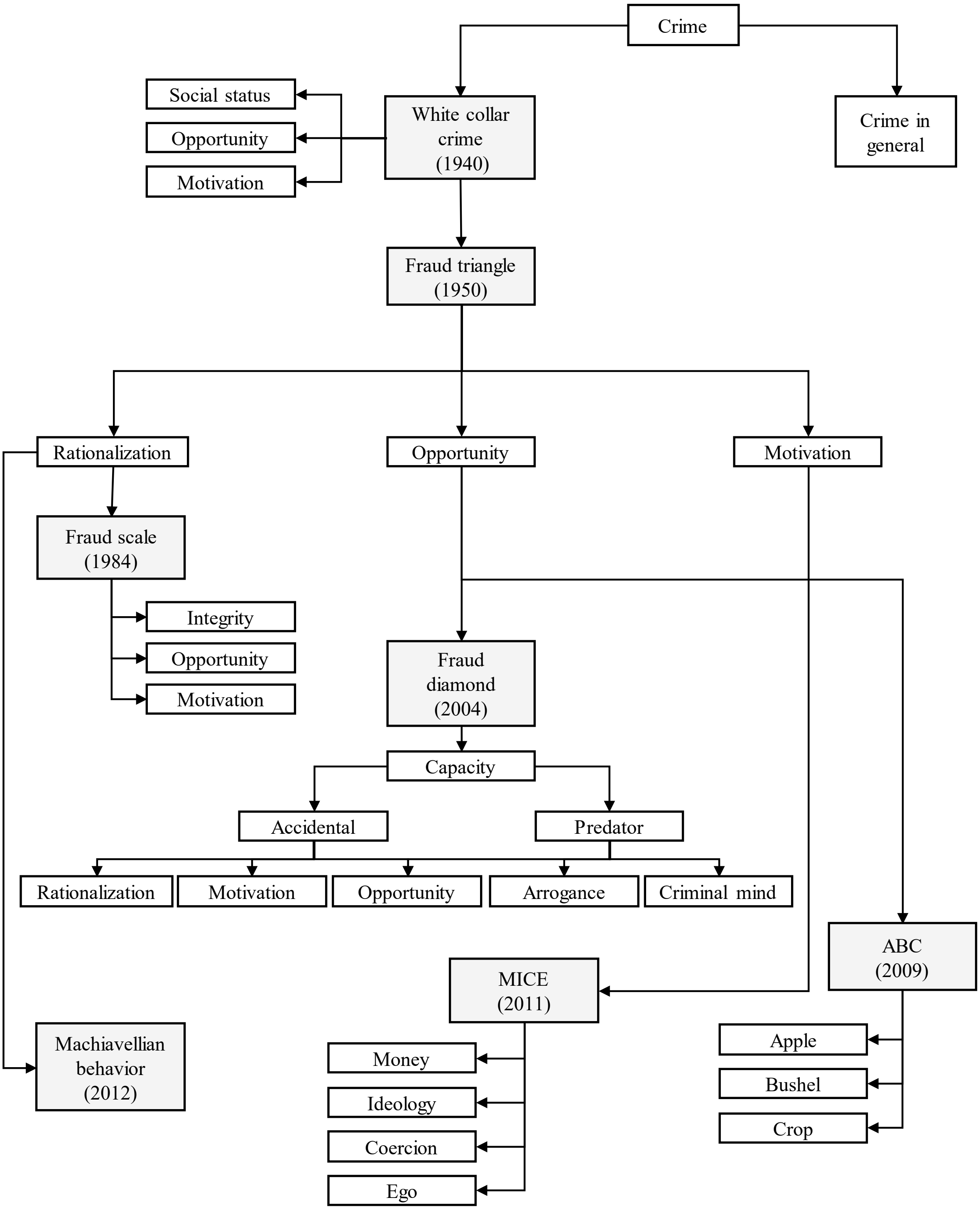


Figure 4 – Fraud theory overview[31]

[31] The fraud theory overview of Maragno and Borba (2017, p. 44) was adopted and complemented to reflect the additional theories discussed in this study.

2.2.1 White-Collar Crime

As well as introducing the term “white-collar crime”, Sutherland (1940) provided one of the first explanatory attempts of the subject. He examined crimes covering economic and business-related activities and differentiated a “new” type of crime from the predominant topics, which mostly focused on street crime and violence primarily caused by poverty and associated psychopathic and sociopathic conditions. Regarding his definition, white collar-crime is connected with violations of implied and/or delegated trust. It is obvious that the definition is closely connected to agency theory (Dorminey et al., 2012, p. 557). Sutherland distinguished two categories of white-collar crime: misrepresentation of asset values and duplicity in the manipulation of power. His work plays an essential role in the perception and awareness of white-collar crime, developing the influential differential association theory from it (Sutherland, 1939; Sutherland, 1940; Sutherland, 1944; Sutherland, 1947).

Sutherland’s work sought to fulfil two goals: to define and prove the existence and relevance of crime in an upper socioeconomic group; and to provide a generalizable theory behind crime in general (Dixon, 1995, pp. 561–562). Sutherland’s findings suggested that the main peculiarity defining and distinguishing white-collar crime from typical criminal activities at that time was the social status of the perpetrators. Indeed, the perpetrators were regarded as professionals with a high status in society, resulting in admiration that can lead to intimidation. He furthermore suggested that status would lead to fewer and less severe punishments because the criminal justice system was supposed to be less reliable in cases of white-collar crime, a supposition that was empirically confirmed by Snider (1982).[32] Lastly, he noted that the consequences of white-collar crimes were more difficult to grasp than for common crimes. This was mostly due to an inability to identify victims in the diluted timeframe in which the delinquency took place and the lack of organization of the victims (Dorminey et al., 2012, p. 557).

Besides the characterization of white-collar crime, Sutherland’s development of the differential association theory has provided an explanation of how a perpetrator becomes a criminal. He hypothesized that a person could learn criminal behaviour through interaction with others. This process encompasses values, motives, attitudes, and techniques. Hence, in an appropriate situation, when a favourable definition of crime outweighs an unfavourable

[32] Snider (1982) compared the sanctions for white-collar crimes to other traditional nonviolent property offences (e.g. theft or possession of stolen goods), finding that white-collar crime seemed to be less severely punished in most cases, as he hypothesized.

one, a person may engage in criminal behaviour. In this regard, Sutherland created the basis for most subsequent theories by explaining how a person develops the capacity to commit fraud and how he or she finds the opportunity to realize the fraudulent scheme.

Considerable criticism of the white-collar crime theory and its adaptation has been formulated. Perri, Lichtenwald, and Mieczkowska (2014, p. 77) criticize the lack of empirical evidence and the active work of Sutherland against the development of an interdisciplinary multi-factor approach.[33] Although Sutherland's model appears to constitute a multi-factorial approach to explain fraud, it mostly focuses on sociological factors while barely considering fundamental ideas from other disciplines like biology, economics or law (Perri et al., 2014, p. 74). Coleman (1987, pp. 434–435) argues that the theory has outlived reality and that the very broad concept of white-collar crime needs to be separated into different kinds of crimes. Moreover, the theory can only explain how the perpetrator becomes a criminal, rather than providing insights into the motivation behind it. Shapiro (1990, pp. 362–363) mentions the imprisoning nature of the framework of white-collar crime, which has dominated the fraud literature for many years and has in the process constrained the debate. Especially the focus on the sociological background of the perpetrator has limited the outcome to individual characteristics rather than the *modus operandi* of the crime (Dixon, 1995, p. 565). With the fraud triangle, which will be explained in the next section, the behavioural aspects of crime have been considered and some shortcomings of the white-collar crime theory have been overcome (Ramamoorti, 2008, pp. 524–526).

2.2.2 The Fraud Triangle

The fraud triangle, as depicted in Figure 5, is a popular model explaining the preconditions behind the fraudulent behaviour of individuals. It was developed from the same theory behind white-collar crime but seeks to generalize the factors to a greater extent and offers a broader explanatory fundament (Dorminey et al., 2012, pp. 557–558). The theory is based on a survey among prisoners convicted of embezzlement and has become the foundation of a large body of literature dealing with white-collar crimes (e.g. Cressey, 1950; Cressey, 1953). Similarities in answers to the survey were condensed to non-shareable financial/economic problems, the opportunity and knowledge to commit the violation, and

[33] Sutherland's standing in the sociological and criminological fields dominated the debate for years. For a comprehensive review, see Perri et al. (2014).

the ability to adapt one's self-perception in order to justify the violation as non-criminal. Therefore, the three factors every fraudulent action had in common were extracted and the triangle of pressure, rationalization, and opportunity was created.[34]

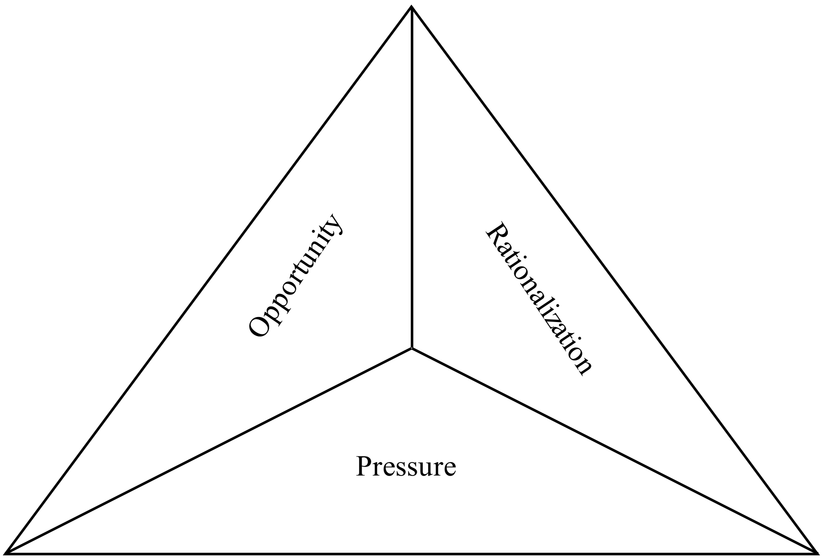


Figure 5 – Fraud triangle

Pressure, sometimes referred to as motivation or incentive, is the cause of the resultant action and consists of individual perceived needs (Cressey, 1950, p. 742). In the business context, pressure revolves around financial motives at an individual or company level (Schuchter & Levi, 2016, pp. 109–111). Overstating revenues to artificially boost the company's performance could be an incentive that affects the company level but may also have an impact on the individual sphere when the related performance goals are defined in a bonus scheme. Hence, the individual and business levels are difficult to separate in the underlying context, as they are generally closely connected. The self-enriching behaviour in this example can be altered to self-preservation when the risk of a negative impact on one's position in the company and associated financial and social status is concerned (Dorminey et al., 2010, pp. 18–19). In the aforementioned example of incentives, by not meeting the defined performance goal. Beyond plain financial motives, pressure may also arise through personal traits. The origin of fraudulent behaviour can rest in an individual's personality when egocentric behaviour like aggrandizing one's ego or an increased desire for power may result in illegitimate manipulations (Singleton & Singleton, 2010, p. 45). In general,

[34] Opportunity and pressure are sometimes described as perceived factors of the fraud triangle (e.g. Dorminey, Fleming, Kranacher, & Riley Jr., 2010) owing to the arbitrary and individual nature of the two factors. In the following, the two factors will only be termed "opportunity" and "pressure". Also see Lokanan (2015) for a deconstruction of arguments from the literature.

motivations, incentives or pressure occur in different ways and need to be identified on a personal level (Gottschalk, 2018, pp. 8–11).[35]

Rationalization, the second factor of the fraud triangle, addresses the justification of the perpetrators. Rationalizing the fraudulent behaviour must be in accordance with individual moral values. Violators often do not see themselves as criminals and justify their actions by external factors such as underlying circumstances. When referring to the example of overstated revenues and the resulting bonus payments, the justification could be the individual's belief that he or she deserves the gratification. Another commonly observed justification from Cressey (1953) deals with the fraudster's perception of the damage or harm resulting from the action. Given that the company is usually the victim, the consequent harm to indirectly affected individuals is not taken into consideration or is valued less. The reasoning may even involve benevolent justifications, where the fraud committed is done for the good of others or for a greater good in general (Singleton & Singleton, 2010, p. 46). Gottschalk (2018, pp. 24–28) has reviewed the neutralization theory, which was introduced by Sykes and Matza (1957) but only became popular in the field of fraud and white-collar crime in recent years. The theory explains the reasoning of fraudsters behind their violations. These individuals tend to deny the damage caused or ignore the victims associated with the damage and lack responsibility for the violation in general. The neutralization theory may help in identifying rationalization attempts and relies on 13 neutralization techniques. Some of these attempts relevant for financial statement fraud that have not been examined to date include the individual's "opinion" that the violation was a legal mistake that should not be forbidden or regulated. With regard to the complex nature of accounting rules, altering financial statements beyond legal permissions for whatever reason might be rationalized by the opinion of having the right to do so. Another closely related neutralization technique is a dilemma leading to a trade-off between benefits and costs, where the violation is accepted following consideration of all interests and alternatives.[36] Overall, the range of possible justifications is vast and is based on a multitude of factors that influence each individual's moral standards, rendering them difficult to identify.

The third factor of the fraud triangle is related to the knowledge and the opportunity to perform the actions causing or related to the fraud scheme. Knowing the internal control mechanisms and the associated probability of being detected plays an important role in

[35] Gottschalk (2018, p. 9) relied on Maslow's hierarchy of needs to identify motives on different levels.

[36] Per definition, a dilemma represents a state of mind in which it is not obvious what is wrong or right.

influencing the potential perpetrator to engage in fraudulent behaviour (Rezaee & Riley, 2009, pp. 67–68). The factor opportunity also deals with the inherent possibility of committing the crime, implying access to relevant resources and an ability to manipulate them accordingly. Access is usually granted through trust and a level of faith in the skills and honesty of an individual. In terms of overstated revenues, manipulators most likely come from the accounting department or are managers and executives with respective access to and knowledge of the systems. The development of colluding groups may also be possible (Ramamoorti et al., 2013, p. 52).

Cressey was not the only scholar to conduct field research by surveying fraudsters. Indeed, Dellaportas (2013) interviewed 10 male accountants who were serving custodial sentences for committing fraud or related offences. His goal was to examine the reasons behind fraudulent activities as well as the fraudsters' point of view as professionals in the field of accounting, bringing the answers in relation with traditional theories like the fraud triangle. His results suggested that opportunity rather than motivation (pressure/incentive) play a decisive role and may be the key to countering fraudulent actions. Schuchter and Levi (2016) gathered additional empirical evidence. In their study of 13 cases of fraud from Switzerland and Austria, they found that not all factors from the fraud triangle are a necessary precondition for fraud. In contrast to Dellaportas (2013), they noted that from a motivational standpoint, a monetary incentive is not sufficient, while pressure from inside the company is an essential driver behind fraudulent violations. Another empirical study was conducted by Skousen, Smith, and Wright (2009), who operationalized the three factors of the fraud triangle using 26 variables, mainly constructed of information from financial statements. They matched 86 fraud firms with non-fraudulent firms to test the influence of the suggested proxies using logit regression analysis. Five proxies for pressure and two for opportunity showed a significant impact. Overall, empirical evidence specifically testing factors from the fraud triangle remain rather scarce.[37]

Amongst professionals, the fraud triangle has become very popular in understanding fraud and developing the techniques to detect and deter it (Dorminey et al., 2010, p. 19). With the implementation of the fraud triangle in the Statement on Auditing Standard 99 – Consideration of Fraud in a Financial Statement Audit (2002), the fraud triangle has even

[37] However, the variables from Skousen et al. (2009) are based on other studies' results, e.g. Persons (1995) and Kaminski, Sterling Wetzel, and Guan (2004), which are not directly related to the fraud triangle factors. In the present study, a similar set of variables is tested, with further explanations provided in section 4.1.4.

become part of the auditing process, raising auditors' awareness of the factors behind fraudulent behaviour.

However, scepticism about the fraud triangle's (and to some extent, its conceptual offspring's) appropriateness in explaining fraudulent actions at more than a simplistic level (e.g. Lokanan, 2015; Huber, 2017). Moreover, the basic concept from Cressey (1953) was extracted from cases of embezzlement and is arguably not transferable to every kind of fraud (Huber, 2017, p. 31). Lokanan (2015) has noted that the focus on a strict and individualized framework like the fraud triangle may draw attention from other decision-making approaches that may explain fraud more effectively in particular cases. This is especially true for different types of fraud that cannot be explained by the fraud triangle, such as the existence of predatorial fraudsters. For such predators, motivation and rationalization play a minor if not unimportant role, as they will commit fraud by opportunity (Dorminey et al., 2012). Moreover is the implementation of the fraud triangle into the auditing standard criticised. Wilks and Zimbelman (2004) have compiled the fraud triangle and game theory to develop a best practice approach for auditors with the assistance of the fraud triangle framework. They suggest the basic implementation of the fraud triangle into the auditing standard to hardly represent a successful way of utilising fraud theory in the auditing practice and hint at individual, unpredictable auditing procedures to be best suited for the task of detecting fraud. The following sections will highlight improvements to the fraud triangle with the aim of tackling the aforementioned shortcomings and enhancing understanding of fraudulent behaviour.

2.2.3 Derivatives of the Fraud Triangle

The fraud triangle has served as the parent of a number of thematically related offspring, which have been developed since its origins in the middle of the 20th century.[38] The limitations of the fraud triangle lie in its perspective and in the typecast of the fraudster (Dorminey et al., 2010, pp. 19–22). Indeed, the fraud triangle operates from the fraudsters' perspective, rendering two factors (pressure and rationalization) unobservable. Furthermore, the fraud triangle does not consider pathological fraudsters or at least does not explain their actions. Dorminey et al. (2010, p. 21) postulate that the fraud triangle works well for accidental fraudsters but is less suitable for more deliberate types of fraud (predators or

[38] For a detailed discussion about the critics of Cressey's work, further read Rogovin and Martens (1992).

collusion). For those types, the explaining factor is mostly opportunity, rather than rationalization or pressure. The shortcomings were revised and the model was further developed into a fraud diamond that additionally includes incentive and capability (sometimes referred to as capacity) instead of pressure (Wolfe & Hermanson, 2004).[39] Capability and incentive are easier to observe and are also able to explain the fraudulent actions of different types of fraudsters, like predators. Incentive corresponds to pressure from the older model but is more closely tied to the business environment, where the observable pressure is more likely to be based on incentives due to compensation schemes. The factor of capability focuses on personal traits and abilities. In contrast to opportunity, which exists in the context of the fraud diamond rather than being associated with environmental factors, capability is directly targeted at the perpetrator (Schuchter & Levi, 2016, pp. 111–112). Capability plays an important role given opportunity and incentive. It is associated with the knowledge to commit the violation as well as character traits that might help during the fraudulent act and the process of rationalization.[40] Hence, Wolfe and Hermanson (2004, p. 38) have argued that although pressure and opportunity exist to commit fraud, alongside a potential rationalization of the crime, it is essential for the fraudster to have the capability to conduct and conceal the action. With the aforementioned changes to the fraud triangle, the fraud diamond seems to be better suited to explain fraudulent actions and develop detection and deterrence strategies.[41]

Another offspring from the fraud triangle that seeks to enhance understanding of some of its factors is the fraud scale. The fraud scale, akin to other theories like Machiavellianism, attempts to improve knowledge about fraud by facilitating understanding of one of the more complex factors derived from the fraud triangle: the rationalization of the fraudster. Albrecht, Howe, and Romney (1984) introduced the fraud scale as an improvement to the fraud triangle in the case of occupational fraud, especially for financial statement misrepresentation. In their extensive study, they surveyed internal auditors (internal audit directors) on fraud demographics, perpetrator characteristics, specific causes and auditor directors' perspectives (Albrecht et al., 1984). They attained data from 212 internal auditors during the 1980s and were able to build their fraud theory based on 82 different variables, mostly focusing on company and perpetrator characteristics. They argued that motivation

[39] Different names have been developed for the fraud diamond. For instance, Kassem and Higson (2012) refer to the fraud diamond as the "New Fraud Triangle".

[40] See section 2.2.6 and the effectuations on Machiavellianism for further understanding.

[41] For a detailed comparison of both models, see Abdullahi and Mansor (2015).

and pressure to meet analyst perceptions were already easily observable in this context. To better assess the likelihood of fraud, they omitted the rationalization factor and instead introduced a new factor, integrity, referring to ethical behaviour based on the personal code each person adopts. According to these authors, integrity is observable through a person's decisions and related decision-making processes. When the three factors are considered simultaneously, the fraud scale can provide a tool to assess the probability of fraud (e.g. high pressure and motivation and low integrity indicates a high probability of fraud).

2.2.4 ABC Fraud Theory

Social psychology offers explanation attempts of fraud committed through the management by studying the relationship between individuals and the organization (Greller, 1980, pp. 171–172). Potential aspects cover the individual's role in a group, basic group processes and the identification of an individual inside the organization (Greller, 1980, pp. 172–173). As the nature of management usually revolves around group processes, studying fraud from this very perspective may deliver additional insight into the occurrence of fraud.

The ABC theory for fraud was introduced by Ramamoorti et al. (2013) and refers to the organizational factors most closely associated with interpersonal relationships in companies. Ramamoorti et al. (2013, p. 23) suggest that fraud is a natural human activity that must be studied in terms of emotions and the state of mind, with greater emphasis thus placed on behavioural science rather than Sutherland's sociological approaches. The acronym ABC stands for apple, bushel, and crop. The three elements describe the three levels that should be considered when analyzing fraudulent behaviour. The first level is the individual sphere, the second level describes a colluding group, and the third level expands to the entire organizational culture. The wording originally referenced the bad apple spoiling its neighbours. Two main questions that need to be asked when examining fraud arise from this metaphor.

The first deals with the number of persons involved in a fraudulent action. Ramamoorti et al. (2013, pp. 49–50) present anecdotal evidence from cases of rogue traders who caused immense losses for their respective corporations. In most cases, it seems that the perpetrator acted on his or her own and without any help from others. However, it is to question if this is actually true, or if colluding groups were able to conceal their involvement, resulting in the incomplete exposure of the perpetrators. Greller (1980, p. 172) suggests that the three functions of a group, namely the task function, sentient function and the reference function play a major role in forming colluding groups of perpetrators in the case of management fraud. As Greller mentions, for example does the reference function help granting a level of rectitude of the fraudulent actions, while the sentient function offers a sense of relief and understanding and reduces alienating tendencies of the perpetrator through manipulating financial statements (which is the task function in this scenario because of the management's involvement in the preparation of the financial statements). These aspects might spread from certain groups and negatively influence the entire company, providing organizational and environmental factors that foster fraudulent behaviour. Especially the tone at the top may induce a trickle-down effect through the organization and result in a culture that may benefit the development of unethical behaviour like fraud (Rezaee & Riley, 2009, pp. 79–82). As mentioned in section 2.1.1, when categorizing the different types of fraud, poor tone at the top is deemed one of the most prevalent internal weaknesses in the case of financial statement fraud.

The second question arises from the most recent view of behavioural forensic on the ABC theory. With the bad apple spoiling its neighbours, the existence of remaining good apples is of interest for forensics. Macey (2013, pp. 83–88) argues that the coexistence of "good and bad apples" can very well be seen in recent scandals like Arthur Anderson, where employees remain valuable to the industry despite the company's collapse. In addition, employees who are reluctant or unwilling to converge towards the colluding fraudulent groups yield high potential for fraud detection through whistle-blowing mechanisms (Rezaee & Riley, 2009, p. 81). These mechanisms seem to operate as cases like WorldCom have shown, where whistle-blowing of the Vice President's internal audit led to the discovery of a major accounting scandal (Scharff, 2005). With regard to the business level, it has furthermore been assumed and found that individuals or groups further up in the management chain commit larger and more severe fraudulent actions (e.g. Heath, 2008, p. 600; Rezaee & Riley, 2009, pp. 59–63). Hansen (2009, p. 33) argues that occupational fraud committed by top-level employees or groups of top-level employees is especially difficult to detect because higher positions are associated with more trust. These positions are more difficult to oversee or supervise and provide a confined space for actions and arrangements. Overall, organizational structure and individual idiosyncrasies are regarded as a major factor within the ABC theory.

The implications of the ABC theory for fraud detection and deterrence efforts primarily rely on human resource management. Background checks and ethics training are regarded as viable solutions to counter the emergence or spreading of such behaviour. In particular, the spreading of unwanted behaviour should be tackled by establishing an ethical foundation that does not tolerate fraudulent behaviour. Internal controls are regarded as ineffective when colluding groups with top-level support engage in schemes (Ramamoorti et al., 2013, p. 52). Therefore, fraud factors capturing the characteristics of internal control systems may also be suited to help detect weaknesses and the greater opportunity for fraudulent behaviour.

2.2.5 MICE and SCORE Model

Another theoretical concept that provides explanations for fraudulent actions has the acronym MICE (Kranacher, Riley, & Wells, op. 2011, pp. 13–14). MICE is hardly comparable to the fraud triangle or diamond because it focuses on a less holistic approach, and instead focussing onto motivational factors for fraudsters (Dorminey et al., 2010, pp. 20–21; Kassem & Higson, 2012, p. 194). The model divides motivation into four factors: money, ideology, coercion, and entitlement (ego). Whilst money is rather straightforward, ideology captures actions of some (perceived) greater good, in line with the personal belief of fraud being justified in this very case. Examples of ideology include tax evasion and terrorist financing (Dorminey et al., 2010, p. 21). Coercion occurs when individuals become accidentally or unwillingly part of a fraud scheme and are often forced to assist the fraudulent actions. Ego or entitlement refers to the psychological theory of criminal behaviour rooted in the mental processes of individuals. Stotland (1977) provides examples of the self-centred factor, suggesting that a sense of superiority and admiration are essential for white-collar crimes.[42] The recent literature dealing with prominent cases like Enron, WorldCom, Adelphia, Phar-Mor and ZZZZ assumes that motivation behind recent cases is best explained via a combination of money and entitlement (Dorminey et al., 2012, p. 563). Overall, the MICE concept is rather simplistic but nevertheless provides an additional and potentially even complementary framework to the fraud triangle/diamond to explain fraudulent actions.

Vousinas (2019) has introduced the fraud pentagon model, with the acronym SCORE. The five letters stand for stimulus, capability, opportunity, rationalization, and ego. The five factors are derived from a combination of the fraud diamond and the MICE model. Vousinas (2019) further expanded upon his initial model by adding collusion as a sixth factor, resulting in the fraud hexagon, also called the SCCORE model. Collusion plays an important role, as already highlighted in the ABC theory, when groups inside the company act together,

[42] This can even be related to Sutherland (1940), whose white-collar crime theory was primarily based on the social status phenomena of a socioeconomic upper class.

making it even harder to detect the fraudulent violation. The fraud pentagon/hexagon seeks to provide a better and more complete approach by combining popular theory. However, as of now, there is no empirical justification for the model being better suited than its predecessors are.

2.2.6 Machiavellian Behaviour

The factors of opportunity and pressure from the fraud triangle have been researched extensively and are widely accepted as being associated with fraud, however, the third factor, rationalization, is only tangible at the level of individuals, rendering it rather difficult to observe (Murphy & Dacin, 2011, pp. 604–605). Rationalization is important in reducing the negative emotional effects triggered by fraudulent behaviour. These negative effects may include encountering self-conscious moral emotion, for example in the form of guilt, caused by the known violation of societal norms or by discomfort with performing a counter-attitudinal behaviour. Sloane (1944, p. 12) defines rationalization as the "mental process of justifying conduct by adducing false motives". Therefore, it is important for an effective rationalization to be believable to the person constructing it (Murphy & Dacin, 2011, p. 613).

The predisposition to misreport may be observed in the form of an attitude or a character trait (Murphy, 2012, p. 244). Whereas attitudes are less stable and may change over time, character traits are said to be largely consistent over a person's lifetime. Machiavellianism is regarded as a character trait that describes people engaging in manipulative behaviour towards others for their own purposes (Christie & Geis, 1970, pp. 1–10). Such individuals tend to be more opportunistic and act in self-interest, while being more inclined to fraudulent behaviour (cheating) when the probability of detection is low (e.g. Cooper & Peterson, 1980; Gunnthorsdottir, McCabe, & Smith, 2002; Murphy, 2012).

Murphy (2012) conducted an experiment in which participants were given the opportunity and the motivation to misreport. Based on the theory, Machiavellianism should be associated with a higher probability to misreport, as also supported by her findings.[43] In addition, Machiavellians were found to feel less guilty after the violation. Murphy further argues that misreporting is rational behaviour for a Machiavellian in the sense of an individual acting in his or her own best interest, like a *homo economicus* maximizing utility. The Machiavellian theory underlines the importance of character traits as a potential

[43] Murphy mentioned that her experimental design did not account for potential punishments for misreporting, which is a typical deterrence mechanism that may influence Machiavellians in their decision-making.

predictor for fraudulent behaviour (and there might be more traits involved, too) (Paulhus & Williams, 2002). Therefore, signals of character traits may be taken into consideration when designing fraud-detection mechanisms. However, in reality, this might be difficult and one is dependent on more accessible clues (Skousen et al., 2009). One source with the potential to be translated into character traits and thus constitute a potential predisposition towards fraudulent behaviour is the written word, which tends to convey deceptive clues (e.g. Goel et al., 2010).[44]

2.2.7 Fraud and the Principle Agent Theory

Management fraud can also be studied from the perspective of agency theory, especially when focusing on the conflict between the parties involved, the so-called principal-agent problem (Elsayed, 2017, pp. 4–6). The principal-agent problem arises from the very core of agency theory when one person or party (agent/s) is assigned to make decisions on behalf of another person or party (principal/s) in a contractual relationship (Jensen & Meckling, 1976; Fama & Jensen, 1983). This may occur in different settings, for example in the business context with the manager as an agent and the stockholders as principals but is also applicable to political science, where politicians and voters form the counterparties.

One of the causes rests in the potential existence of information asymmetries constituting qualitative or quantitative information advantages of one of the involved parties (Akerlof, 1970, pp. 489–491). What causes information asymmetry are differences in access to information, the expertise to understand it and physical (geographical) and social distances between the involved parties (McGuire, 1988, p. 7). For example, in the business context, the manager has considerably better access to company information, especially regarding the financial situation. Information asymmetry and opportunity to commit fraud are closely connected (Ndofor, Wesley, & Priem, 2012, p. 1780). Opportunity as part of the fraud triangle is a necessary and sufficient precondition of fraudulent behaviour and is especially pronounced in an environment with high information asymmetry.

Moreover, agency theory is based on several assumptions that may help to explain fraudulent actions. Besides being typically rational and risk-averse, people (agents) are also assumed to act in self-interest (Eisenhardt, 1988, p. 492). When acting in self-interest, the agent might not be inclined to report the true and fair view of the company. The existence of

[44] A detailed discussion is provided in the sections 3.1. and 3.2.

information asymmetry makes it impossible for principals to detect whether the reported numbers are truthfully reflecting the company situation. However, the consequences of the fraudulently reported numbers do not only harm the principals but all stakeholders with different levels of severity based on the dependence between the parties (Hill & Jones, 1992, pp. 134–136). In addition, information asymmetry may lead to moral hazards. These describe agents' tendency to engage in unwanted behaviour, which may lead to individual benefits, knowing that someone else is bearing most of the risk (Arrow, 1963, pp. 961–962). Moral hazards originated in the economics literature on welfare and insurance, where insured people might increase their risk exposure, knowing that they are secure and disassociated from potential costs (Baker, 1996, pp. 252–253). The difference in knowledge of the risk-taking party against the risk-bearing party about its own intentions and actions constitutes an information asymmetry.

One possibility to solve the aforementioned conflicts is to align agents' interests with those of the principals. This can, for example, be achieved by establishing appropriate compensation schemes that act as an incentive to pursue the desired outcome, or by internal and external corporate governance tools in the form of monitoring or disciplinary mechanisms (Wiseman, Rodríguez, & Gomez-Mejia, 2012, p. 206).[45] In addition, the market for corporate control or other internal and external corporate governance mechanisms to monitor and discipline agents may help overcome the agency problem (Dalton, Hitt, Certo, & Dalton, 2007, pp. 23–24). Typically, agents face a combined set of mechanisms that must be adjusted on a regular basis (Ndofor et al., 2012, p. 1794). For example, stock options could be a possibility to expand the risk sphere from the shareholders towards the managers. In addition to the direct interweaving of the agents' and the principals' fate, monitoring the execution of stock options by managers may yield important information. Indeed, Gerard and Weber (2014) identified patterns in the execution of stock options of managers shortly before they engaged in fraudulent schemes.

Although, stock options resulting in the agent potentially becoming a principal and tying their fate closely together, the recent accounting scandals are said to be negatively influenced by the excessive use of stock options as a form of management compensation. Hake (2016) argues that stock options are rather leading to a short-term than a long-term view and increase the likelihood of accounting manipulations in order to profit from capital gains

[45] A comprehensive overview of the relationship of corporate governance and fraud can be found in section 2.4.

rather than long-term profitability. Besides efforts to counter the unwanted behaviour of managers by establishing suitable mechanisms, principals can also decide to limit the leeway in which this might appear. Especially in cases where principals have little to no control over agents, such figures might reduce their vulnerability by limiting agents' discretion (McGuire, 1988, p. 8).

Overall, applying agency theory to financial statement fraud can help to build an understanding of the relationship between agents and principals and the mechanisms that may reduce the risk of fraudulently altered corporate reports. Moreover, it can assist in identifying potential clues for fraud detection purposes, which will be discussed in the following sections.

2.2.8 Implications for Fraud Detection

The fraud theories discussed in the previous sections have sought to explain fraudulent actions from different angles. Given the existence of multiple points of view, when trying to create effective fraud deterrence, prevention and detection mechanisms seem to be of vital importance in capturing as many factors as possible. This section will identify how fraud theory has been operationalized so far as well as the extent to which this study can rely on the theoretical foundations of the models presented to date. The availability of and access to the required information to capture the factors may constrain efforts in the first place. Moreover, some factors of fraud theory are rather difficult to capture, especially those meant to be surveyed on an individual level (Skousen et al., 2009, pp. 66–67; Nakashima, 2017, p. 27).

However, given that the type of fraud examined in this study influences a communication vehicle, namely the annual report in which the perpetrator is more or less directly communicating with the deceived receivers, communication patterns that are potentially derived from the individual level may be deducible (e.g. Buller & Burgoon, 1996). Hence, not every factor of each model can be implemented into the detection approach, or at least not at the same level of depth and most likely not with the same suitability. Presumably, a detection model building upon a holistic view and therefore covering many of the factors in different dimensions by enriching the data set with additional sources performs better than a mere focus on isolated factors (Dorminey et al., 2012).

Operationalizing variables from fraud theory is important, yet not always possible. In the case of quantitative predictors, this study relies on previous findings and adopts the association between measurements and fraud factors. Few studies have actively sought to build upon whole concepts of fraud theory like Skousen et al. (2009) did with the fraud triangle. In most cases, only single factors such as pressure are examined (and even then, the link to related theory is often missing).

Three basic factors, namely rationalization, opportunity, and motivation/pressure constitute the core of fraud theory. These factors, which made up the original fraud triangle, form the basis for further substantiations. Relative the well-studied and therefore established factors of fraud detection models of opportunity and motivation/pressure, rationalization has received little attention (Murphy, 2012). Rationalization can scarcely be observed from the outside. Even with theories that seek to expand upon rationalization (like Machiavellianism), the fraud diamond or the fraud scale, finding an appropriate way to operationalize this factor by exploiting publicly available information in general or from the company financials (as is most common) in particular represents a major challenge. In contrast to opportunity and motivation, rationalization is mainly rooted in the mindset of each person and is thus manifested in diverse forms. The closest we can potentially get with the data from annual reports is by analyzing the narratives, especially those subject to a high degree of management discretion (Goel et al., 2010, pp. 28–29). These may capture patterns that can be associated with character traits and attitudes, like deceptive nature, that scratch the surface of this fraud factor. Lendez and Lorevec (1999, pp. 48–49) suggest that management characteristics influence the financial reporting procedure of a company, potentially enabling the possibility of identifying fraudulent clues through reporting peculiarities in financial reports.

The factors of opportunity and motivation or pressure can be captured more easily, as they are typically closely tied to the company, hence necessary data are more readily available. In particular, financial pressure is observable through financial information available in the balance sheet or income statement. Opportunity as a manifestation of institutional peculiarities, like the design of the internal control mechanisms, is to a certain degree also readily observable through publicly available information. However, when expanding both factors by adding more explanatory theories, like MICE for identifying motivational elements or ABC theory for group cohesion and interpersonal relationships, the operationalization of these factors becomes significantly more difficult. Elliott and Willingham (1980, pp. 35–39) already suggested to rely on the work of Sutherland and Cressey and operationalize the hints from sociology like differential association theory to assess the likelihood of learning and spreading of unwanted behaviour. In general, as soon as the personal (and interpersonal in the case of ABC theory) sphere of managers is reached, determining appropriate predictors from publicly available information is hardly possible. Therefore, to come as close to a holistic approach as possible, relying on corporate narratives may help proxy for the factors that cannot otherwise be captured.

Despite attempting to translate the fraud factors into abstract variables, fraud theory literature has also created far more directly observable clues (Albrecht et al., 2016, p. 175). The clues are intended to sensitize auditors in particular to hints that may be associated with fraudulent activities. Generating an extensive list of such hints, also referred to as red flags, has been undertaken extensively in the fraud literature. Red flags are the fingerprints of fraud. They are derived from fraud factors or gathered on a case basis from empirical investigations (Singleton & Singleton, 2010, pp. 95–110). The literature has conducted extensive research on fraudulent cases over a considerable timeframe. The compiled results have led to the identification of red flags, which can be associated with factors of fraud theory. Albrecht et al. (1984) have composed a list of 82 potential red flags from an extensive range of different data sources. The red flags were not exclusive to financial statement fraud but fraud in general. Their hints could be classified into three levels: societal red flags, organizational red flags and personal red flags. Examples for the red flags include peer pressure to succeed on a social level, unrealistic performance goals on an organizational level and a feeling of being underpaid for the job on a personal level. Singleton and Singleton (2010) built upon the initial work of Albrecht et al. (1984), gathering a comprehensive list of fraud-specific red flags on the level of different fraudulent schemes. An example of a red flag for fictitious revenue schemes would be missing, incomplete, or fictitious customer data. For a scheme surrounding concealed liabilities, different auditors for different subsidiaries could constitute a red flag. Red flags have also found a way into professional standards. For example, SAS 99 – Consideration of Fraud in a Financial Statement Audit incorporates a red flag matrix for different types of fraud and fraud factors. Overall, the tenets of SOX have led to the consideration of red flags in standards in different styles (Singleton & Singleton, 2010, p. 301). Based upon fraud theory, it can be hypothesized that detection models covering more factors or the same factors in greater detail yield better detection performance compared to models with less coverage (Elliott & Willingham, 1980, p. 35). Therefore, this study relies on the identification and extraction of fraudulent patterns from all parts of the annual report, specifically a combination of quantitative and qualitative data, to capture potential clues.

Given that annual or interim reports are concerned with fraudulent manipulations, the financial reporting process – with its underlying framework as well as its typical schemes – needs to be discussed before identifying the potential detection mechanisms. Therefore, in the following sections, the foundation of an accounting and auditing standpoint will be laid, before explaining the feature generation and detection approaches.

2.3 Fraud through Manipulations of Accounts

The accounting standards according to which the reports are prepared and to which they declare conformity are willingly violated in the case of financial statement fraud. The difference in regards to an honest mistake rests in the intention of the preparers of the statements to manipulate the reports to deceive their readers (Zack, 2013, pp. 239–244). When detecting fraud, it is difficult to prove the intention behind altered financial statements (Singleton & Singleton, 2010, p. 40). The effectuations in this section will cover the most common practices of fraudulent alterations of financial statements that due to the effort and complexity that go into the execution of the schemes, very likely are not only an honest mistake (Wells, 2017, pp. 328–329).

Based on the fraud tree in Figure 2, the schemes of financial statement fraud can be divided into six categories: net worth/income overstatements and understatements; timing differences; understated and overstated revenues; understated and overstated liabilities/expenses; improper asset valuations; and improper disclosure. Each of these categories will be covered following the structure of Zack (2013) and dividing fraud types into revenue bases schemes, asset-based schemes, liability- and expense-based schemes and derived schemes, thus accounting for the remaining types of fraudulent misrepresentations. However, as will be shown in the following sections, fraudulent cases are not exclusive to a certain category that can be explained by the interrelated nature of the financial statements and the surrounding disclosure. Fraudulent schemes typically affect balance sheets and income statements at the same time (Wells, 2017, p. 49). Furthermore, in practice, multiple fraud schemes are often conducted in conjunction, making it difficult to separate the cases. In anticipation of the upcoming development of a fraud detection model, Figure 6 depicts the distribution of the different fraud categories in this study’s sample.

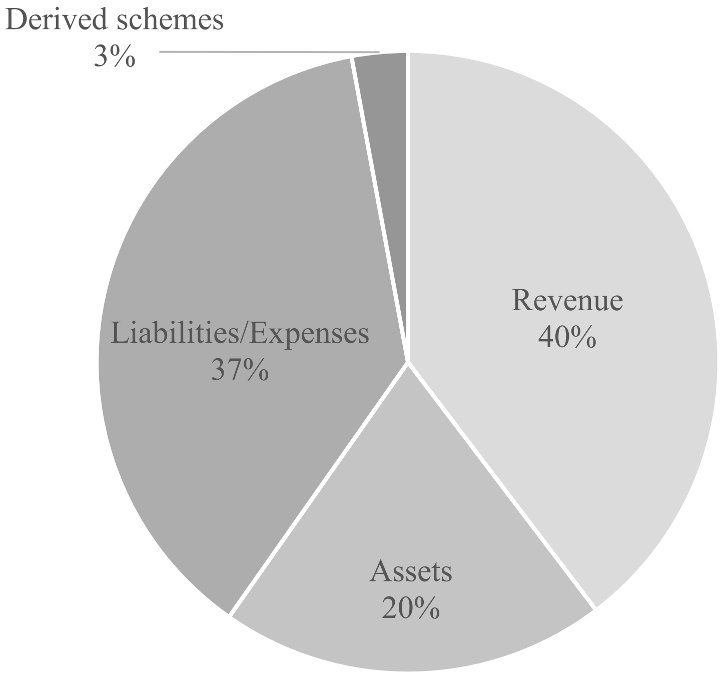


Figure 6 – Fraud schemes in the sample

The cases from the sample can mostly be assigned to revenue and liability/expense schemes, which together equate to almost 80% of all cases. Asset-based schemes seem to be rather rare, with only 20% of occurrence, while derived schemes are almost non-existent in this study's sample. Dividing the cases into single categories is rather arbitrary, as about one third of the cases refer to several AAERs with multiple misstatement events and can therefore hardly be allocated. The distribution of schemes of financial statement fraud in this study's sample is similar to the distribution of the Global Fraud Survey (2015), as published in Wells (2017, p. 321), in which expenses and liabilities schemes constitute the largest portion of cases with fictitious revenue schemes following slightly behind on the second place.[46] The following sections will provide a basic insight into the financial statement reporting process and the most common fraudulent schemes.

[46] The differences are most likely attributable to the scope of the study. The Global Fraud Survey, as the name suggests, incorporates a larger number of cases from different countries.

2.3.1 Financial Reporting and Annual Reports

It has been shown that the financial reporting process requires a certain degree of oversight to guarantee the correctness of financial statements (Wells, 2017, pp. 309–311). The developments of the accounting and auditing professions, as partially depicted in section 2.1.3, are the result of the need for reliable financial information issued by companies. The complexity of the financial reporting process has grown significantly over time and reached new heights in the aftermath of the severe fraud cases of the early 2000s and SOX (Rezaee & Riley, 2009, pp. 20–21). The participants involved and depicted in Figure 7 are the companies and their directors and officers, independent public accountants and oversight bodies, as broken down by the Treadway Commission in its report.

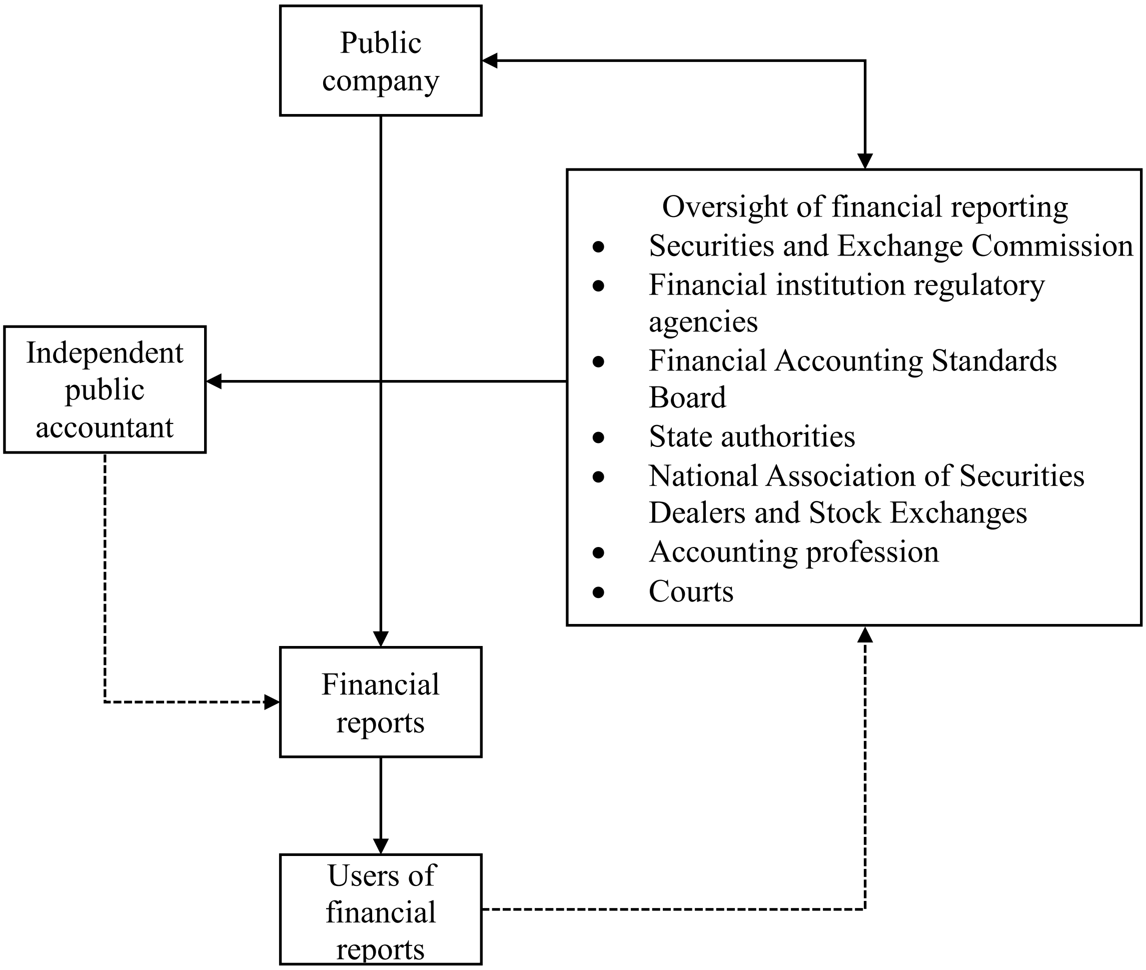

Public company
Oversight of financial reporting
• Securities and Exchange Commission
• Financial institution regulatory agencies
• Financial Accounting Standards Board
• State authorities
• National Association of Securities Dealers and Stock Exchanges
• Accounting profession
• Courts
Independent public accountant
Financial reports
Users of financial reports

Figure 7 – Financial reporting system[47]

[47] In reference to Rezaee and Riley (2009, p. 32) and the Report of the National Commission on Fraudulent Financial Reporting (1987, p.18). Retrieved from https://www.coso.org/Documents/NCFFR.pdf. Altered to reflect the users' influence on the oversight bodies.

The directors are responsible for true and fair presentation and thus conformity with applicable accounting standards. Therefore, the integrity and quality of financial reports reflect the commitment of the management to issue reliable, relevant and useful information about the company's financial status (Rezaee & Riley, 2009, pp. 184–185). In this regard, management prepares and certifies financial statements, while the board of directors, including the audit committee, oversees the process. Independent public accountants engage in the process by rendering an opinion about the fair presentation of the financial reports and their conformity with applicable accounting principles. The purpose of lending more credibility and objectivity to the financial reports and therefore reducing information risk can only be achieved if the auditors are without any doubt independent and knowledgeable to fulfil the task (Wells, 2017, pp. 310–311). Despite those two directly involved parties (management and public accountants), several oversight bodies are concerned with the financial reporting process (Rezaee & Riley, 2009, p. 33). The oversight bodies influence the whole process, for example by releasing accounting standards and requirements as well as auditing standards. Moreover, oversight institutions like the SEC are responsible for capital market supervision, including the financial reporting process of publicly traded companies. However, the SEC may delegate its authority to other institutions, for example, the Financial Accounting Standards Board (FASB) regarding the issuance of accounting standards or the PCAOB to regulate auditing firms (in the post-SOX era) (Wells, 2017, pp. 309–311). Finally, the recipients of the financial reports absorb the information and make decisions based upon the true and fairly presented view of the company. Potential feedback from the users to the oversight bodies exists, due to the possible influence for instance of the accounting standard-setting process according to the needs of the financial statement's recipients (Harding & Mckinnon, 1997).

The primary financial report and therefore the main communication vehicle is a company's annual report. In the USA, under Rule 14a-3 of the Securities and Exchange Act of 1934, a company is required to provide annual reports to shareholders in conjunction with annual shareholder meetings. The annual report must consist of audited financial statements, the balance sheet of the two most recent fiscal years and statements of income and cash flow of the three most recent years (Securities and Exchange Act 1933, Chapter 240.14a-3 (b) 1). In addition, Regulation S-K and especially subpart 229.300 regarding further financial information organized in items 301-308 should be included in the annual report. Such information must include selected quarterly financial data, summaries of financial data for the last five years, segment information, management discussion and analysis of financial condition and results of operations, quantitative and qualitative disclosure on market risks, market price of company's stock over the last two years for each quarter, description of business activities and changes in disagreements with accountants on accounting and financial disclosure. The Sarbanes-Oxley Act of 2002 requires additional information in conjunction with the annual report. SOX Section 301 and 401 and the related SEC implementations and rules require an annual audit committee report in conjunction with the annual shareholder meeting in the publication of the proxy statement, stating the mandatory activities of the audit committee (Wells, 2017, pp. 309–313). The disclosed information covers a review of the audited financial statements from the audit committee with the management, a judgement on the auditors' independence and a recommendation that the financial statements be included in the annual report on Form 10-K (or 10-KSB for small businesses).

Figure 7 is intentionally limited to the financial reporting process for public companies, as private companies in the USA are not required *per se* to issue annual reports or to release financial information to the public (Reardon, 2017). Due to special requirements or when negotiated with investors, private companies may have to comply with the same set of rules as public companies. In this regard, SEC enforcement actions apply both to public and private companies alike. In the press release 50-2019, the SEC charged the Chief Executive Officer (CEO) of a start-up private company (called Jumio) for defrauding its investors.[48] However, such cases are rather rare and due to the limited availability of financial information, private companies are seldom included in studies dealing with fraud. An unusual example has been provided by Fleming, Hermanson, Kranacher, and Riley (2016), who examined differences in financial statement fraud between private and public companies. Their findings suggest that due to the stronger anti-fraud environments in public companies, fraudulent schemes are less obvious than their counterparts in private companies are. Moreover, cases in public companies appear to be larger, have a higher number of perpetrators and are less likely to be discovered by accident, requiring more sophisticated detection mechanisms.[49]

The importance of truthful financial reports is vital for capital market efficiency. Therefore, the requirements for the financial reporting process and reports are closely

[48] Retrieved from https://www.sec.gov/news/press-release/2019-50.

[49] Forming colluding groups to circumvent controls and cover fraudulent actions has been covered in section 2.2.4. Depending on the size of the colluding group, internal controls can be rendered rather ineffective, requiring additional detection mechanisms to unveil the fraud.

supervised to ensure high-quality financial reports. However, financial statement fraud still occurs, even after enhancements to supervision and reporting requirements (Ogoun & Obara, 2013). To further examine the challenge of fraud detection, the following sections will discuss financial reporting standards and accounting and auditing fundamentals before presenting an overview of the most prevalent fraudulent schemes through the manipulation of accounts. Afterwards, detection sources will be highlighted, before summarizing the theoretical background and developing a comprehensive detection model from there.

2.3.2 Financial Statements and GAAP

When examining financial statement fraud, it is important to elaborate on the very object of manipulation and its underlying principles. In the definition of financial statement fraud and its substantiations, the generally accepted accounting principles are often specifically mentioned, e.g. AS 2401.06: “[m]isstatements arising from fraudulent financial reporting are intentional misstatements or omissions of amounts or disclosures in financial statements designed to deceive financial statement users where the effect causes the financial statements not to be presented, in all material respects, in conformity with generally accepted accounting principles”. The loss of confidence in audited financial statements after the major accounting scandals shocked the accounting profession to its very core. Reporting truthful and reliable information is the foundation of the profession and is stated in numerous core principles issued by standard-setting bodies. The importance of these principles, especially in conjunction with fraudulent manipulations of financials statements and their schemes will be highlighted in this and the following sections.

The generally accepted accounting principles (GAAP) are standards regulating the preparation and presentation of financial reports, including financial statements (e.g. Wells, 2017, p. 303). The standards are developed and overseen by private accounting institutions and governmental regulators (e.g. Sunder, 1988, pp. 37–39). Besides nation-specific standards like the US GAAP or Canadian GAAP, there exist international accounting standards like the International Financial Reporting Standards (IFRS) that may be endorsed by countries to complement or replace existing nation-specific standards. However, with attention to global relevance, competition among accounting standards has led to the development of a duopoly, formed by the FASB and the International Accounting Standards Board (IASB) (Meeks & Swann, 2009). Thus, the leading accounting standards today are the IFRS and US-GAAP (Leuz, 2003, p. 446). The FASB is a private, non-profit organization designated by the SEC as the accounting standard setter for public companies. Companies with an interest in being publicly traded on US stock exchanges or in issuing securities for US markets have to comply with US GAAP. The IASB is the standard-setting body of the International Financial Reporting Standards Foundation, developing accounting standards to be endorsed around the world. As of now, the IFRS are required by over 140 jurisdictions.

Accounting standards not only differ in their elaboration of rules but also in their underlying core principles. These core principles are the foundation of financial reporting, guiding the standard-setting bodies in developing additional or revising existing standards and serve as a guideline for users of the standards.[50] The FASB has issued the Statement of Financial Accounting Concepts No. 2 – Qualitative Characteristics of Accounting Information in 1980, stating the requirements of high-quality financial reports.[51] The FASB presents a hierarchy of qualities that make accounting information a desirable commodity (Wells, 2017, p. 303). According to FASB, relevance and faithful representation are the fundamental qualitative characteristics that make financial information useful (Statement of Financial Accounting Concepts 8, QC17). Moreover, the enhancing qualitative characteristics comparability, verifiability, timeliness, and understandability enhance the usefulness of information that is relevant and faithfully represented (QC19).

Financial information is relevant if it makes a difference in the decision-making process. It should help recipients to assess past performance, predict future performance, confirm or correct expectations and provide feedback in earlier expectations (QC6-QC10). An entity-specific aspect of relevance is materiality (QC11). Materiality of information could result in the change of a decision if the underlying information is omitted or misstated. Financial information is faithfully represented if recipients deem the information to represent the conditions that it purports. Therefore, information is faithfully represented if it is complete, neutral and free from error (QC12). Information is complete if the depiction includes all that is necessary for a user to understand the phenomenon being depicted (QC13). The neutrality of financial information specifies choices leading to different outcomes, to not be biased towards those outcomes but taken in order to reflect the true situation in the best way possible

[50] For further information, see IFRS Conceptual Framework, available through https://www.ifrs.org/issued-standards/list-of-standards/conceptual-framework/ and for US-GAAP the Statements of Financial Accounting Concepts, available through https://www.fasb.org/resources/ccurl/816/894/aop_CON1.pdf.

[51] The Concept No. 2 was superdeeded by the Statement of Financial Accounting Concepts 8, issued in 2010. Chapter 3 of the new concept is titeld Qualitative Characteristics of useful Financial Information and has amended the original concept and will be discussed hereafter. Concept 8 is available on the official website www.fasb.org.

(QC14). Error-free information is concerned with the description of the information and the process responsible for providing the information, both of which have to be free from omissions or errors. (QC15).

The enhancing qualitative characteristics comparability, verifiability, timeliness, and understandability may help to improve the usefulness of information and furthermore can determine how a phenomenon is depicted if alternatives are equally satisfying the fundamental qualitative characteristics (QC19). Comparability is achieved if information of one entity can be compared to similar information of another entity and with similar information of the same entity but from previous periods (QC20). Identifying and understanding similarities and differences among items is a core purpose of comparability (QC21). Comparability is thereby related to consistency, which refers to the consistent application of the same method for the same item of the same entity across periods or in the same period across entities (QC22). The verifiability of information is given if independent individuals using the same measurements arrive at similar conclusions or outcomes (QC26). Timeliness is concerned with the availability of information before it loses its capacity and capability of influencing the decision-making process (QC29). However, timeliness cannot make information relevant, even though a lack of timeliness can render information irrelevant. Finally, understandability of information is given if it is clearly and concisely presented (QC30). This does not imply that information has to be understood by everybody, which would potentially lead to the omission of complex topics, leading to incomplete and potentially misleading financial information (QC31). Besides the fundamental and enhancing qualitative characteristics, the FASB is concerned with the cost constraint on useful financial reporting (QC35-QC39). The cost constraint considers the trade-off between perceived benefits and the associated perceived costs of a particular disclosure. The benefits should exceed the costs in order to justify the adaption. However, quantification, especially of the benefits of information, is rather difficult due to the relative nature of the value attached to it through different individuals. Furthermore, practical implementation must be reasonable and appropriate for companies.

Although effectuations are provided with regard to the FASB’s core concept, similarities to other standard setters are high, especially regarding the two primary characteristics of relevance and faithful representation.[52] Nevertheless, qualitative characteristics may differ slightly and shift the emphasis to different qualities, like understandability or transparency. The general problem with these sets of characteristics is that they are hardly simultaneously achievable without a certain trade-off (Smith, 1996). Smith (1996) has surveyed different user groups of financial statements, finding that reliability and relevance are the most important characteristics across groups, even when sacrificing the understandability, timeliness, completeness and comparability of the reports. The results suggest that truthfulness as an aspect of reliability is valued highly and that fraud prevention and detection are important in guaranteeing the credibility of financial reports.[53]

Another problem with GAAP and financial statement fraud is attributable to differences in accounting information resulting from the application of different country-specific versions. A prominent and often discussed case surrounded Daimler Benz and the adoption of US GAAP in 1993. The net income reported for 1993 under German GAAP accumulated to 0.6 billion DM while under US GAAP Daimler Benz reported a loss of 1.8 billion DM (historical exchange rate was 1.65 DM/USD for 1993).[54] This rather extreme example emphasizes the potential differences in the outcome regardless of the fact that the company did comply correctly with US GAAP and German GAAP (Shil, Das, & Pramanik, 2009, p. 196).

To account for differences across accounting standard-setting bodies, the following section will highlight the accounting standard convergence process in the light of financial statement fraud, before explaining accounting-based fraudulent schemes.

2.3.3 Accounting Standardization, Convergence and Fraud

The requirements set by accounting standards provide the framework when distinguishing between correct and misstated financial statements. Tweedie and Seidenstein (2005) claim that the uncertainty that results from low-quality financial reporting and corporate governance is severely punished by investors, as evidenced in the aftermath of the accounting scandals of the 2000s. Therefore, ensuring high-quality financial reporting

[52] For IFRS, the Conceptual Framework (issued in September 2010 and revised in Match 2018) Chapter 2: Qualitative Characteristics of useful Financial Information suggests similar (albeit not identical) requirements. In the Appendix to Chapter 3 of the Statement of Fonancial Accounting Concepts No 8., the FASB mentions the joint efforts with the IASB and links to references to the IASB literature (BC3.2)

[53] The examination was concerned with characteristics from different standard setters, including the FASB and the ASB.

[54] Annual reports retrieved from https://www.daimler.com/documents/investors/berichte/geschaeftsberichte.

through accounting standards represents the basis for efficient resource allocation and economic growth. Accounting standards vary over jurisdictions and must constantly be revised to keep pace with the constant development of the business environment. This has resulted in a multitude of standards available to companies, although through countrywide corporate law restrictions, the effective number of standards considered for use in crafting financial statements is limited (Shil et al., 2009, pp. 197–198). The trend towards internationalization and professionalization can be related to a wide range of factors, including the globalization of capital markets as well as demographic changes and the increased complexity of business transactions (Volmer, Werner, & Zimmermann, 2007, pp. 457–459). According to Gordon and Bovenberg (1996), the most likely explanation for the international immobility of capital flows lies in the informational advantage that domestic investors enjoy over foreign investors. The effort to understand multiple accounting standards in order to confidently invest in a foreign country has a negative impact on transaction costs. Having a single set of accounting standards or converged standards has the potential to decrease information asymmetry and transaction costs and thus strengthen international capital flows (Chen, Ding, & Xu, 2014). The positive impact of the adoption of international accounting standards like IFRS on information barriers has been extensively proved at a theoretical and empirical level. Easley and O'Hara (2004) have used a rational equilibrium model to demonstrate that companies can influence their capital costs by adopting different accounting standards, as the quality (precision) and quantity of information vary across standards. Tarca (2004) has studied the voluntary adoption of international standards like IFRS or US GAAP in the United Kingdom (UK), France, Germany, Japan and Australia, finding that especially large and internationally operating companies engage in the application of international standards, emphasizing their usefulness in accounting convergence due to the greater comparability of financial reports. Lambert, Leuz, and Verrecchia (2007) have noted that higher accounting information quality can result in lower risk premiums within a model, consistent with the CAPM. In general, empirical evidence on IFRS adoption reveals rather positive effects (Chen et al., 2014). Nevertheless, some findings report higher costs of capital for certain countries and periods, such as Daske (2006) for a sample of German firms and the IFRS transition phase between 1993 and 2002.

In this study, a sample is drawn from companies that are required to file in accordance with US-GAAP between 1996 and 2010. During this period, the convergence of accounting standards accelerated immensely in line with the first joint efforts of FASB and IASB to release common international accounting standards (e.g. Zack, 2009, pp. 17–18). Regarding financial statement fraud research, the convergence project and the establishment of international accounting standards have facilitated a broader view of the topic and enabled studying the subject in greater detail. The following paragraph will highlight important milestones in the convergence process, before discussing their influence on financial statement fraud.

During the International Congress of Accountants from 1904 in St. Louis, USA, the first ideas for international accounting standards were presented (Shil et al., 2009, p. 196).[55] The call for internationally agreed accounting standards intensified in the 1960s, related to an increase in international investments. Indeed, differences in country-specific regulations had resulted in a lack of comparability of financial statements. In the aftermath of the 8th International Congress of Accountants in 1962, the expressed need for internationally accepted standards led the AICPA to establish the Committee on International Relations to start working on a potential harmonization process. The International Accounting Standard Committee (IASC) started to develop accounting standards of international relevance in 1973, successfully releasing its 25th standard by 1987. The steady work of the IASC in releasing new and improving on existing standards in the 1990s contributed to their adoption by companies, accounting professionals and other standard-setting institutions. However, until that point (mainly during the 1970s and 1980s) the IASC and other major Anglo-American standard setters achieved only minor compatibilities in the existing standards (Street & Shaughnessy, 1998b, p. 203). Harmonization attempts were intensified by the FASB after 1991 when its first strategic plan for international activities was released (Herz & Petrone, 2004, p. 634). The strategic plan involved joint projects with other accounting standard setters and active participation in the IASC's accounting standard-setting process, among others. In 1994, the first joint project concerning the revision of the earnings per share standard between the IASC and the FASB marked a milestone in the harmonization efforts. Besides the collaboration between the FASB and the IASC, the formation of the G4 in 1993 set the starting point for joint projects between the FASB and its counterparts from the UK, Canada and Australia (Street & Shaughnessy, 1998a, p. 132).[56] The IASC joined the group as an observer member, strengthening its relationship with national standard-setting bodies.

[55] It is now known as the World Congress of Accountants and the International Congress of Accountants from 1904 St. Louis (USA), or sometimes referred to as the 1st World Congress of Accountants.

[56] The "G4" later formed the "G4+1" when New Zealand joined the group. It abandoned its work in 2001 after the IASB was founded.

Between 1999 and 2001, the IASC restructured itself as the International Accounting Standard Board (IASB).[57] The 2000s saw improved collaboration between the major standard setters and increased recognition for international accounting standards. Two important milestones of the early 2000s were met, when in 2002 the European Union (EU) decided to adopt IFRS, according to which companies would prepare their consolidated financial statements from 2005. In the same year, the Norwalk Agreement between the FASB and the IASB was settled. This was seen as one of the most important steps in the convergence process, declaring four goals, namely the joint development of standards, the short-term elimination of differences, the elimination of differences remaining after the initial convergence phase until 2005 and the vision to stay converged (Tweedie & Seidenstein, 2005, pp. 597–601). The Memorandum of Understanding (MoU) from 2006 elaborated on the Norwalk Agreement by adding precise topics to the list of convergence projects on which both standard setters had to work. In the opinion of the boards involved, remaining differences should be reduced by new common standards, although projects outside of the agreements did exist.[58] The next big driver in the convergence process was the SEC's affirmation of the IFRS for foreign issuers. However, at that time (around 2011), US-GAAP was regarded as the higher quality standard by the SEC, leading to increased effort by the IASB to conform its standards to GAAP (Baudot, 2014, p. 981). Greater pressure for improvements in accounting standards and the convergence process also came from the governmental actors of the G20 in the aftermath of the financial crisis (Baudot, 2014, pp. 974–976). In 2009, these demanded noticeable convergence progress until 2011, which was not met by the standard setters. During the period of intensified efforts, less and less common ground could be found as problems started to emerge related to underlying principles, leading to a stagnation of the convergence process (Baudot, 2014, pp. 982–984). It should be remembered that despite the collaboration, the two standards operated in a competitive relationship. In 2013, the Accounting Standards Advisory Forum (ASAF) was founded by the IFRS Foundation to enhance the cooperation of worldwide accounting standard setters. With the FASB as one of the members of the ASAF, the exchange between the major standard setters is now also possible through the forum that was created.[59]

[57] An official chronological overview is available at https://www.ifrs.org/about-us/who-we-are/#history.

[58] The MoU was updated in 2008. Both versions are available on the FASB website at http://www.fasb.org.

[59] The ASAF holds 12 seats, one for Africa, three for the Americas, three for Europe, three for Asia and two appointed members, subject to maintaining the geographical balance.

The influence of different accounting standards and convergence on financial statement fraud has yet to be extensively studied. Nevertheless, some vital contributions have been made to the fraud literature. Chen, Hu, Lin, and Xiao (2015) have studied fraudulent cases of dual-listed companies, identifying GAAP differences between the financial reports of companies filing for the same periods in different countries that might be used to detect financial statement fraud. They divided the differences into an expected or explainable and an abnormal gap. Their findings suggest that larger abnormal gaps are associated with fraudulent manipulations of financial statements. Meeks and Swann (2009) have exposed the costs and benefits of standardization and their contributions to wealth creation. They argue that monopoly standards have a negative impact on market discipline, which may lead to less transparent and more manipulated financial reports, among other disadvantages. McAfee and Guth (2014) discussed similarities and differences between US GAAP and IFRS and discussed opportunities of fraudulent manipulations in a conversion process from the rather rule-based US GAAP to the principle-based IFRS. They further assume that the highly prescriptive nature of US GAAP hampers the accounting and auditing profession in preventing financial statement fraud, as especially apparent during the 2000s accounting scandals.

In the following four sections, the most prevalent fraudulent reporting schemes will be highlighted. Thereby, the focus will be on US GAAP, as it represents the applicable accounting standard for the companies in this study, although as stated in the aforementioned remarks, through convergence and the international relevance of US GAAP, this study is not limited to the US market.

2.3.4 Revenue and Sales Schemes

Many cases of financial statement fraud are associated with the manipulation of revenue. Revenue and its recognition have been an important topic for accounting standard setters ever since (Wells, 2017, p. 332). In defining general underpinnings or specifying the details of particular transactions, standard setters are in a constant struggle to keep up with the development of businesses. This is not only the case for revenue recognition but also all standards in general. For example, as a major project of the US-GAAP/IFRS convergence, the IASB released IFRS 15 in 2014, combining the former IAS 11 (construction contracts) and IAS 18 (revenue). The FASB issued the new ASU 2014-09 (codified in ASC 606) in the same year, completing the joint effort to improve revenue recognition. This study relies on a sample of cases covering 15 years, which is why the following sections will not go into specific details regarding accounting standards and will instead provide a broader scope.

The ACFE's fraud tree, which has been used as an introduction to fraud in general and financial statement fraud (see section 2.1.1 and Figure 2), will again be utilized to structure the most prevalent types of fraud schemes. The fraud tree distinguishes between net worth/net income overstatements and understatements. The major objects of manipulation are revenues through overstating, assets through improper valuation and expenses and liabilities through understating, which can also be taken from Figure 6, presenting the fraction of each of the schemes in this study's sample (Wells, 2017, p. 321). Each of the following sections will deal with one of the objects, starting with revenue.

Zack (2013, p. 6) classifies revenue manipulation schemes into four categories: timing schemes, fictitious or inflated revenue, misclassification schemes and gross-up schemes. These will be followed in the upcoming effectuation, as they answer the when, why, where and how of revenue recognition, thereby providing a comprehensive overview.[60] As previously mentioned, most schemes in practice are not exclusive to one category and instead extend across multiple categories through the interrelated nature of financial statements (Wells, 2017, p. 332).

Timing, the when of fraud schemes, deals with shifting revenue between periods outside of the legal possibilities that accounting regulations offer (Wells, 2017, pp. 335–340). Most commonly, revenue is recognized too early, boosting the current period's performance and leading to a problematic lack of revenue in the period that should actually be under study. The practice often results in a downward spiral, when additional manipulations are necessary for later periods to cover up for the revenue that has been recognized too early. This short-sided manipulation, often termed "management myopia", is induced by the expectations and goals that one expects to meet (Merchant, 1990, pp. 297–299). Timing schemes can be established in different ways, the most straightforward being the alteration of records. Transaction documents are dated backwards to ensure the possibility of recognizing the revenue in the required period. This alteration can be done with or without the knowledge of the transaction party. Construction contracts offer another possibility to shift revenues. When revenue recognition is based on the percentage of completion method, as commonly used under US-GAAP and IFRS, the amount of total revenue recognized in each period depends

[60] Zack (2013) offers an in-depth review of common financial statement fraud cases. Some examples to clarify the execution of different schemes based on his review will be brought up in the upcoming sections.

on the amount of costs accrued in the period in relation to the total estimated costs (Zack, 2013, pp. 12–15). Especially the accrued costs in any period may be subject to manipulations when overstating results in high revenue recognition for the respective period. The double-booking or misclassification of costs related to other projects is often manipulated in the percentage of completion context (Sidorsky, 2006, p. 12). Boosting sales by channel stuffing is another possible scheme based on timing irregularities to artificially pretend that one is running a successful business. Channel stuffing is a scheme in which sales are generated by pushing excess inventory along the distribution line, for example to retailers, at the end of the period or quarter (Jackson, 2015, p. 83). Knowing that a substantial fraction will most likely be returned in the following period, the scheme empties one's inventory while generating fictitious sales (Singleton & Singleton, 2010, pp. 80–81).

Fictitious and inflated revenues provide two possibilities to boost firm performance (Zack, 2013, pp. 33–42). The former refers to fabricated transactions that have not happened; the latter to actual transactions that have been artificially inflated in scale (Singleton & Singleton, 2010, p. 81). Compared to timing schemes, where the date of recognition rather than the amount of revenue from transactions is altered, fictitious and inflated revenue schemes affect the underlying value of the transaction. In practice, transaction partners may exist, yet false sales are recorded or the transaction partners and their respective sales are made up. Completely fabricated transactions, especially when fictitious customers are involved, are usually more easily detectable by irregularities in customer master files. These types of transactions may be more difficult to recognize, when regular customers, which are covering the perpetrator, are involved. Another possibility to fabricate fictitious revenue is the top-side adjustment in which entries are recorded in the financial statements but fail to be found in formal accounting records like the general ledger (Jackson, 2015, p. 127).

Misclassification schemes deal with intentionally wrongfully classified transactions. Misclassification can have a material impact on financial statements when incorrectly classified transactions misstate positions. This would not have an impact on the bottom-line outcome. However, key performance indicators referring to manipulated lines may present misleading information and influence the economic decisions of the financial statement's recipients. One-time income transactions or non-recurring costs that occur outside of regular business operations and are unlikely to reoccur in the future are shifted to positions representing core business activities (Zack, 2013, pp. 47–50). Another possibility is the false recording of financing agreements as revenue. Thus, a company sells to another entity (e.g. another company or a bank) with the intention of repurchasing the goods at a later date at a premium price. This transaction results in additional funds for the selling company, for which it pays interest in the form of the repurchase premium (Zack, 2013, p. 48). However, under ASC 470-40-25, the company selling a product to another company and in the related transaction agrees that the repurchase of the sold product (or a substantially identical one) must record a liability to the extent that the repurchase and the financing arrangements are covered. The company is not allowed to record the transactions as sales and must remove the goods from the balance sheet.

In gross-up schemes, revenue and associated costs are inflated simultaneously, making the company appear larger, without the primary goal of influencing profitability (Zack, 2013, p. 57). In addition to generating artificial transactions, round-trip transactions are another example of gross-up schemes. Round-trip transactions boost the revenue of two or more companies, without creating any economic benefit, by arranging a series of transactions, starting and ending at the same point (Wells, 2017, p. 333). Company A may sell unused assets to Company B, with the arrangement of Company B selling the same assets back at a similar price in the future. Such deals are also called "lazy Susans", deriving from the rotatable services that used to distribute food among people. Distinguishing between legitimate transactions and illegitimate "lazy Susans" requires an in-depth investigation of transactions to identify sequences associable with a gross-up scheme via round-trip transactions (Zack, 2013, pp. 59–61).

2.3.5 Asset Schemes

Following the structure of Zack (2013), asset-based schemes are divided into four categories: improper capitalization of costs, inventory schemes, overvaluation in the context of fair value accounting and improper asset impairments.

The improper capitalization of costs leads to an overstated net income, as expenses are illegitimately capitalized as assets, taking them from the income statement to the balance sheet (Wells, 2017, pp. 343–344). For capitalized assets, expenses are recognized as depreciation or amortization over future periods. The schemes are commonly related to assets that are developed internally by the company, assets failing to provide future benefits, the capitalization of research and development costs or administrative costs, the capitalization of start-up costs, and the illegitimate capitalization of advertising costs. Ryerson (2009) has examined AAERs dealing with improper capitalization for the years 2001 to 2008 and suggests that distinguishing misstatements between accounting errors and fraud is difficult because the degree of uncertainty in the requirements is rather high. For fraudulent cases, the intention could mostly be referred to as the necessity of meeting internal or external expectations, leading to earnings manipulation through improper capitalization.

Inventory schemes revolve around the concept of misreporting the quantity or value of inventories in order to misstate (usually to overstate) their true condition, which is usually reinforced via the manipulation of records (Wells, 2017, pp. 347–349). Inventory-related schemes are common due to the complex and challenging nature of inventory valuation and the difficulties involved in the auditing of inventories (Kirkos, Spathis, & Manolopoulos, 2007, p. 998). There are plenty of possibilities available to manipulate inventories, the most common regarding quantity being counting items multiple times or reporting fictitious inventory not owned by the company (Wells, 2017, p. 348). Misstating the value of inventories can be achieved by allocating inventory to higher manufacturing overhead costs or failing to recognize an impairment loss. Inventory schemes are often used for earnings manipulations, affecting the balance sheet and income statement alike (Rezaee & Riley, 2009, p. 98).

Fraudulent schemes based on the manipulation of fair values to illegitimately benefit from wrongfully achieved asset valuations also fall into asset-based schemes. Determining fair values involves a high degree of estimations and judgements, rendering the process susceptible to fraudulent alterations (Zack, 2009, pp. 9–11). In most accounting standards, fair values can be achieved using a number of valuation approaches in a hierarchical order.[61] The highest priority class consists of valuations based on market prices in active markets or identical assets (level 1). Second-priority valuations are based on market observables, due to a common shortage of level 1-quality data (level 2).[62] The last priority contains valuations based on un- or scarcely observable inputs, where the company must rely on a significant number of assumptions to ascertain a valuation result (level 3). It is possible to determine fair values either internally or through external parties, but due to the complexity of the subject, involved parties require proper expertise and knowledge. Under fraudulent intentions, the management might lean towards a preferred fair value, thus trying to influence the valuation process (Zack, 2009, pp. 211–212). Potential manipulations may occur through the appraisers, who might be bribed to deliver the desired outcome, or even be fictitious, the valuation and especially the valuation report being entirely fabricated. The fair value

[61] See IFRS 13 or SFAS 152 for specific details on fair value valuation.

[62] Lower level valuations are only applicable if the prerequisites for higher level valuations cannot be achieved.

measurement may also be affected through the available data basis. According to the hierarchical order of the valuation approaches, the potential for manipulations increases with the extent to which the measurement parameters become less observable, augmenting the degree of estimations and judgements on the valuation outcome. In most actual cases, companies exploit fair value manipulations to increase their book value. Benston (2006) has criticized the extensive use of the level 3 valuation of Enron and has assumed that by tying compensation schemes to targets evaluated through fair values, its managers were inclined to engage in fraudulent manipulations.

Furthermore, in the light of fair value accounting, impairments and especially impairment losses are susceptible to manipulations (Zack, 2009, pp. 99–102). The general idea behind impairments is the consideration of a permanent reduction in the value of an asset; for example, in the case of investments, this may occur when the carrying amount exceeds the price that it could be sold for, leading to a potential fair value manipulation. The difference between both values is written off, resulting in the decline of the value of the respective asset on the balance sheet. In practice, determining the impaired value differs in detail between accounting standards.[63] However, assessing the new value contains a high degree of estimation and judgement, in a similar way to the fair value valuation. Manipulations usually occur by limiting the resulting loss or the classification of the losses.[64]

2.3.6 Expense and Liability Schemes

Fraudulently boosting corporate performance or conveying the impression of a stronger financial situation can also be achieved by manipulating expenses and liabilities. Similarly to revenue, the alteration of expenses can lead to timing differences and understatements and the counterpart to the generation of artificial revenue or assets and the omission of expenses and liabilities (Zack, 2013, p. 131). Pushing expenses in later periods may induce the same downward spiral effect, which becomes apparent in the case of shifted revenue. The need to cover up the manipulation with additional future periods has the potential to lead to a longer-lasting scheme. Compared to revenue schemes that rely on the manipulation of sales, expense and liability schemes are said to be much easier to be committed, because falsifying

[63] See IAS 36 or ASC 360-10 for specific information on impairment.

[64] Under US-GAAP and IFRS, depending on the assets, impairments may be distinguished between temporary and non-temporary, resulting in different classifications of impairment losses. This change may affect financial ratios where losses are no longer part of another comprehensive impact and influence net income through profit and loss.

the required documents is a considerably higher effort for a comparable effect in the case of sales (Wells, 2017, p. 342).

Shifting expenses in later periods can usually be achieved by manipulating invoices and in the case of omission, hiding invoices and therefore not reporting the respective liability under the accounts payable. Similarly, related parties can be involved in scamming relevant documents to defer expenses and/or payments (Zack, 2013, p. 132). Depending on whether accrual or cash accounting is the underlying basis for recognition purposes, the correct moment of the recognition of expenses may diverge. However, under US-GAAP and IFRS only accrual accounting is legitimate, where the recognition is taking place when the expenses (or revenues) are accrued. The manipulation of expenses is often tied to revenue schemes, further stimulating a boost in performance in time and value (Singleton & Singleton, 2010, p. 81).

In the case of understated liabilities, the subject of the manipulations is usually derived from the type of liability and the constituent features that determine its initial and subsequent measurement (Zack, 2013, pp. 141–144). These are usually the respective interest rate, the time over which it is amortized or manipulations to the process of fair value measurement if the fair value option under US-GAAP and IFRS is chosen. Understating can then also be achieved by simply altering the documents or as part of the fair value fraud schemes. Another possibility, which is especially hard to detect and is usually undertaken by large and internationally operating companies, is the moving of liabilities to a subsidiary that is either not audited or audited by a different auditor (Wells, 2017, pp. 342–343). In this case, the liability is completely omitted from the company’s balance sheet.

2.3.7 Derived Schemes

In the final section of this chapter, relevant schemes derived from those discussed so far will be highlighted. These include the concealment of illegal acts as well as the often closely related disclosure fraud. As part of disclosure fraud, the section discusses illegitimately prepared consolidated statements or business combinations like mergers and acquisitions.

Asset misappropriations, as part of the ACFE fraud tree and explained in section 2.1.1, are not directly referable to financial statement fraud, as they do not *per se* lead to an illegitimate presentation or material alteration of said statements. However, to conceal the asset misappropriation, typically the manipulation of accounts is necessary, potentially leading to financial statement fraud in the process, which can also be recalled in Figure 3.

These actions are usually carried out by higher level employees or executives with access to the accounting tools (Wells, 2017, pp. 299–302). In the case of embezzlements, usually invoices are manipulated to appear business-related and are then paid by the company instead of the respective employee (Zack, 2013, pp. 172–173). In this regard, the concealment of illegitimate bribes by hiding them in other transactions or making up fake transactions to cover up the delinquency affects the financial statements. However, manipulations like those previously mentioned do not represent financial statement fraud, as they are not typically intended to manipulate in order to deceive the users of the financial statements but rather to conceal other schemes of occupational fraud, thereby altering transactions in the process. Moreover, concealing acts are usually carried out in a precautious manner to avoid drawing excessive attention, leading to little meaningful alteration of the financial statements (Zack, 2013, p. 247).

Nevertheless, disclosure schemes, like consolidation schemes, represent a material manipulation to deceive the users of financial statements (Zack, 2013, p. 157). Through the consolidation of the financial statements of the parent company and its subsidiaries, the corporate group is presented as a single economic entity. Regardless of the applicable accounting standards, the general idea behind consolidation fraud schemes revolves around the illegitimate omission of financially weaker subsidiaries and in counterpoint the inclusion of stronger ones. In both cases, the requirements for consolidation are not met by the respective accounting standard, rendering the consolidated statement misstated or fraudulent.

Akin to the types of fraud that surround consolidations, schemes arising from mergers and acquisitions have a material impact on the resultant financial statements once the acquisition procedure is complete. These schemes involve the illegitimate consideration of estimations and judgements in the course of the accounting process to combine different entities (Zack, 2009, pp. 91–93). In this way, fraud can occur in different stages. The first stage involves differentiation between a business combination and an acquisition of assets. Whereas a business combination involves the transfer of control over the business, an asset acquisition does not constitute a handover of control but only the respective assets (Zack, 2013, p. 165). The process can result in different outcomes. For example, the acquired business can become a wholly owned subsidiary of the acquirer. Second, only the assets of another company are acquired. Third, in the case of a roll-up merger, a new entity is created from the parties involved.[65] The difference between the three exemplary outcomes with regard to fraudulent schemes is the associated price of the transaction. In the case where assets alone are acquired, only the assets are priced, whereas a business combination usually does result in additional goodwill. The potential impact of misclassifying asset acquisitions as business combinations is the long-term inflation of assets.[66]

For business combinations under US-GAAP and IFRS, the so-called pooling-of-interest method, where the book values of the assets and liabilities are taken over in the allocation process, has been prohibited since 2001 (Zack, 2013, p. 165). Under the purchase method (acquisition method) of accounting, a fair value measurement of the respective positions is required. The fraud risk mainly revolves around the wrongful allocation of the purchase price to assets not subject to depreciation or amortization as well as assets with longer lifespans.[67] In this regard, the allocation of a portion of the purchase price to intangible assets that do not qualify for separate recognition would also be an example of a fraud risk factor that needs to be considered.

Finally, fraud can also be carried out when creating disclosure. According to Zack (2013, p. 189), disclosure fraud can be divided into four categories: omissions, incomplete disclosure, misrepresentation of information and confusing disclosure. Omitted and incomplete disclosure both share that the requirements of the underlying accounting standards regarding the disclosure of information are not met but differ in the extent to which information is omitted (Rezaee & Riley, 2009, p. 103). Incomplete disclosure ignores certain pieces of information about a topic, whereas omitted disclosure leaves out the topic completely. Misrepresented and confused disclosure occurs when the given information is incorrect or is presented in such a way that the recipient is unable to clearly assess and absorb it. Confusing disclosure may not be a fraudulent misrepresentation *per se* but serves as a potential indicator of other fraud-related actions (Wells, 2017, pp. 344–345). In practice, the omitted or altered information is usually of a negative connotation, thereby presenting the company's situation in a better light. A good example again is Enron, which in compliance with GAAP disclosed information on special purpose entities but intentionally obfuscated

[65] Assets acquired or transferred do not *per se* constitute a business but require the associated set of activities capable of managing the underlying assets for the purpose of providing an economic benefit Zack (2009, pp. 91–92).

[66] In most standards, goodwill is subject to impairment rather than amortization.

[67] This study will not go into further detail; see Zack (2009, pp. 93–95) for detailed analysis of manipulations of fair values in the case of business combinations.

the information so that hardly anyone could understand its ramifications (Jackson, 2015, p. 250).

2.4 Sources and Participants of Fraud Detection

As discussed in the previous sections, the reliability of standardized financial statements is important. In the following, fraud deterrence, prevention, and detection are discussed and different approaches are highlighted. In this regard, the need for audits to ensure compliance with accounting standards and different types of audits will be highlighted. This foundation serves as an initial evaluation of the demand for a reliable fraud detection tool based on publicly available accounting information.

2.4.1 Crime Signal Detection Theory

Crime signal detection theory provides a foundation for explaining the ability to detect fraudulent actions (Gottschalk, 2018, pp. 38–41). The ability of the detector to actually detect the stimulus is affected by the characteristics of both. On the side of the detector, perceptual sensitivity, which is determined by the physical and psychological state, specifies one's ability to discriminate signal from non-signal events (Szalma & Hancock, 2013, p. 1741). Beyond that, detectors may have different capabilities to distinguish between information-bearing recognition (patterns) and random patterns that only distract from actual information (noise) (Gottschalk, 2018, p. 38). This may be due to differences in competences, experience, expectations, and signal alertness. Pattern recognition is contingent on one's ability to contextualize. By being able to understand and explain relationships between information elements, patterns can be identified. In addition, psychological factors like personal bias in judgements may distort the result. The detector should not be a single person, but rather individuals working together in a team with the same goal. Indeed, team cognition may improve detection performance if converging perspectives and knowledge exists, but where it is ineffective, it can lead to failures in coordination, resulting in poorer detection capabilities (Wildman, Salas, & Scott, 2014).

On the level of the signal, the primary factor is intensity: stronger, more intense signals are easier for the detector to recognize. Another factor that is particularly important in reality is the signal value. This depicts the uncertainty of a stimulus to fall into either category. For example, the categories may be manipulated or unmanipulated. At the time of the decision, the detector acts with uncertainty, as the true nature of the stimulus is usually revealed later. When combining both signal and detector characteristics, a threshold can be determined, which is translatable to the likelihood of fraud being detected. Gottschalk (2018, p. 39) has created a ranking of potential detectors for white-collar crime, based on certain basic factors from crime signal detection theory. He suggests that journalists, for professional reasons, tend to have the greatest potential to act as crime detection sources, as compared to the internal control function and external auditors, which are ranking lower.[68]

Determining optimal decision thresholds, such as when a stimulus results in an action by the detector and a potential hit, is difficult. In many cases, hits are primarily influenced by intuition and subjective judgements, even when objective points of comparison exist (Phillips, Saks, & Peterson, 2001, p. 300). Thus, detection quality is mainly influenced by the discretion of the respective individual detector (the examiner). Verification procedures, such as the use of additional examiners are common and usually increase the positive hit rate but do not necessarily ensure better results. Technological advances have helped to identify patterns that would not otherwise have been uncovered through human examiners, or human examiners would have encountered severe difficulties in correctly recalling patterns through memorization and comparison (Phillips et al., 2001, p. 294). Even with the help of multiple examiners and the utilization of objective clues and hints, it is doubtful that detection results are comparable.

This study builds upon the idea that examinations of financial statement fraud require technological support, as examiners are constrained by materiality in general and the complexity of fraudulent behaviour in particular. As crime signal detection theory suggests, numerous factors influence detection capability, especially when stimuli are difficult to capture. Besides applying fraud theory and utilizing hints that have proved to be successful in previous studies, the high dimensional pattern analysis of corporate narratives may increase detection performance. The number of predictors and length of patterns provides one of the most extensive analyses available to date and may offer the potential to incorporate unknown or unrecognizable patterns. With the help of a machine learning approach and a comprehensive sample, the results may be of interest for everyone concerned with financial statement fraud.[69] In practice, individuals concerned with the detection of

[68] However, the factors have been examined and assessed in a rather subjective way, rendering the validity of the examination questionable.

[69] Machine learning is a subfield of artificial intelligence as it fulfils tasks based on the fundamentals of human intelligence, for example, the recognition of patterns from extracted data. Section 4.2 lays out the machine learning approach of this study.

fraudulent manipulations of company financial statements can be identified at an internal or external company level. Depending on the stage, the different groups that may ensure the truthfulness of the financial statements or uncover the manipulation will be highlighted hereafter.

2.4.2 Fraud Prevention, Detection, and Deterrence

Unwanted behaviour like fraudulent manipulations must be countered to avoid negative consequences for capital market efficiency. Deterrence, prevention and detection are the three major overarching fraud-related control categories under which specific approaches fall (Rezaee & Riley, 2009, pp. 84–89; Wells, 2017, pp. 359–361). Therefore, the three categories are separated based on their effects and relative to the timing of the fraudulent event. Prevention and deterrence are techniques to *ex ante* reduce the likelihood of fraudulent actions, whereas detection mechanisms are designed to *ex post* respond to fraud and uncover fraud, as can be seen in Figure 8. However, approaches do not exclusively belong to one category. For example, knowledge about a specific reliable detection tool or high detection rates can deter and therefore prevent fraudulent behaviour from happening (Rezaee & Riley, 2009, p. 88; Wells, 2017, pp. 370–371).

Deterrence is based on three factors: certainty, swiftness and severity of punishment (Wells, 2004, p. 2). Certainty is the perception and fear of potentially being caught, with swiftness depicting the immediacy of the actions against the perpetrator. The severity of punishment covers consequences like disciplinary, criminal, and civil actions. Deterrence approaches are mostly about establishing a culture and environment that does not accept or tolerate fraudulent behaviour and proactively training employees about the established mechanisms to detect fraud and how fraud can be reported (Wells, 2017, pp. 370–371). Fraud deterrence mechanisms, such as releasing a code of conduct, establishing an internal audit department, installing a whistle-blowing hotline, or relying on the four-eyes principle for crucial processes can be seen as a matter of discouragement to potential perpetrators. With SOX, regulators have reacted to increasingly severe accounting scandals by devoting greater energy to fraud prevention and detection as well as deterrence (Wells, 2017, pp. 309–310). Indeed, with SOX Section 807: Criminal Penalties for Fraud, the factor severity of punishment has been addressed and increased significantly. The efficiency of deterrence is mainly influenced by the interaction of the three factors in the occupational fraud context (Wells, 2004, p. 2). Wells (2004) argues that swift and certain punishment does not need to be severe. In contrast, Friesen (2012) provides evidence of an increase in the severity of punishment constituting a more effective fraud deterrence factor than an increase in certainty of being caught. Conflicting evidence, especially with regard to the experimental student design of Friesen (2012), was previously found by Block and Gerety (1995). Given that most empirical evidence comes from various fields of criminological studies, it is rather difficult to say how efficiently the three factors actually deter fraud and what represents a good method of deterring financial statement fraud in particular. This study is mainly concerned with the detection of fraud, where deterrence plays a minor role and is only touched upon with regard to the identification and creation of a comprehensive detection model based on publicly available information.

Fraud prevention and deterrence are often used synonymously (Wilhelm, 2004). The difference between the two lies in their order of influence on the fraudulent action, although a strict separation is not possible and is potentially unwanted due to the interaction between fraud deterrence, prevention and detection mechanisms (Jans, Lybaert, & Vanhoof, 2009). Figure 8 depicts the basic order of the three concepts. In contrast to deterrence, which is based on discouragement, prevention is related to hindering or stopping individuals from committing fraudulent behaviour (Wilhelm, 2004, p. 10). Therefore, prevention occurs when deterrence fails. Prevention mechanisms are designed to create barriers to make it more difficult to successfully commit a fraudulent act and furthermore to increase the likelihood of being caught afterwards. For example, verification and authentication procedures represent an important means of fraud prevention. The literature suggests to use a wide variety of prevention mechanisms, as preventing the fraud before its occurrence is usually less costly than the consequential damages (e.g. Wells, 2004; Bierstaker, Brody, & Pacini, 2006; Ogoun & Obara, 2013). However, determining and balancing the optimal amount of deterrence and prevention is difficult. The costs associated with implementing and maintaining mechanisms on the business side and also on the regulator's or lawmaker's side have to be justified against savings from deterrence and prevention (Rose, 2010).

While fraud prevention is based upon the goal of not letting fraud happen in the first place, fraud detection assumes that an unknown number of fraudulent actions have already happened and need to be uncovered (Wilhelm, 2004, pp. 10–11). Detection is about the identification of fraudulent behaviour. In the case of financial statement fraud, this refers to the intentionally misstated nature of financial statements, with examples including whistle-blower hotlines or regular and surprise internal and external audits. According to Bishop (2004, p. 123), the balance between prevention/deterrence and detection with regard to the effort companies usually devote to the respective mechanisms is 20% to 80%. He argues that although detection is a vital part of fighting fraud, more evenly balancing the two (as he barely distinguishes between deterrence and prevention) would help to reduce fraud. Bierstaker et al. (2006) have emphasized the more frequent and excessive application of modern data analytic approaches, which are seldom deployed according to their survey of accountants and auditors regarding the detection and prevention of fraudulent behaviour.

Figure 8 presents an overview of different fraud detection stages after the emergence of a fraudulent financial report.[70] This will serve as the foundation for the upcoming sections, which focus on fraud detection possibilities and responsibilities and how this study may help to improve the fraud detection process. Given that this study seeks to develop a comprehensive fraud detection approach, the fraud detection stages in the context of a fraudulent event will be highlighted, rather than an in-depth discussion of deterrence and prevention mechanisms.

According to Figure 8, once deterrence and prevention have failed, a fraudulent manipulation of financial reports occurs and sets the starting point for the five stages of detection. In the second stage, effective corporate governance (for example through a vigilant board of directors or an adequate internal audit function) may uncover the fraudulent action, if it has not already prevented it. In stage three, an independent external auditor provides reasonable assurance through control and substantial tests to provide assurance that financial statements are free of material misstatements. At this point, if undiscovered, a misstated financial report will be publicly released. It is then possible that misstatements are uncovered by other sources, like users or formal investigations by regulators. At stages four and five, the discovery of a misstatement will result in an SEC enforcement action. If the misstatement is discovered at stages two or three, it can be corrected and the truthful financial report issued in its place.

The following section will go through the stages and discuss the individuals who are involved in fraud detection. The section closes with a summary and an assessment of the need for fraud detection that relies on publicly available information

[70] Combining the view of Wilhelm (2004, p. 15) and Rezaee and Riley (2009, p. 23) to reflect the three stages of deterrence, prevention and detection while emphasizing on the latter.

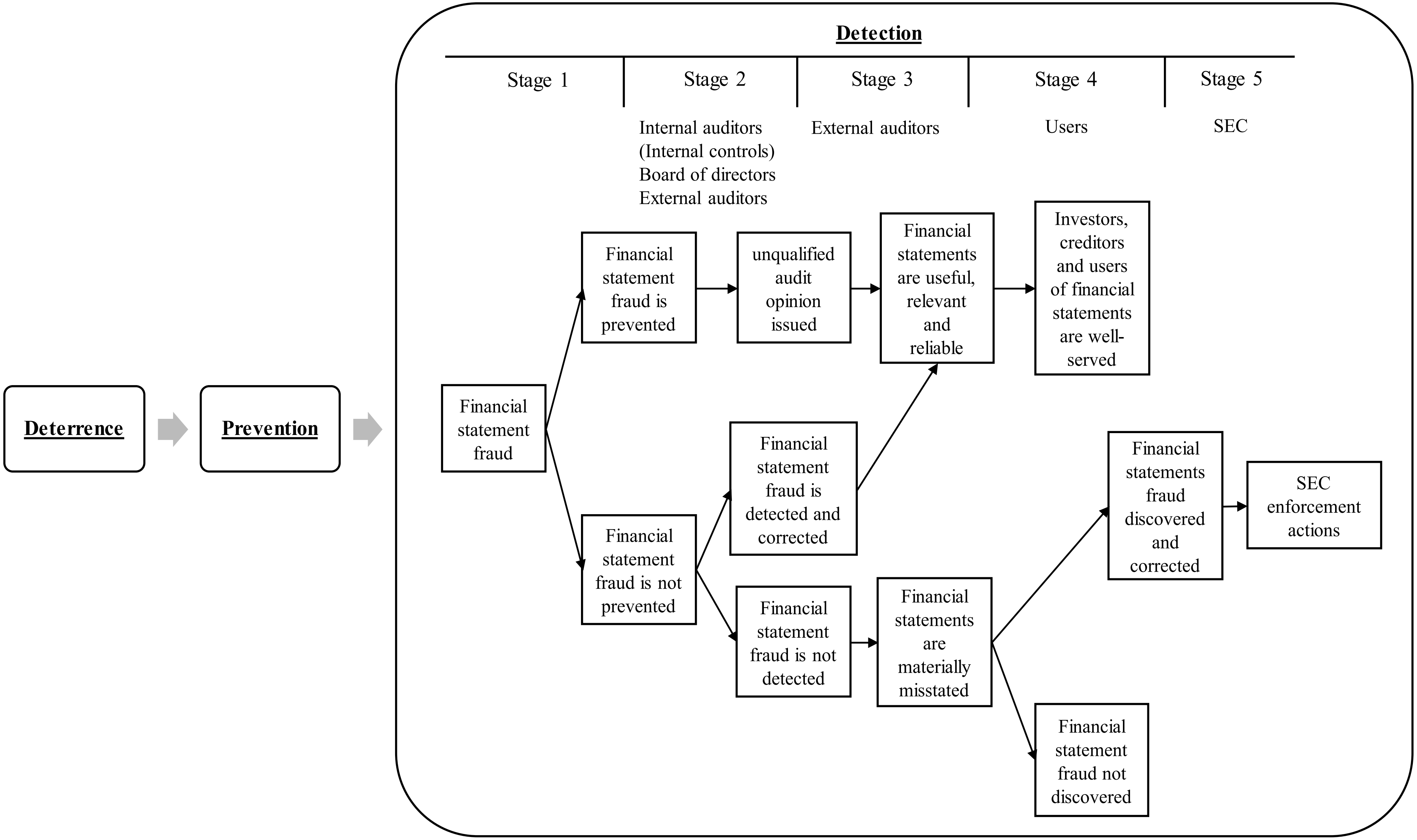

Deterrence
Prevention
Detection
Stage 1
Stage 2
Stage 3
Stage 4
Stage 5
Internal auditors
(Internal controls)
Board of directors
External auditors
External auditors
Users
SEC
Financial statement fraud
Financial statement fraud is prevented
unqualified audit opinion issued
Financial statements are useful, relevant and reliable
Investors, creditors and users of financial statements are well-served
Financial statement fraud is not prevented
Financial statement fraud is detected and corrected
Financial statement fraud is not detected
Financial statements are materially misstated
Financial statements fraud discovered and corrected
SEC enforcement actions
Financial statement fraud not discovered

Figure 8 – Fraud deterrence, prevention and detection process

2.4.3 Corporate Governance

Providing a general definition of corporate governance is difficult, even though it has constituted an omnipresent concept since the 1990s (Keasey, Thompson, & Wright, 2005, pp. 1–7). Owing to the universality of the term and the almost countless perspectives and interpretations available from different fields, including but not limited to accounting, finance, the regulatory field and economics, narrow and broad definitions have been formulated (Thomsen & Conyon, 2012, pp. 4–5).

From a broader angle, corporate governance reflects "the combination of applicable laws, regulations, and listing rules that facilitate, direct, and monitor corporations' affairs in attracting capital, performing effectively and efficiently, increasing shareholder values, and meeting both legal requirements and general social expectations" (Rezaee & Riley, 2009, p. 122). Within the fraud detection context and with origins in principal-agent theory, corporate governance can more easily be addressed as mechanisms that shareholders and stakeholders "require to protect their interest in a world of imperfectly verifiable actions" (Keasey et al., 2005, p. 2).

Corporate governance approaches differ around the world, albeit mostly between the USA and other Anglo-Saxon countries on the one hand and most of mainland Europe and Japan on the other (Gilson, 2001, p. 329).[71] The three main institutional factors that drive corporate governance in general and to a certain extent, the differences between those two groups are ownership structure, legal systems and capital markets (Rezaee & Riley, 2009, p. 125). The ownership structure can either be dispersed or concentrated. Dispersed ownership structure pertains to the situation in the USA or UK, where ownership is spread across institutional investors, as is typical for cross-border investments, whereas concentrated ownership can be found in countries such as Germany and revolves around family-owned companies, for example. The ownership structure determines the relevance and organization of internal (board composition) and external (rules, laws and regulation) control mechanisms.

Regarding the legal system, corporate governance can also be separated into two major categories: one-tier and two-tier systems. This divide is associated with the two traditional

[71] Although the differences seem to be of basic nature, a trend towards convergence of corporate governance approaches (usually in terms of their constitution of standards) has emerged in recent years (Gilson, 2001). However, an observable effect cannot easily be measured as Yoshikawa and Rasheed (2009) have stated. As these authors suggest, convergence seems to be more focused on formality than on actual substance. An extensive literature review can be found in Thomsen (2003) and Thomsen and Conyon (2012, pp. 78–82).

legal systems of common and civil law (Rezaee & Riley, 2009, p. 126). Civil law has its origins in Roman law and is today practised in countries such as Germany, France, Italy, and Spain; in contrast, common law came from the UK and is practised today in the USA and Australia. The two-tier system has two bodies that operate side by side, namely the board of directors and the supervisory board, whereas the one-tier system combines the duties of both bodies in one board. Given that this study relies on cases regulated by the SEC (which are typically US companies), the corporate governance structures discussed in the following refer to the one-tier system. The US corporate governance approach is widely regarded as being regulatory-led through policy-makers (Congress, SOX), regulators (SEC) as well as stock exchange listing standards and specific state laws (Rezaee & Riley, 2009, p. 126).

The third factor revolves around the importance of corporate governance for capital markets and their participants and vice versa. Capital markets' purpose to efficiently allocate scarce financial resources can be facilitated through suitable corporate governance mechanisms (Thomsen & Conyon, 2012, pp. 46–61). Corporate governance can lead to the positive development of capital markets by enhancing investor protection and lowering the cost of capital by reducing the barriers that constrain access to capital (Haque, Arun, & Kirkpatrick, 2008, p. 266).[72] Moreover, it may strengthen a firm's information environment and enhance stock valuation (Rezaee & Riley, 2009, pp. 126–127). Different capital market characteristics like the US stock market-centred capital market or Germany's and Japan's bank-centred capital market have also led to different developments regarding corporate governance.[73]

Corporate governance is a high-dimensional construct that differs by country through legal requirements and is subject to constant change. In the following section, the relevant mechanisms and participants of corporate governance in general and their influence and function in fraud prevention will be outlined in the detection and deterrence context.

2.4.4 Participants of Corporate Governance

The participants of corporate governance are derived from its mechanisms and principles and vice versa. Three have already been covered when discussing the factors that drive the

[72] A summary of the literature on the consequences of characteristics of corporate governance can be found in Haque et al. (2008).

[73] An extensive overview of differences in corporate governance approaches can be found in Gilson (2001). This study will not go into further detail as the sample is based on SEC regulated companies, hence the US approach is the primary focus.

approaches of corporate governance in different countries. In addition to legal and regulatory, market-based and ownership mechanisms, informal governance through social norms, reputation and trust or other codes build the foundation of corporate governance on a very basic level (Thomsen & Conyon, 2012, pp. 47–48). The combined set of mechanisms constitutes the participants of corporate governance. Each participant successfully fulfils his or her responsibility and particular role if their participation results in an added value that is often associated with shareholder (investors) interests (Rezaee & Riley, 2009, pp. 122–123). Therefore, defining an optimal set of mechanisms is important because every mechanism is associated with costs and benefits (Thomsen & Conyon, 2012, pp. 60–61). Regulatory standards like SOX considerably influenced the costs of corporate governance compliance through defining a higher quality framework for corporate governance (Coates, 2007, pp. 107–108).[74] Figure 9 depicts corporate governance participants, especially focusing on participants who are directly concerned with financial statement fraud and its detection.

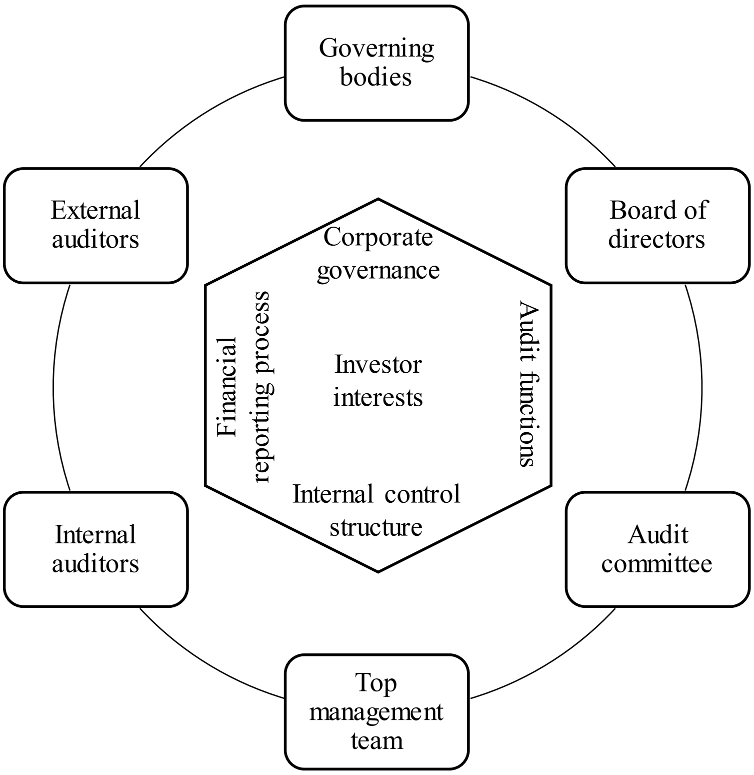


Figure 9 – Corporate governance participants[75]

[74] According to Coates (2007, p. 107), the audit fees and internal audit costs increased by approximately $1 million per $1 billion of revenue, leading to a modest trend of going private to avoid compliance costs, especially for smaller firms.

[75] In reference to Rezaee (2003) and Rezaee and Riley (2009, p. 123).

The participants can be divided into internal and external participants depending on their sphere of influence on the company. External participants are foremost governing bodies, like the SEC, PCAOB, FASB or the New York Stock Exchange (NYSE) and external financial statement auditors.

In the financial statement fraud context, the governing bodies have different purposes, based upon their general goals (Rezaee & Riley, 2009, pp. 138–279). Thus, the participants share complementary and interrelated roles and substitute for each other to a certain extent (Ogoun & Obara, 2013, pp. 126–127). Their common goal is to provide and ensure capital market efficiency, integrity, and safety while remaining globally competitive. Governing bodies lay the cornerstone for external financial statement auditors. Besides issuing accounting and auditing standards, they also oversee auditors.[76] External financial auditors are responsible for assessing the conformity of a company's financial statements with GAAP.[77] The conformity of the statements is essentially important for functioning capital markets, as market participants trust and rely on them. Overall, external governance bodies build the framework in which companies operate and serve as supervisors. In the aftermath of the recent major accounting scandals, governing bodies have reacted with a multitude of programmes to counter fraudulent behaviour and strengthen market participants' trust (Coates, 2007, pp. 91–92). Although governing bodies are not usually directly concerned with the detection of financial statement fraud, the programmes were created to strengthen the fraud deterrence, prevention, and detection environment, for example through the creation of the PCAOB to oversee auditors.

Internal participants, as depicted in Figure 9, comprise the board of directors, the audit committee, the top management team, and internal auditors. The board of directors' function is to create a system of checks and balances for the management team, which is necessary due to the separation of control and ownership (Thomsen & Conyon, 2012, pp. 142–143). In practice, the overarching function of the board of directors is the "overseeing, monitoring and controlling of management activities" (Rezaee, 2003, p. 27). The board of directors can consist of external and internal executive and non-executive members. The effectiveness of the board of directors in fulfilling its functions is highly dependent on the quality, reputation, and independence of the members. As the (top) management is typically involved in fraudulent actions like financial statement fraud, the board of directors has significant scope

[76] In the case of the PCAOB after the release of SOX.

[77] More on the role of external financial statement audits in fraud detection can be found in sections 2.4.6 and 2.4.7.

to counter fraud (Wells, 2017, p. 360). Empirical evidence has suggested that weaker corporate governance, for example in the form of a board of directors with the chairman also acting as the CEO, or boards that have fewer outside members, are more likely to bear fraud (e.g. Farber, 2005). The embodiment of corporate governance in general and the board of directors in particular has often proved to be significantly related to fraudulent activity (e.g. Uzun, Szewczyk, & Varma, 2019). Grove and Cook (2007) have developed a comprehensive list of red flags associated with corporate governance weaknesses derived from the cases of financial statement fraud around the early 2000s. They suggest that the focus on short-term performance and weak internal controls amongst others primarily have negatively influenced the occurrence of the scandalous cases and assume that the reactions from external corporate governance bodies like SOX to provide the required remedy.

The National Commission on Fraudulent Financial Reporting issued a report in 1987 that recommended that public companies implement and configure an effective audit committee.[78] The audit committee is an operating committee formed through the board of directors. The audit committee, as the name suggests, oversees internal and external audits and is responsible for maintaining an effective audit environment (Rezaee, 2003, p. 27). In this context, it must also ensure the integrity of the financial reporting process. Despite having the knowledge to understand the financial statements, the financial reporting process and being familiar with applicable laws, standards and regulations, members also have to be independent of the management to be able to fulfil their function. The audit committee arguably plays one of the most important roles in the context of fraud prevention, detection and deterrence through its centrality between other participants of corporate governance (Rezaee & Riley, 2009, pp. 157–159). The Treadway Commission regards the audit committee as an effective part of corporate governance, especially to combat fraud.[79] Therefore, audit committees' responsibilities and powers have been revised extensively through governing bodies in the aftermath of SOX to improve the quality of their financial statements and to ensure truthfulness (Coates, 2007, pp. 104–105). Empirical evidence on the role of audit committees regarding the occurrence of fraudulent manipulations is rather mixed. Whereas Beasley (1996) finds that firms that do not engage in fraudulent schemes have a higher proportion of outside members, the mere existence of an audit committee does not affect the likelihood of fraud in his sample of 75 fraud and 75 non-fraud companies. The

[78] The report is available at https://www.coso.org/Documents/NCFFR.pdf.

[79] SEC (1989), The Treadway Commission Report: Two years later. Retrieved from https://www.sec.gov/news/speech/1989/012689grundfest.pdf.

findings of Abbott, Park, and Parker (2000) suggest that audit committee independence and regular meetings can reduce the likelihood of fraudulent manipulations. The integrity of the audit committee is another factor that negatively affects the risk of fraud. Personal ties to top management compromise the audit committee's effectiveness (Wilbanks, Hermanson, & Sharma, 2017).

Management, which is installed through the board of directors, is responsible for executing the corporate strategy and managing the resources of the company. In this regard, it is also responsible for the financial reporting process and the conformity of statements with GAAP as well as the quality, integrity and reliability of the financial reporting process as a whole (Rezaee & Riley, 2009, pp. 184–185). This additionally implies that financial statements have to be free of irregularities, material errors, or misleading information. In the post-SOX era, for example, management has become responsible for implementing effective internal control over financial reporting (SOX 302.4 and SOX 404.b). These and other internal control mechanisms should be able to reduce the risk of fraudulent and similar unwanted behaviour if they are effectively and efficiently implemented (Walsh & Seward, 1990). Therefore, adequate implementation is crucial, as fraud risks remain if internal control systems are not adapted to company peculiarities and evaluated correctly, as one of the primary threats to internal control systems is management override (Rezaee & Riley, 2009, pp. 192–194). In general, management characteristics, board structure and control of management have often been shown to be associated with fraudulent actions and should be considered as a fundamental source when assessing fraud risk (e.g. Dechow, Sloan, & Sweeney, 1996; Sharma, 2004).

The management is supported and controlled by a number of corporate governance participants. Internal auditors assist the management, the audit committee and other participants of corporate governance in assuring that control systems are properly designed and maintained (Lin, Pizzini, Vargus, & Bardhan, 2011, pp. 288–290). Internal auditors have received increased attention in the aftermath of the WorldCom accounting scandal, in which internal auditor Cynthia Cooper discovered and reported fraudulent actions (Scharff, 2005). Internal auditors have also become an integral part of corporate governance through regulatory requirements and standards. For example, the NYSE required the installation of internal audit functions in 2004.[80] The impact of internal auditors on the financial reporting

[80] NYSE (2014), Corporate Governance Guide. Retrieved from https://www.nyse.com/publicdocs/nyse/listing/NYSE_Corporate_Governance_Guide.pdf.

system, especially on the outcome, has been empirically examined in detail. Prawitt, Smith, and Wood (2009) among others have noted that a higher quality of internal control function, especially internal audits, leads to better financial reporting in general in their examination of the probability of meeting analysts' perceptions and the quality of external financial reporting as a whole. Coram, Ferguson, and Moroney (2008) provide evidence of internal audits contributing extensively to fraud detection, especially to the self-reporting of fraud. They furthermore suggest that outsourcing the internal audit function is jeopardizing this effect.

Besides the positive effects of corporate governance mechanisms, there exist a number of downsides associated with the increasingly powerful presence of the related mechanisms. Shi, Connelly, and Hoskisson (2017) suggest that the powerful expectations of external governance participants can limit the autonomy and intrinsic motivation of managers, potentially even increasing the likelihood of fraudulent actions. Weir, Laing, and McKnight (2002) offer evidence of the impact of corporate governance mechanisms on firm performance in the UK. Although most of the mechanisms seemed to have little to no effect on company performance, the mixed results could be discussed with reference to similar studies, raising a question about the necessity of imposed governance mechanisms without evidence of positive effects. However, the authors also state that the effectiveness of corporate governance mechanisms as a whole is difficult to ascertain, given that most companies comply with the code of best practice and thus leave little room for comparison. Jain and Jamali (2016) also present mixed results in their comprehensive literature review of the effects of corporate governance.

Among corporate governance participants, internal and external auditors are a major source of fraud detection (Gottschalk, 2018, p. 32). In the brief discussion of corporate governance participants, the positive effects of both have already been suggested. In the following sections, internal and external audits will be discussed and their contributions to the prevention, detection, and deterrence of financial statement fraud presented.

2.4.5 Internal Audits

Internal auditors are positioned as an internal participant of corporate governance and thus have considerable potential to assist in detecting fraudulent actions.[81] Internal auditors

[81] Although the internal audit function can be outsourced rather than being provided by in-house sources (Coram et al., 2008, p. 544).

provide the internal control structure and are involved with operational and financial reporting systems (Rezaee & Riley, 2009, pp. 208–211). The internal audit function (IAF) has evolved and grown in recent years in many countries, mainly through regulatory novelties like SOX in 2002 (Soh & Martinov-Bennie, 2011, p. 605). For example, internal auditors may provide consulting and assurance services to comply with the provisions of SOX, like those regarding internal control systems, risk assessment and financial reporting.

Internal auditors and their role in corporate governance have been addressed extensively in authoritative reports and professional standards. The integration of internal control systems is attributable to the Foreign Corrupt Practices Act of 1977, a US federal law addressing the bribery of foreign officials as well as accounting transparency. In this concern, the Committee of Sponsoring Organizations of the Treadway Commission (COSO) was established by five private organizations in 1985 as a platform for the Treadway Commission, which was formed to inspect, analyse and formulate recommendations on corporate financial reporting (Vanasco, 1998, p. 24).[82] COSO adopted a framework in 1992 to design and test the effectiveness of internal control. This framework with the name "Internal Control – Integrated Framework" has been developed to become the predominant framework, which is globally applied and was updated in 2013 to incorporate changes to business and the operating environment (Janvrin, Payne, Byrnes, Schneider, & Curtis, 2012). The mission of COSO and its funders is "to provide thought leadership through the development of comprehensive frameworks and guidance on enterprise risk management, internal control and fraud deterrence designed to improve organizational performance and governance and to reduce the extent of fraud in organizations" (COSO mission statement).[83]

The general framework provided by COSO has influenced regulators and standard setters and has been incorporated into auditing standards. The PCAOB has released the AS2 – An Audit on Internal Control Over Financial Reporting Performed in Conjunction with an Audit of Financial Statements and the AICPA the SAS 55 and 78 – Consideration of Internal Control in a Financial Statement Audit, suggesting the COSO framework as a basis for auditors.[84] In addition, the Institute of Internal Auditors (IIA), which was founded in 1941 as a non-governmental organization, has issued a Statement on Internal Auditing Standards

[82] The funders of the Committee were professional accounting and auditing institutions based in the USA, namely: the American Institute of Certified Public Accountants (AICPA), the American Accounting Association (AAA), the Financial Executives International (FEI), the Institute of Internal Auditors (IIA) and the Institute of Management Accountants (IMA).

[83] The COSO mission statement as of 2019 is available at https://www.coso.org/Pages/aboutus.aspx.

[84] AS2 was superseded by AS5 in 2007.

3 (SIAS3) – Deterrence, Detection, Investigation and Reporting of Fraud in 1985 to counter specific types of fraud and to ensure that internal auditors prioritize engaging and assisting in fraud detection (Vanasco, 1998, pp. 22–23).[85] Although it is the most prominent framework, COSO is not the only one, especially with regard to the risk of fraud. Certainly, an alternative framework for internal auditing with the objective to deter, prevent and detect fraud is the so-called IFR² (Jans et al., 2009). The IFR² framework heavily relies on data mining techniques coupled with domain knowledge to fulfil its task. The frameworks in general hint at the importance of a structured detection approach, which is dependent on a foundation in fraud theory and the exploitation of technological advances.

Similar elaborations can also be found in the respective standards. The responsibilities of internal auditing departments regarding fraud based on the SIAS3 are the identification of red flags signalling the possibility of fraud, investigating symptoms of fraud and reporting findings to an audit committee (or another appropriate level of management or the board). Therefore, the internal auditor should possess sufficient knowledge regard fraud factors, think critically about control weaknesses, determine whether the corporate culture and environment fosters related behaviour, determine and assess corporate policies describing violations and how they should be tackled and assess and determine potential improvements (Vanasco, 1998, pp. 42–43). The SIAS3 also states that through the tradeoff of benefits and costs, not all fraud can be deterred and prevented and therefore detection mechanisms and investigative approaches need to be established. Consequently, standard setters, regulators and the literature suggest that internal auditors are more successful when they play a proactive role (Rezaee & Riley, 2009, pp. 214–218).

The internal audit function is viewed as the first-line defence against fraud. It can prevent and deter fraud through its bare existence but unfolds its true potential through thorough implementation and compliance with standards (Rezaee & Riley, 2009, pp. 212–214), especially when creating an environment in which potential perpetrators are unable to find the ground for fraudulent actions due to an anti-fraud culture and effective controls (Coram et al., 2008, p. 546). With regard to fraud detection, internal auditors might have an advantage over external auditors because they have a broader scope of insight into the company and are less constrained through the factors of time and associated costs (Rezaee & Riley, 2009). However, internal auditors do not act as additional or surrogate external

[85] The IIA sees itself as standard setter, educator and global voice of the internal audit profession. The information on the self-perception have been gathered from the official website: https://na.theiia.org/about-us/Pages/About-The-Institute-of-Internal-Auditors.aspx

auditors and thus do not ensure the fair presentation of financial statements (Rezaee & Riley, 2009, p. 215). Regarding this concern, the interrelated nature of corporate governance participants is significant, as the audit committee should ensure that the internal and external auditors' functions are complementary and coordinate them accordingly (Rezaee, 2003, p. 27). The positive influence of the audit function on fraud detection has also been found empirically. Beasley, Carcello, Hermanson, and Lapides (2000), as well as Coram et al. (2008), have found that the existence of an IAF and the support of internal auditors are associated with non-fraudulent firms, whereas fraudulent firms exhibit a less pronounced IAF. The audit committee should ensure the objectivity and independence of internal auditors (Christopher, Sarens, & Leung, 2009). For example, internal auditors may run into conflicts of interest, as they are hired, promoted, and evaluated through top managements, for whom they operate. After SOX, the task to hire and compensate the director of internal audit has therefore moved to the audit committee.

Overall, the internal audit function plays an important role in itself and provides supportive services for other participants of corporate governance. Standard setters and regulators have also engaged in strengthening the IAF by improving its effectiveness through requirements regarding the function's structure and the persons acting within it, building a fundamental basis for the detection of fraud in general and financial statement fraud in particular. The different frameworks imply the use of state-of-the-art data mining techniques for fraud deterrence, prevention, and detection, supporting the goal of this study in developing a comprehensive fraud detection approach that utilizes different data sources and modern detection models.

2.4.6 External Audits and the Expectation Gap

The auditing profession has a vital task to ensure market efficiency and has developed alongside companies over time. Nevertheless, the most noticeable changes to the profession have been based on uncovered cases of fraud. Following the stock market crash of 1929 and the issuance of the Securities and Exchange Act in 1934, it became obligatory for public companies to be independently audited for the first time in the USA (Singleton & Singleton, 2010, p. 284). An increasing number of companies wanted to be listed on the stock market and sell shares. Potential investors' need for reliable information about a company's situation led to an increase in demand for audited reports. Soon, private accounting firms emerged and took care of testing the adherence to GAAP. In this regard, a call for auditing rules could soon be heard after fraudulent activities at McKesson & Robbins 1939 went undiscovered (Byrnes et al., 2018, pp. 286–289). Most of the tasks were optional, like the physical inspection of inventories or the confirmation of receivables. The American Institute of Certified Public Accountants (AICPA) issued the Statement on Auditing Procedure (SAP) No. 1 in the subsequent year and significantly increased auditors' responsibilities. The next step was influenced by the integration of computers into the business sphere. In 1973, the Equity Funding Corp. scandal revealed a need for electronic auditing. Indeed, this company boosted profits via falsely reported commission income, which was not uncovered by the audits due to a lack of electronic auditing tools. This and technological advances induced by the access to computer-processed data led to the continuous increase in computer-guided audit routines on the one side and the obligatory introduction of the internal control system on the side of the companies.

Alongside changes to auditing tasks and tools, the expectations of auditors have changed considerably, especially concerning fraud detection (e.g. Epstein & Geiger, 1994; Kinney & Nelson, 1996). Initially, auditors were expected to provide (almost) absolute assurance against fraudulent manipulations. However, with increased business activity, auditors were somewhat constrained in determining the fairness of financial statements and in verifying every transaction (Byrnes et al., 2018, pp. 286–287). Absolute assurance has hence changed to a reasonable level of assurance. Although efforts have been made to communicate the (self-perceived) responsibilities of auditors, there was and still is a different perception between the public, regulators, jurors of what auditors are intended to do and what is required for them to do (Albrecht & Hoopes, 2014). This discrepancy is called the "expectation gap" and describes a fundamental issue of the audit profession.[86] The term was introduced by Liggio (1974). Since then, the literature has argued whether and to what extent it actually exists, with cumulative evidence proffered over time. Focusing on audit report references for loss contingencies, Kinney and Nelson (1996) have found that SEC, GAO and financial analysts (among others) hold auditors more responsible for audit effectiveness and expect a higher level of disclosure than auditors think is reasonable.[87] In 1988, a number of SAS were released in the course of public criticism to partly address the expectation gap issue. In particular, SAS No. 58 – Reports on Audited Financial Statements introduced an explanation stating that an audit (only) provides reasonable assurance regarding whether the financial

[86] See Chye Koh and Woo (1998) for a review of the literature.

[87] GAO stands for the US Government Accountability Office, providing services like auditing and evaluations for the US Congress.

statements are free of material misstatements.[88] SAS No. 53 – The Auditor's Responsibility to Detect and Report Errors and Irregularities[89] addresses fraudulent manipulations of financial statements and requires that an audit in accordance with GAAS (generally accepted auditing standards) be designed to detect errors and irregularities, albeit under the concept of reasonable assurance and not guaranteed.[90] However, a survey among investors by Epstein and Geiger (1994) subsequently concluded that around 70% of participants demand absolute assurance against fraudulent misstated financial statements.

The expectation gap is difficult to bridge (Porter, 1993). The accounting and auditing literature demands the design, implementation and maintenance of internal control mechanisms that are able to reasonably assure that reliable financial statements are crafted (Albrecht & Hoopes, 2014, p. 16). Auditors then have the responsibility to reasonably assure that those statements are prepared fairly in accordance with GAAP. Despite the public perception, it is mostly understood in the literature that auditors have the duty to ensure that financial statements are accurately represented and that CPAs do not need to act as crime or police detectives (Farrell & Healy, 2000, p. 25), despite the fact that the literature attributes a greater potential for fraud detection to auditors (Gottschalk, 2018, pp. 35–38, for a comprehensive literature review). AU section 316.12 Consideration of Fraud in a Financial Statement Audit even reads, "a properly planned and performed audit may not detect a material misstatement resulting from fraud".

With the aforementioned discussion in mind, it may be unsurprising that external auditors are not among the top sources of fraud detection. Gottschalk (2018) has examined 369 cases of white-collar crime between 2009 and 2015, finding that 287 of these were attributable to fraud or manipulation. The best source of detection for both categories were journalists, with about 28% of the detected cases. Of other typical mechanisms to detect fraud, internal control systems and audits of official authorities (e.g. tax controls) appear to be slightly better than external auditors in the overall comparison. Similar results have been attained by Albrecht, Albrecht, and Dunn (2001), who found around 18-20% of fraudulent cases are revealed by external or internal audits. Pedneault, Rudewicz, Silverstone, and Sheetz (2012) has estimated that around 12% of initial hints of fraud cases are discovered by external audits

[88] The effective date was 1 January 1989. SAS No. 58 is still in effect today.

[89] Errors are regarded as honest (unintentional) mistakes, whereas irregularities are deemed intentional misstatements.

[90] SAS No. 82 – The Auditor's Responsibility to Detect and Report Error and Irregularities superseded the old version. The effective date was 15 December 1997. In 2002, SAS 99 – Consideration of Fraud in a Financial Statement Audit superseded SAS 82 and placed greater emphasis on fraud detection in the aftermath of the major accounting scandals (Moyes & Baker, 2003).

and an additional 19% through internal audits. Other studies dealing with the ability of auditors to detect fraud have also yielded rather negative results. According to Beasley (2003), the ability of financial statement audits to detect fraud after the numerous auditing standard revisions has not improved. Albrecht and Hoopes (2014) have inspected cases in which external auditors were unable to detect fraud and have analysed the factors that most likely lead to failure. In 10 of the 19 cases, the auditors were conducting the audits in accordance with applicable auditing standards but were unable to detect the fraudulent manipulation. In five cases, the auditors were negligent in performing the audit, whereas in four an early settlement was reached. The reasons were manifold in both groups. In the case of negligent audit practices, the reasons ranged from inadequate training and poor audit execution to a failure to exercise due care and a lack of independence, for example through bribes. In the 10 cases where the auditors were unable to detect the fraudulent manipulation but were not negligent, the reasons were either attributable to the fraudster or to participating (colluding) parties. Reasons can be plentiful, for example through lying or people being reluctant to disclose what they know, or alternatively through the nature of accounting and auditing, which mainly focuses on the inability to detect fraudulent records in a sample-based audit in the voluminous accounting records.

To summarize, external financial statement audits are, in spite of public opinion to the contrary, unable or unintended to detect every instance of financial statement fraud. However, detection results by external auditors seem to be rather disappointing, as they most likely have considerable potential to detect manipulations, leading to a need for additional mechanisms and tools as standard setters and the literature have already demanded. Therefore, in the following section, the audits specifically designed to detect and investigate fraud will be discussed.

2.4.7 Fraud Audits

Despite the existence of regular financial statements and internal audits, an additional type, the so-called fraud audit, has emerged. Where suspicion is raised at the conclusion of an aforementioned audit, a fraud audit may be undertaken to examine the case in greater detail and with respect to fraudulent actions (Singleton & Singleton, 2010, p. 12). Fraud audits require experts from the field of accounting and auditing. In this regard, the term forensic or investigative accounting has appeared (Vanasco, 1998, p. 23). The difference between fraud auditing and forensic accounting is not clearly defined but according to Singleton and

Singleton (2010, pp. 12–15), fraud auditing is related to the specialized approaches and methodologies necessary to discern fraud in an audit, whereas forensic accounting is the general term concerned with investigative work in the domain of accounting with a broader scope. Forensic accounting may also involve court-related work like gathering evidence and translating complex financial transactions to laypersons.

Regular financial statement audits and fraud audits differ. A fraud audit is neither a defined term nor a defined professional service and may also be called a fraud examination or fraud investigation (Singleton & Singleton, 2010, pp. 12–14). Fraud audits are usually performed in the aftermath of financial statement audits and/or if suspicion of potential or actual fraud is raised by any source. As mentioned in section 2.4.6, fraudulent manipulations are not only detected by auditors but often other sources that provide the relevant clues leading to suspicion (Gottschalk, 2018, p. 32). One example of a fraud audit would be the case of an auditor being sued by shareholders after fraudulent manipulations were discovered in audited financial statements. The fraud audits may reveal whether the manipulation could have potentially been uncovered by the regular audit.

Fraud audits are much more in-depth, time-consuming and therefore costly than regular financial statement audits. They are closer to the concept of absolute assurance and thus the inspection of every transaction. Moreover, fraud audits also involve investigations of the accounting and controls system in general, to identify potential weak spots and ways in which safety procedures might be circumvented (Vanasco, 1998, pp. 42–43). Cases have shown that fraud audits to consume more than 50 times the amount of hours spent on regular audits (Albrecht & Hoopes, 2014, p. 15). Albrecht and Hoopes (2014, p. 14) have argued that fraud audits represent "the only way to satisfy the expectation that financial statement auditors always detect material financial statement fraud". The procedures of these audits are not considered by GAAS or PCAOB auditing standards. Moreover, Albrecht and Hoopes (2014, p. 15) claim that these standards would also not be feasible to fulfil the task.

Besides the already mentioned differences in the timing and the scope of the audits, additional peculiarities are important. Whereas regular financial statement audits are performed by CPAs, fraud audits are more likely to be executed by Certified Fraud Examiners (CFEs) or forensic accountants (FA) (Singleton & Singleton, 2010, pp. 14–15). CFEs and FAs usually share knowledge of accounting and auditing but are additionally trained in fraud investigation. Moreover, their tooling may involve data mining and analysis techniques, which are usually not part of a regular financial statement audit (Bierstaker et al., 2006, p. 530).[91] The Association of Certified Fraud Examiners (ACFE) offers programmes to accredit CFEs and may help to establish fraud policies in companies (Bierstaker et al., 2006, p. 523).[92] Among the big accounting firms (at least for the big four), fraud examinations have become typical services that are offered to clients. In spite of their differences and similarities, their most obvious divergence is their purpose. Indeed, whereas a regular financial statement audit provides reasonable assurance of the financial statements being prepared in accordance with applicable accounting standards and in its course uncovers misstatements and potential fraud, a fraud audit is solely used to uncover the latter.

2.4.8 Publicly Available Information and Fraud Detection

The availability of useful data is essential in designing a reliable and well-performing detection model. The different types of audits that have been discussed so far can rely on access to internal and publicly available data alike. In terms of financial records, internal data can be accessed at the transaction or journal level, yielding a higher level of insight than aggregated financial statements (Gee, 2015, p. 320).[93] In addition, gathering non-financial data inside the company may help to improve the detection results, especially when red-flags can be utilized (Singleton & Singleton, 2010, pp. 110–111). Examples cover clues related to characteristics of corporate governance or top executives. Eining, Jones, and Loebbecke (1997, p. 10) present a list of red-flags in their assessment of decision aids for auditors with entries like: “the company has a weak control environment” or “the chief executive officer and the chief financial officer both have an aggressive attitude toward financial reporting”. These data may be associated with factors from fraud models that could not be incorporated using only publicly available data, or at least not to a satisfactory extent.[94]

However, this barely limits the usefulness of the results of this study. It is important to note that detection performance could most likely only be improved further with clues extracted from additional, preferably company internal data sources. If the detection models in this study can potentially distinguish annual reports into truthful and fraudulent categories,

[91] Although the technological tools of external auditors are also improving, their focus may differ and be designed to assist in the audit process rather than to detect fraud.

[92] The ACFE was founded in the USA in 1988 to counter white-collar crime. It has developed to become the world’s largest anti-fraud organization, with 85,000 members in 125 countries as of 2019, offering training, education, and anti-fraud policies. Retrieved from https://www.acfe.com/about-the-acfe.aspx.

[93] Although, Gee (2015, p. 320) states that one should start an audit or an investigation with the financial statements and intensify in certain areas using additional data from journal entries, transactions and also outside of normal business transactions when necessary.

[94] See the explanations on fraud models and their translation into detection models in section 2.2.

this process would be even easier with more predictors extracted from additional data sources or from complementary decision aids (Eining et al., 1997, pp. 16–17). Moreover, such a detection model is not only relevant for auditors and their contributions to functioning capital markets but to all stakeholders and shareholders alike. It may be especially relevant in cases where auditors support the fraudulent scheme or are involved in some way. Additionally, auditors have shown not to be the most reliable source of fraud detection. Therefore, when having stakes in a company, further determining the truthfulness of the audited publicly available financial statements like the annual or quarterly reports may be helpful. From an investor or analyst standpoint, for example, such detection models could be exploited to evaluate truthfulness on an additional basis before translating the disclosed information into decisions.

Furthermore, the detection models developed and tested in the following chapters could easily be extended to incorporate additional data sources if the identification of the company and year are provided. In this regard, the models are not limited to a US sample but can be employed for all companies reporting in English. The detection performance might suffer if the models cannot be trained on foreign data but should still be able to detect cases if the annual reports are similar in nature.[95]

[95] See section 5.2.1 for a comparison of qualitative features from this and similar studies.

3 Literature Review of Financial Statement Fraud Detection Studies

The following literature review will focus on relevant studies in the field of textual analysis in accounting and finance and financial fraud detection. The literature review will be used to identify the shortcomings and research gaps of contemporary work in the field of financial statement fraud detection and prepare the research goals for this study. After identifying the research gap and formulating the detailed tasks for this study, the methodology will be outlined, before presenting the results.

3.1 Textual Analysis in Finance and Accounting

Textual analysis in the accounting, finance and auditing literature has become increasingly popular over the past 20 years (Fisher, Garnsey, & Hughes, 2016, p. 160).[96] Especially the accounting domain, by definition being concerned with communicating corporate information, offers extensive possibilities in applying textual analysis. Narratives in corporate disclosure can be studied from various perspectives. On the one hand, the focus might be on the recipients' side, i.e. by focusing on perceived information. This literature often explores economic implications like changes in equity evaluation, the cost of capital, changes in analysts' forecasts and so on. On the other hand, disclosure might be examined from the perspective of creators. Textual analysis facilitates consideration of how textual information is crafted by management and opens the door to analyzing the interaction of attributes of corporate disclosure and the underlying management and firm characteristics. In order to predict fraud, this perspective on the management's decision-making process is vital. It is plausible that managers are the ones committing or instructing the fraudulent actions and therefore the focus should be on the textual documents in their present condition.

Li (2010a) has found empirical evidence for self-serving attribution bias in the MD&A section of annual reports. Communication patterns reveal an association between superior firm performance and managers' self-expression in the MD&A section. Moreover, self-serving attribution bias is associated with overconfident behaviour, resulting in more optimistic forecasts. Overall, the findings have revealed that management characteristics that result in corporate policy-making can be captured by textual patterns. Narrative peculiarities and management intentions have also been the focus of Abrahamson and Park (1994) and Li

[96] Fisher et al. (2016) compiled an extensive overview of natural language processing in the accounting, auditing and finance literature. Also see Lewis and Young (2019) for an explanation of commonly used methods for automated analysis of financial texts and their implications on corporate reporting.

(2008), who examined concealment based on the obfuscation theory, which focuses on management's intention either to hide information completely or to obfuscate readers using largely incomprehensible language, increasing ambiguities and information asymmetry as a result.

Li (2008) applied readability scoring to measure the relationship between reading ease and firm performance as well as earnings persistence. He found empirical evidence indicating a negative relationship between firm performance and readability. The results of Abrahamson and Park (1994) support the potential opportunistic behaviour of managers when crafting corporate disclosure narratives. They find evidence of managers concealing or obfuscating negative outcomes if the institutional environment provides greater opportunities to do so. Othman et al. (2012) have also shown that the readability of a company's annual report reduces in the two years prior to fraudulent manipulation, revealing potential evidence of management obfuscation theory in the financial statement fraud context. Social psychology suggests that lying and deception can be identified through linguistic peculiarities for example differences in narrative complexity, tone, or the existence of less self-relevant and self-referencing language (e.g. Newman, Pennebaker, Berry, & Richards, 2003). Similarly, Buller and Burgoon (1996) discussed and examined the interpersonal deception theory, which deals with deceit in interpersonal communication and found evidence for a variety of linguistic features both in written and spoken language that indicate deceptive intentions. In this regard, Zhou, Burgoon, Nunamaker, and Twitchell (2004) however, appeal that although it has been shown that clues in communication exist that hint at certain states of the sender, these clues may vary considerably across different types of communication mediums and for different types of states that are studied (e.g. deceit) and require a specific assessment to be successful.

In the context of negative financial situations, a number of studies have focused on companies in peril. Narrative peculiarities in the narratives of affected companies are examined around related events, such as bankruptcy or financial distress (e.g. Tennyson, Ingram, & Dugan, 1990; Smith & Taffler, 2000; Boo & Simnett, 2002). Mayew, Sethuraman, and Venkatachalam (2015) examined the language of MD&A reports to find evidence of linguistic characteristics being associated with subsequent bankruptcy events. More specifically, they revealed tone and the management's opinion of going concern to have predictive power for future bankruptcy. Boo and Simnett (2002) provided similar results for companies in financial distress, finding that companies that do not comment on future prospects are more likely to fail in the future than are firms from the comparison group. Non-disclosure, i.e. an absence of narratives about a certain topic, may also be associated with the subsequent economic condition of firms. This has complemented the findings of Abrahamson and Park (1994) and provides additional useful guidance in establishing forecasting models based on linguistic characteristics. Loughran and McDonald (2011) developed wordlists to measure tone in corporate disclosure and linked them to firm characteristics of their sample including trading volume, unexpected earnings and material weaknesses amongst others, providing evidence of tone in corporate disclosure hinting at corporate peculiarities for a large sample of 10-Ks between 1994 and 2008.

Other studies have focussed on the influence of management and corporate governance characteristics on different types of corporate disclosure. Bamber, Jiang, and Wang (2010) have examined the influence of demographic characteristics like age cohort, career track, military experience and qualification of managers on corporate disclosure. They find that manager specific fixed-effects exist. For example are managers from the accounting, finance and legal career track and those with military experience associated with rather precise disclosure styles. Rouf (2011) has studied voluntary disclosure and finds empirical evidence for corporate governance and management characteristics being associated with the amount of voluntarily disclosed information. Especially board size and independence, as well as the existence of an audit committee, influence voluntary disclosure positively. Similar results have been found Jizi, Salama, Dixon, and Stratling (2014) for corporate social responsibility (CSR) disclosure. For their study in the banking industry, besides common corporate governance characteristics like frequency of board meetings and size of boards being positively related to CSR disclosure, they suggest powerful CEOs (as of the existence board duality) counterintuitively to increase transparency for personal interests like reputation. Trotman and Bradley (1981) studied voluntary corporate social responsibility disclosure and found that the managements' decision horizon influences the existence and amount of voluntary disclosure. Davidson, Dey, and Smith (2019) have examined the impact of CEO materialism on CSR and the underlying disclosure and find that materialists are negatively associated with their company's CSR score. Their results depict how a character trait of the management directly influences disclosure and hence may be captured by textual analysis.

On the other hand, a vast amount of studies have focused on the recipient side, revealing the impact of linguistic peculiarities on recipient behaviour. For example, Lawrence (2013) have presented evidence that suggests that the readability and complexity of financial disclosure are vital for private investors' decision-making. Private investors are more likely to invest in firms that provide disclosures with higher readability. In contrast, he has found no evidence of institutional investors' decision-making being affected by the readability of narratives. Additional evidence on the influence of readability of corporate disclosure on investors was presented by Lee (2012), who showed that longer and harder to read 10-Qs reports result in delayed stock price reactions for the three-day post-filing drift window. The results from Lee (2012) imply that management can affect information efficiency by influencing textual characteristics in their communication instruments. Lehavy, Li, and Merkley (2011) focused on financial intermediaries and corporate narratives. They recognized greater analyst dispersion for companies providing MD&As with lower readability. Moreover, Li (2010b) and Davis, Piger, and Sedor (2012) have provided empirical evidence of sentiment in corporate disclosure being associated with future firm performance, with the market responding to sentiment signals. This finding has further implications for the credibility of textual information and their value-relevance for investors. Given that investors react according to the positive signal given by management, they value the credibility of information over management's opportunistic behaviour.

From a wider perspective, the results suggest that textual components of corporate disclosure can reveal subtle details of a firm and its management. Moreover, managers can and (probably) actively do influence recipients' ease of absorbing the information provided by manipulating its textual components. For the purpose of financial statement fraud detection, where subtle details beyond the raw financials and from the inside of the firm and the management could potentially be important, textual analysis offers considerable potential.

3.2 Financial Statement Fraud Detection

Fraud detection is a complex topic that necessitates a multidisciplinary approach (Ramamoorti et al., 2013, pp. 6–7). Indeed, it is unlikely that fraud can be countered using isolated and narrow approaches (Singleton & Singleton, 2010, p. 145). The fraud detection literature has examined manifold issues regarding financial statement fraud and its detection.

Over the past years, two streams have emerged in the financial statement fraud detection literature. Early research focused on quantitative information in annual reports or other external communication instruments, supplemented by capital market data (e.g. Persons, 1995; Beneish, 1999; Kaminski et al., 2004; Dechow et al., 2011; Abbasi et al., 2012). The typical design extracts financial and non-financial metrics for companies or firm years where their disclosures are known to be truthful and fraudulent.[97] A comprehensive set of potential predictors exists, although these indicators vary considerably between studies.[98] Over time, additional and more sophisticated metrics, as well as more advanced classification approaches, have been developed and applied to detect fraudulent cases (e.g. Alden, Bryan, Lessley, & Tripathy, 2012). The potential disadvantage of financial metrics is the fact that they are the subject of manipulations and may hide more than they actually reveal, moreover contain texts more diverse and denser information than the plain numbers (Goel & Gangolly, 2012, p. 81). Therefore, besides the commonly used financial metrics, some research has also started to incorporate non-financial measures. Identifying non-financial measures and creating a detection model by building upon the relationship of non-financial and financial metrics (for example, facilities' growth versus revenue growth) may help to identify manipulations to the raw financials (Brazel, Jones, & Zimbelman, 2009).

Besides constructing predictors from company financials, a number of studies have focused on exploiting the distributional characteristics of the underlying business-related numerical data (e.g. Nigrini, 1996; Durtschi, Hillison, & Pacini, 2004; Saville, 2006; Watrin, Struffert, & Ullmann, 2008). Thereby, the data has been examined on different levels like journal entries or the aggregated results as disclosed in the financial statements to detect anomalies that deviate from the expected frequency distributions, hinting at potential manipulations. The most common methods that have been utilized for fraud detection relied on Benford's Law.[99] Benford (1938) noted that lower digits occur more often than higher digits, based on his impression of the condition of pages in books covering common logarithmic tables, which he later confirmed by 20 different lists of numbers from different sources. Fraud detection utilizing Benford's Law is conducted by calculating deviations of occurrences of digits at certain places in numbers from the expected occurrence provided by the Benford distribution (e.g. in ~30% of the time the first digit is a 1) (Nigrini, 1996). At this point, it is important to note that the basic assumption of this type of investigation is that the unmanipulated cases actually follow the Benford distribution (Watrin et al., 2008, pp. 235–236). Therefore, it has to be assessed beforehand, if the underlying data like journal

[97] Although there might still exist a number of fraudulent observations in the group of truthful observations due to the fact that fraud is determined based on external data sources like AAERs, which are assumed to be unable to detect every case. See 4.1.1 for a discussion of fraud proxies.

[98] A detailed summary of quantitative predictors will be given in 4.1.4, where the selection for this study will be discussed.

[99] Sugiarto, Noorzaman, Madu, Subagyo, and Amiri (2017) showed that for first digit tests of fraud, frequency distributions resulting from Fibonacci sequences and Lucas sequences are equally well suited as they obey Benford's Law.

entries or numerical data in financial reports do even obey Benford's Law to make this type of fraud detection approach applicable at all.[100] Benford's Law requires the underlying data to satisfy a number of assumptions, which may not be met by all types of numerical data from finance and accounting (for example a built-in minimum or maximum like transaction costs of investments or a threshold value to be met to record an asset, rounded data, etc.).[101] Moreover, some fraudulent schemes cannot be captured by using Benford's Law at all. These schemes involve manipulation to data that is not suitable for examination of Benford due to the fact that it does not suffice the relevant assumptions, does not have the required sample size or is perpetrated through the absence of transactions (Durtschi et al., 2004, p. 27). In the above-mentioned examples, deviations from Benford's Law are seen as anomalies, which need to be investigated further. In those cases, each observation is tested against the distribution. However, there have been efforts to utilize Benford's law further. Saville (2006) tested the potential of distinguishing fraudulent and non-fraudulent companies using Benford's Law on numbers from the income statements of a pair-matched sample of 34 companies from South Africa by using a regression model to assess the level of conformity of each company's financial information to Benford's Law. According to his results, all fraudulent observations occur to be conspicuous, whilst only three of the non-fraudulent observations appear to have manipulated. Under the above mentioned difficulties, especially the questionable assumption that data from financial statements (which are the primary source of quantitative predictors in this study) differs in its distributional characteristics with regard to the Benford distribution between manipulated and non-manipulated observations, this study desists from utilizing a similar approach and rather relies on the more common and approved predictor variables.

Regardless of the nature of the detection model, most studies concerned with financial statement fraud detection rely on samples from the USA, but there is also plenty of empirical evidence available from other countries using similar models. For example, Spathis (2002) and Spathis, Doumpos, and Zopounidis (2002) have constructed a sample based on Greek companies, Dikmen and Küçükkocaoğlu (2010) have relied on a Turkish sample and Liu, Chan, Alam Kazmi, and Fu (2015) and Ravisankar, Ravi, Raghava Rao, and Bose (2011)

[100] Henselmann, Scherr, and Ditter (2013) have examined all monetary line items in 10-K XBRL (eXtensible Business Reporting Language) filings in 2012, suggesting that those are obeying to Benford's Law. Similar results have been found by Quick and Wolz (2003) for annual report data of large listed German companies between 1994-1998. However, this does not necessarily imply that Benford's Law is able to distinguish between manipulated and non-manipulated observations.

[101] See Drank and Nigrini (2000, p. 132) and Watrin et al. (2008, p. 222) for further effectuations on the assumptions.

have examined Chinese companies. The results suggest that regardless of country and regulatory sphere, detection models created from similar predictors seem to be able to detect fraud regardless of the origin of the company.

Perols, Bowen, Zimmermann, and Samba (2017) have produced one of the most comprehensive studies and rely on prior quantitative predictors from Cecchini, Aytug, Koehler, and Pathak (2010b), Dechow et al. (2011) and Perols (2011). Their findings have a broad scope of implications for the financial statement fraud detection literature. First, they reveal how different sampling methodologies can be adopted to address the problem of imbalanced data sets in the fraud detection context, where the cases of interest represent rare events compared to the overall population. Moreover, they boost existing models by complementing them with additional predictors. Finally, they show that addressing different types of schemes of financial statement fraud rather than fraud as a whole may lead to a better understanding of the subject overall.

Abbasi et al. (2012) have set their goal in enhancing the performance of fraud detection approaches relying on quantitative predictors. Before introducing their MetaFraud framework to increase the detection results, they highlight the poor detection performance of baseline quantitative predictors from annual reports. By using additional contextual information like industry comparisons and enriching the vector with data from previous years as well as introducing an adaptive learning algorithm (learning from year to year), the results could be improved considerably, scoring extraordinarily well (area under the receiver operating characteristic, AUC of 0.931) for their final results, which is outstanding compared to other approaches.[102] However, the results might be somewhat skewed owing to sampling issues, as every fraudulent report is regarded as a single instance of fraud, resulting in the same company being detected over and over again in longer-lasting schemes. Regardless of this potential issue, they demonstrate how far the results can be improved by enriching the set of predictors with additional information.

The second stream complements the indicators from quantitative data by exploiting additional sources to extract qualitative data (textual information) from corporate reports (e.g. Goel et al., 2010; Humpherys, Moffitt, Burns, Burgoon, & Felix, 2011; Purda & Skillicorn, 2015; Hoberg & Lewis, 2017). In addition to these studies, Cecchini et al. (2010a), Dong, Liao, and Liang (2016) and Brown, Crowely, and Elliott (2018) have combined predictors from both qualitative and quantitative data to significantly improve

[102] Common metrics for evaluating detection outcomes are discussed in section 4.2.5.

their prediction results compared to single data-type detection models. The most relevant studies will be discussed in the following to highlight the state of the art, present the results attained and hint at the contribution of this work. However, comparing the raw prediction results across studies is rather difficult as methodology and sample composition vary considerably.[103] Moreover, a best practice solution for reporting the results has yet to be found. Nevertheless, the individual results of each study will be presented to raise general awareness of the possibility of detecting fraud in the respective studies, while keeping the differences between them in mind.

Cecchini et al. (2010a) extracted textual predictors from the MD&A section of annual reports of a matched sample consisting of 61 truthful and 61 fraudulent observations. To discriminate between the groups, they selected words and multi-word phrases of up to three words with a different frequency of occurrence in both groups. The qualitative predictors were preprocessed by removing stop words, stemming and part-of-speech tagging. They complemented their qualitative predictor list with financial variables based on Beneish (1999). Their results revealed a significant increase in accuracy from 75% to 82% when enriching the list of predictors based on qualitative information with quantitative financial information.

Goel et al. (2010) noted that textual analysis yields the best prediction results when the texts are preprocessed by removing stop words and pruning the textual elements. With the resulting bag-of-words, they achieved better classification results than with the original version of textual elements. Furthermore, they divided the sample into several stages of fraud, thus capturing changes in language before and after the actual fraud period. With their best attempt, they were able to classify almost 90% of the 405 fraudulently observations in their mildly balanced sample correctly.

Humpherys et al. (2011) examined the linguistic peculiarities of 202 publicly available financial disclosures. In contrast to most other studies, they used metrics to measure the attributes of the language in deceptive and truthful texts. The 24 metrics captured affect, complexity, diversity, expressivity, non-immediacy, quantity, specificity, and uncertainty of the textual information. They found that fraudulent disclosures consist of a greater portion of active language words, imagery, pleasantness, and group references and have less lexical diversity than truthful ones. They assumed that the managers of fraudulently manipulated financial statements try to make their reports appear credible while avoiding actual content

[103] See West and Bhattacharya (2016) for a comparison of methodologies of fraud studies.

in the texts with the change in language. In their best approach, the indicators were able to classify 67% of the 202 pair-matched observations correctly.

Goel and Gangolly (2012) examined indicators of fraud in the language of annual report narratives and beyond by capturing document presentation style. They found significant evidence that six categories of linguistic clues were associated with fraudulent financial reports. The use of complex sentinel structures, readability, positive tone, passive voice, uncertainty markers, and adverbs proved to be decisive markers in fraudulent reports. In contrast, presentation like the formatting or the punctuation of the annual report seemed to lack indicators of fraudulent behaviour.

Purda and Skillicorn (2015) relied on a similar approach as Cecchini et al. (2010a). They utilized a bag-of-words with textual predictors that discriminated well between fraudulent and truthful reports. They incorporated 4,895 interim 10-Q reports of which 23% were fraudulent and designed their sample to examine fraud detection in a cross-sectional and time-series setting. The results reveal considerable changes in their underlying detection metric prior to a fraudulent event. They compared their results to eight alternative language-based and financial-based fraud detection models and indicated that their language-based approach worked best with an accuracy of 82% of correct classifications. Furthermore, they revealed that language-based models and financial-based models are best used together because they detect different cases.

Dong et al. (2016) extracted predictors from the MD&A sections of 805 fraudulent observations using systemic functional linguistics theory. They relied on seven information types based on the following three metafunctions: ideational, interpersonal, and textual. Each of the metafunctions was deemed to capture different concepts. Under the ideational metafunctions, they utilized topics (generated via latent Dirichlet allocation), opinions, and emotions (studied by sentiment analysis). The interpersonal metafunction covered modality (for example, the number of modal words relative to the number of verbs) and personal pronouns (for instance, the number of pronouns relative to the number of verbs). The textual metafunction covered writing style (for example, predictors associated with readability measurement) and genre (word frequencies based on n-grams). They also combined the three metafunctions with metrics from quantitative data, scoring higher results than only qualitative or quantitative data on 10-fold cross-validation. For their matched sampling approach with 805 fraudulent and 805 truthful observations, they revealed that quantitative predictors scored an accuracy of 65.97%, while the combined model achieved 82.49%.

Chen, Wu, Chen, Li, and Chen (2017) extracted fraudulent feature terms of up to three-word length while using text preprocessing, removing punctuation and stop words. Using a static feature vector size of 240, they distinguished between fraudulent and non-fraudulent Chinese/Taiwanese shareholder reports between 1995 and 2012, covering a sample size of 45 fraudulent and 135 matched non-fraudulent observations.[104] They used decisions trees and support vector machines as classifiers and achieved an accuracy of 85.25%.

Hoberg and Lewis (2017) examined abnormal disclosure in the MD&A sections of annual reports between 1997 and 2008. In their study, they utilized latent Dirichlet allocation to match the content of each document to 75 different topics. By comparing the differences in topics found in fraudulent and truthful reports, they were able to distinguish between the two groups. They revealed verbal clues in fraudulent reports, which were associated with grandstanding the company performance without further mentioning the source of it. Furthermore, their results suggested a tendency among management to self-disassociate in the case of fraudulent behaviour, in line with evidence found by Li (2010a).

Similarly to Hoberg and Lewis (2017), Brown et al. (2018) applied latent Dirichlet allocation to analyse narratives from 10-Ks from 1994-2012 and manually coded 64 topics in their attempt to reveal what is disclosed in contrast to how it is disclosed. In addition to qualitative predictors, they utilized the F-Score from Dechow et al. (2011) to compare their detection results to a set of quantitative predictors, while also combining qualitative and quantitative ones to improve their results further. Robustness was tested by examining alternative topics, model specifications and texts from the MD&A section in contrast to the entire 10-K narratives. Their overall results, as reported by the pooled AUC values, indicated that quantitative predictors were inferior to qualitative counterparts, with a combined approach yielding the best detection quality.

3.3 Research Goals and Hypothesis Development

This study builds upon relevant previous literature by extracting textual predictors with high discriminatory power from fraudulent and truthful reports. In contrast to the existing approaches, the features are extended to a length of up to five subsequent words.[105] In this way, the investigation is more likely to capture complex textual patterns that are related to

[104] The feature vector comprises all predictors (features) utilized in the model. For further explanations, see section 4.1 and the related subsections. Given that this study builds upon a machine learning methodology, the term “feature” will mostly be used instead of “predictor” from here on.

[105] See section 4.1.3 for the generation of qualitative features.

the red flags associated with fraud, such as obfuscation theory, the self-serving attribution bias, or other clues in textual parts. To date, few studies have exploited multi-word phrases to detect fraudulent observations (e.g. Cecchini et al., 2010a). In addition, extensive preprocessing of the texts will be exercised by generating normalized terms, creating more comparable textual features between the reports, which according to Goel et al. (2010) leads to better detection results. To incorporate the results from the stream of literature relying on quantitative predictors and enriching the feature vector with additional financial information, this study additionally utilizes common metrics from financial statements.

When developing a comprehensive detection model, it is necessary to adopt a well-structured approach, going systematically through the process and relying on previous findings from the literature (Geerts, 2011). Thus, this study focuses on the shortcomings of previous work to overcome the sometimes seemingly arbitrary research design choices that may have hindered an understanding of fraud and fraud detection while scoring suboptimal detection performances. In taking different perspectives, this study does not seek to criticize other studies but rather to provide a new overall approach with additional insight. The general structure is divided into design questions and enhancing questions, with the goal of developing a sound detection model for future accounting fraud. Figure 10 presents the overview of questions and tasks that this study seeks to tackle.

Under the general research question focusing on the detection of future financial statement fraud through attaining qualitative and quantitative information from annual reports, the related research tasks are formulated as design and enhancing questions. The design questions, which can also be seen as problems that need to be addressed and answered, are based on the literature review and theoretical constructs of fraud with regard to a possible impact on fraud detection. In the following, the idea behind each question and the expected result will be laid out briefly.

Goal: Detection of future financial statement fraud exploiting qualitative and quantitative information from annual reports.

Design questions

1. Size of qualitative feature vector
2. Stability of qualitative features over time
3. Varying time gaps between training and holdout sample
4. Varying training and holdout set sizes
5. Detection rates over time
6. Feature vector performance
7. Matched and realistic sampling
8. Classifier performance

Enhancing questions

1. First instance of fraud detection
2. Quantitative feature vector enhancements
3. Cost-sensitive results

Figure 10 – Research goals

The first question deals with the rather static design upon which most studies have relied when constructing their qualitative feature vector. Given that the size of the vector reflects the number of textual clues used to distinguish between fraudulent and truthful reports, more features should result in better detection performance. Crime detection signal theory, as presented in section 2.4.1, suggests that the detector needs to be able to absorb a large amount of the signal, which can be very weak while maintaining its ability to distinguish between signal and noise. Therefore, although it is likely that feature vector size will not be arbitrarily big, because not every feature possesses the same discriminatory power and a larger number of features might result in a decrease of detection performance if conflicting features blur the distributional decisiveness of the vector, an optimal feature vector size for the problem has yet to be assessed (Goel et al., 2010, p. 40). The majority of studies have utilized a bag-of-word or similar approach to explain or examine the effect of different sizes on detection quality. For instance, Purda and Skillicorn (2015) used a fixed vector size of 200 features, Chen et al. (2017) relied on 240 different qualitative features and Goel et al. (2010) devised a total of 261,110 features. Cecchini et al. (2010a) have provided the only study to date to report detection performance changes for vectors between 100 and 1,200 features in different increments, showing that larger vectors actually result in better performance, marked by hypothesized saturation for large vectors and even a small drop in performance between the two biggest vectors (containing 500 and 1,200 features). In contrast to Cecchini et al. (2010a), this study utilizes larger textual patterns as well as a realistic probability of fraud approach in addition to the conventional matched sampling, which might influence the results to a certain extent. It can be hypothesized that similar results regarding the vector size and the aforementioned saturation effect will be observed.

H1: Larger qualitative feature vectors result in better detection performance.

The second design question also builds upon the bag-of-words-like design of most studies and the intertemporal stability of the underlying feature vectors. The predictors are based on qualitative features, which have usually been extracted for a single point in time or for a static timeframe. However, a potential problem may occur when utilizing predictors from narratives of corporate communication instruments like annual reports due to the constant change of content and language (Dyer, Lang, & Stice-Lawrence, 2017). Moreover, fraudulent behaviour may also change over time, resulting in different types of clues available and extractable from the reports (Zhou & Kapoor, 2011, p. 570).[106] To demonstrate the impact of altering disclosure on bag-of-words-like approaches, this study uses an initial sample of five rolling subsamples to demonstrate the changes to the bag-of-words covering different timeframes. The idea behind the rolling subsamples lies in the very general research goal, namely the detection of future fraud cases. An artificial future is generated by limiting the observations from which the patterns are extracted and models are trained to a specified date for each subsample while using future observations that are unknown to the models to assess the final performance of the models thereafter.[107] The separation in non-overlapping timeframes is necessary because the model set-up, which involves feature extraction and training, should not be carried out on data that is intended for testing purposes. The procedure should ensure the potential real-world application of the detection models, something that has rarely been the case in previous studies (e.g. Brown et al., 2018). The indicator "future" in Table 2 hints at the design philosophy of the study.

[106] See section 2.1.3 and the discussion about the evolution of fraud. Changing fraud scheme patterns over the course of time may influence the possibility of finding decisive clues in the reports.

[107] See Figure 11 for clarification of the sampling design.

The extent to which the bag-of-words actually differs is difficult to estimate beforehand. However, when utilizing design question 1 and assuming that smaller vectors result in worse performance, a considerable change in relevant clues over time may also cause the detection models to achieve unsatisfactory results. Hence, the detection models would need to be updated and tested regularly to deliver reliable results over time.

H2: Qualitative features are time-dependent.

Question 3 builds upon the first two questions by advancing the scope of the intertemporal stability of feature vectors. To shed light on the actual influence of composition changes in the qualitative feature vectors on the detection results, the initial five subsamples are used to detect future fraudulent observations with varying time gaps between the model set-up and the testing bed.

This is also tied to question 4, in which the results of questions 2 and 3 are briefly used to examine the influence of different-sized timeframes for model set-up and testing. As stated earlier, the model building process involves feature extraction and training, which is greatly influenced by the number of observations reserved to carry out this task, hence the number of years from which the observations are derived. The basic idea behind the procedure rests in the general bag-of-words approach and the sampling design. Creating an artificial future may be problematic when training and testing subsamples cover large timeframes, thereby increasing the time gap between the observations in both groups. Therefore, question 4 seeks to reveal the impact of time gaps and subsample length on detection results. To emphasize the subsequent implication of larger timeframes at a very basic analytical level, the following formulation is presented. A fixed bag-of-words of size F (number of features) captures the patterns of a certain number of cases n_t, with t denoting the period to which the cases refer over a timeframe of length x, ranging from one to multiple periods. Accordingly, formula one depicts the resulting imparity:

$$\frac{F}{\sum_{t=1}^{x+1} n_t} < \frac{F}{\sum_{t=1}^{x} n_t} \quad (1)$$

When increasing the length of the timeframes, there is less room for particular clues in the bag-of-words. Assuming that fraud is not a highly generalizable concept due to the sheer number of possible fraudulent schemes and cases, this should be taken into consideration. Especially when increasing the timeframe but keeping vector size constant, it is unlikely that patterns can be absorbed equally well. As Perols et al. (2017) have stated, different types of fraud require different sets of predictors to be successful. A larger timeframe will likely result in feature vectors being dominated by the most prevalent schemes without taking shifts towards certain schemes into consideration, which to a certain extent may be possible through smaller timeframes.[108]

H3: Time gaps between model creation and model application lead to deteriorating detection results.

Question 5 builds on the previous results to develop the final design and test the detection performance over the entire timeframe utilized in this study (1996–2010). Therefore, this study follows the design of Brown et al. (2018) with rolling subsamples over 15 years to test the reliability of the detection models. The indicator time series in the comparison Table 2 shows that few studies have adopted a similar approach and have examined performance at different points in time over a longer timeframe.

Questions 5 and 6 are answered in conjunction. In question 6, the three different feature vectors representing clues from the financials, from the textual parts and a combination of both, will be tested upon the final sampling approach developed up until question 5. In line with Goel and Gangolly (2012), this study follows the argumentation that linguistic features are better suited to detecting fraudulent behaviour, because quantitative features from company financials tend to conceal fraud and hence do not sufficiently differ from truthful years to yield good results. In general, it can be assumed that a larger feature vector (usually qualitative feature vectors are larger than quantitative ones) should score better results as it has the potential to cover more clues. Following this principle, it can be suggested that a combination of both would score even better results. So far, only few studies (Cecchini et al., 2010a; Dong et al., 2016; Brown et al., 2018) have used combined vectors to detect fraudulent cases, each supporting the assumption. What distinguishes this study from the first two is its design regarding the detection of future cases and the attempt to assess the detection performance over a longer timeframe, while also enlarging the textual patterns. Moreover, this study provides evidence for a realistic probability sampling approach, in addition to the matched sampling approach, which has been used by both studies. Brown et al. (2018) have studied fraud detection from a different angle. Their topic modelling approach captures potential clues at a more aggregated yet more understandable level. Regardless of the research design, all three studies have found similar evidence of qualitative

[108] With regard to the first design question, varying the feature vector size in conjunction with the number of years from which the observations for model building and application are derived can also lead to similar results. This study wants to hint at the influence of such basic design decisions on the outcome.

features being superior to quantitative ones and a combination of both outperforms the use of a single data type.

H4a: The qualitative feature vector achieves better detection results than the quantitative feature vector.

H4b: Combining qualitative and quantitative features improves detection results.

Fraud prediction represents a "needle in the haystack problem", whereby the minority group, here the fraudulent observations, occur considerably less often than the truthful ones (~1% of observations are fraudulent).[109] There are two general approaches to deal with imbalanced data sets.[110] For instance, Kirkos et al. (2007) or Cecchini et al. (2010a) have generated a balanced sample by matching each fraudulent observation with a truthful one. Despite the strict balancing through pair-matching, mildly balancing the sample is also possible. Lin, Hwang, and Becker (2003) or Chen et al. (2017) for example have relied on a matching approach in which each fraudulent observation is matched with several non-fraudulent ones, resulting in a mildly balanced sample. Another possibility is presented by Perols et al. (2017), who suggest different under- and oversampling approaches, which are typically used to adjust the balance between both groups and reveal the impact on detection results.

As an alternative to the aforementioned techniques, realistic probability of fraud sampling (in short, realistic sampling, in contrast to matched sampling) captures the actual probability of fraud in the sample by utilizing all available observations (e.g. Lee et al., 1999; Dechow et al., 2011; Brown et al., 2018). Some problems stemming from the use of a realistic probability of sampling approach pertain to the validity and the measurement of the results.[111] In the realistic sampling approach, detection, which is usually a basic binary classification, might be skewed towards specific predictors that discriminate between the classes (fraudulent and non-fraudulent) at a very basic level, like industry or size.[112] This may be caused by the patterns being dominated by size and industry variations; hence, the classifier would actually distinguish between different-sized companies in different

[109] Perols et al. (2017) suggested the "needle in the haystack" analogy in the context of financial statement fraud to emphasize on sampling difficulties and the different approaches to solve sampling issues.

[110] Table 2 reports the sample sizes for different studies. Large imbalances are usually associated with a realistic probability of fraud sampling, while equal sizes of both groups suggest a matched sampling design.

[111] Another deficit of realistic sampling approaches pertains to the considerably larger sample sizes involved, which hamper model building as feature extraction and model training times increase. However, technological advances to a certain extent have already rendered the argument invalid.

[112] Size has been identified as a significant predictor to distinguish between fraud and non-fraud observations in several studies (e.g. Persons, 1995; Beneish, 1999).

industries than actual fraudulent and truthful observations, while still scoring a satisfying result. Matching companies by size and industry is typically carried out to test whether predictors are significantly different relative to control firms (Dechow et al., 2011, p. 23). Matched sampling is performed by finding similar companies concerning a set of indicators, usually size and industry. Hence, the matched sampling approach may counter the aforementioned constraint. In this study, both sampling approaches are exploited, as the goal to develop a reliable and real-world applicable financial statement fraud detection model should include testing on both sampling approaches. Therefore, a robust detection model should successfully detect observations in both approaches, with a tendency to exhibit worse performance in the matched sampling context due to the fact that potentially decisive predictors are missing. This is the first study to rely on a realistic probability and a matched sampling approach to assess the performance of the detection models.

H5: The pair-matched sampling approach scores inferior results than the realistic sampling approach.

The second problem deals with the measurement and the interpretation of the detection performance in a realistic probability of fraud sampling compared to a matched sampling approach. Basically, not every performance measure is applicable in both scenarios. A detailed discussion will follow in 4.2.5.

The final design question then focuses on the performance of different classifiers. Studies so far have used a wide variety of classifiers. The results are varying considerably but it can be suggested that more sophisticated versions score better results (West & Bhattacharya, 2016). What distinguishes this study is the fact that each classifier is used for the detection of every feature vector across all subsamples and therefore the entire timeframe in both sampling approaches. So far, studies that have applied different classifiers have typically only focused on a single type of data (for example, financial ratios) but have scored different results across multiple classifiers (e.g. Fanning & Cogger, 1998; Lin et al., 2003; Gaganis, 2009; Abbasi et al., 2012). However, differences in feature vectors regarding size (qualitative feature vectors are usually larger than quantitative ones) and underlying data peculiarities resulting from different data types (ranging from traditional financial metrics to qualitative features) are considerable. Therefore, different classifiers are likely to score better on certain types of feature vectors (Sigletos, Paliouras, Spyropoulos, & Hatzopoulos, 2005). A comparison of classifiers across multiple feature vectors and sampling approaches can help to improve understanding of fraud detection on a technical level. Overall, the comprehensive overview provided by the eight initial design questions can enhance the general understanding of financial statement fraud detection in a time series setting and help develop reliable, robust, and efficient detection models.

H6a: More sophisticated classifiers achieve better detection performance.

H6b: Classifiers differ in their detection performance across feature vectors.

In addition to the eight initial design questions, which result in the first essential outcome of this study, three enhancements will be implemented to ensure the robustness and validity of the results, increase performance, and make the results more relatable from an economic point of view. The first enhancement deals with the first instance of fraud test, where schemes lasting over several periods are only taken with the first occurring instance into consideration (e.g. Brown et al., 2018).[113] The problem with schemes lasting over several periods is that this might result in the classifier distinguishing companies rather than actual fraudulent cases, due to the fact that each observation of a fraudulent scheme is regarded as an individual instance of fraud, as often seen in previous literature (e.g. the design of Persons, 1995; Cecchini et al., 2010b; Dikmen & Küçükkocaoğlu, 2010; Abbasi et al., 2012). Furthermore, it should be the goal of a real-world application to always detect the first occurrence of fraud. Through *ex post* examination, fraudulent schemes have often been detected several periods after their initial appearance, resulting in lengthy schemes. According to Summers and Sweeney (1998), it takes about three years on average before fraud is exposed. In this study's sample, the average scheme length is 2.6 years, with 322 (40%) of the 805 instances of fraud referring to a single period. Presumably, through the first instance of fraud detection variation, the results might be slightly decreasing, as the chances of getting a second try through a later instance of the same scheme are reduced.[114] Additionally, it is not possible for a firm with a persistent status (e.g. many subsequent fraudulent firm-year observations of the same company) to negatively influence the results and bias the learner towards that firm.

The second enhancement is associated with improvements to the baseline design of fraud detection models that rely on quantitative features extracted on an annual basis, as is most typical (e.g. Persons, 1995; Kaminski et al., 2004; Cecchini et al., 2010b; Abbasi et al., 2012). Abbasi et al. (2012) have offered several improvements to the traditional approach, especially taking additional contextual information like deviations from industry means or data from previous reports into consideration. By enriching the quantitative feature vector

[113] Perols et al. (2017) have gone even further by excluding all observations from a fraudulent firm (even the years of the firm that were non-fraudulent) but the very first instance of fraud.

[114] Although this is irrelevant for the detection with a single year holdout sample as in this study's final sample.

with additional information from previous years for every observation, this study controls for improvements to a certain extent. It can be assumed that the improvement results in better detection quality for the quantitative feature vector (Abbasi et al., 2012). Given that this study utilizes common quantitative features from previous research, the baseline results from the first eight design questions will not contain improvements to the quantitative feature vector. The examination of feature vector enhancements on a later stage then allows determining the level of improvement possible with the addition of enhanced quantitative features.

H7: Incorporating additional contextual quantitative features results in better detection performance.

While answering the design questions, the study is concerned with assessing the performance of the detection models. Similar studies have often relied on the area under the receiver operating characteristic (AUC) as the basic detection performance metric owing to the universal expressiveness of other characteristics (e.g. Abbasi et al., 2012; Brown et al., 2018).[115] In order to translate the rather abstract and generic performance measure into more relatable results, misclassification costs are introduced last, which is a common technique in the fraud detection context to evaluate the outcome using estimated cost ratios (e.g. Persons, 1995; Beneish, 1999; Abbasi et al., 2012). Studies have often only reported a single metric or a single set of metrics, limiting the comparability of the results and precluding the reporting of their economic benefits when not relying on a cost-sensitive design.

[115] A detailed discussion of different performance metrics and an explanation of the AUC is provided in section 4.2.5.

Quantitative features	**Fraud**	**Non-fraud**	**setting**	**Future**	**Time series**
Loebbecke et al., 1989	77	305	USA		
Persons, 1995	100	100	USA		
Hansen et al., 1996	77	305	USA		
Green & Choi, 1997	46	49	USA		
Fanning & Cogger, 1998	102	102	USA	yes	
Summers & Sweeney, 1998	51	51	USA		
Beneish, 1999	74	2,332	USA		
Lee et al., 1999	56	60,453	USA		
Bell & Carcello, 2010	77	305	USA		
Feroz et al., 2000	42	90	USA		
Spathis, 2002	38	38	Greece		
Spathis et al., 2002	38	38	Greece		
Lin et al., 2003	40	160	USA		
Kaminski et al., 2004	79	79	USA		
Kirkos et al., 2007	38	38	USA		
Gaganis, 2009	199	199	USA		
Lou & Wang, 2009	94	467	Taiwan		
Cecchini et al., 2010(b)	132	3,187	USA	yes	
Dikmen & Küçükkocaoğlu, 2010	17	109	Turkey	yes	
Dechow et al., 2011	293	79,358	USA	yes	
Ravisankar et al., 2011	101	101	China		
Alden et al., 2012	229	229	USA		
Abbasi et al., 2012	815	8,191	USA	yes	(yes)
Liu et al., 2015	138	160	China		
Perols et al., 2017	51	15,934	USA		
Qualitative features					
Goel et al., 2010	450	622	USA		
Glancy & Yadav, 2011	11	20	USA		
Humpherys et al., 2011	101	101	USA		
Purda & Skillicorn, 2015	1,407	4,708	USA		
Chen et al., 2017	45	135	China/Taiwan		
Qualitative and quantitative features combined					
Cecchini et al., 2010(a)	61	61	USA		
Dong et al., 2016	805	805	USA		
Brown et al., 2018	459	37,806	USA	yes	yes

The table reports a selection of relevant studies segmented according to their primary source of predictors, with additional information on sample size, setting, and test environment.

Future: where the design explicitly states that the observations of the holdout set are from future periods and feature extraction is limited to previous years

Time series: the results are assessed at different points in time in order to create a reliable detection model over the course of time

Table 2 – Comparison of similar studies

4 Methodology

In the upcoming sections, the methodology and the results will be presented. This study relies on the development of a comprehensive financial statement fraud detection model, utilizing techniques related to machine learning and textual analysis in the accounting research domain. Therefore, to discuss the relevant topics, the following effectuations start with the feature extraction and essentially depict how the narratives of annual reports are analysed to identify relevant textual patterns (which are referred to as qualitative features) as well as the financial metrics borrowed from the literature (described as quantitative features). Following the feature extraction and generation process, the machine learning framework utilized in this study is highlighted, focusing on validation techniques and the classifiers applied, before the results are presented. Thereafter, the structure of the presentation of the results is based on the research goals and the respective hypothesis. Lastly, the limitations of the research design are discussed and the results compared to similar studies.

4.1 Sampling and Feature Extraction

This section on sampling and feature extraction provides explanations regarding the data gathering process, especially focusing on the different sources of data, their peculiarities, and how they potentially influence the analysis. Afterwards, a detailed exposition of how the features of this study are extracted from the final sample is given.

4.1.1 Sample Composition

The underlying sampling process of this study is depicted in step 1 of Figure 11. The overall process is based on common text mining procedures (Feldman & Sanger, 2007, p. 15). The three primary sources are the EDGAR database,[116] the AAER archive[117] accessible through the SEC website and Compustat for supplementary company financials. Fraudulent financial statements are identified using the Accounting and Auditing Enforcement Releases (AAER) No 1-3810 issued by the SEC in the case of violations against the financial reporting requirements of the Securities Exchange Act of 1934. AAERs have been extensively studied in the literature in order to identify fraudulent observations (e.g. Persons, 1995; Green &

[116] EDGAR services are available at https://www.sec.gov/edgar/searchedgar/companysearch.html.

[117] AAERs are available at https://www.sec.gov/divisions/enforce/friactions.shtml.

Choi, 1997; Beasley et al., 2000; Cecchini et al., 2010b; Abbasi et al., 2012; Dong et al., 2016). The use of AAERs as proxies for fraud has potential downsides, which are rarely discussed but should be mentioned before the analysis. Dechow et al. (2011) have used the term “misstatement” rather than “fraud” when utilizing AAERs to identify observations of interest and has mentioned that although the SEC’s allegations often imply fraud, firms’ managers typically do not deny or admit guilt. Another limitation is the fact that the use of AAERs means relying on the SEC’s activity in the field of fraud detection and the publication of AAERs, which may be hindered for example by budget constraints or agenda setting and was in the past limited to publicly traded companies (Rezaee & Riley, 2009, pp. 270–272). Dyck, Morse, and Zingales (2013) and Dyck, Morse, and Zingales (2017) have estimated the number of fraudulent cases that go undetected: according to their model, 13% of US public companies engage in fraudulent actions in general and 7.3% in financial fraud in particular. When assessing the probability of fraud in this study’s sample, about 1% of observations are described as fraudulent, (from Table 3: 805/(805+84,960)≈0.009), potentially with a larger number of cases going undetected in the sample.[118] Dyck et al. (2017) furthermore suggest that SEC-related fraud proxies are highly biased by the SEC activity level. Karpoff, Koester, Lee, and Martin (2017) have compared four different sources and databases of financial statement misconduct research and ranked them based on four different features relevant for the purpose of quality research.[119] Samples based on AAERs are said to have the best scope (% of all enforcement actions by the SEC or the Department of Justice relating to financial misrepresentation under Section 13 (b) of the Securities and Exchange Act of 1934) while being the worst at timeliness (time lag to event date). Moreover, AAERs are suggested to be best suited at capturing most cases of fraud (this greatly depends on the definition of fraud) while exhibiting the second least amount of omissions. Karpoff et al. (2017) point out that the choice for either database is mostly depending on the research goals and should carefully be determined. As timeliness can be assumed to be rather unimportant for this type of study but coverage of cases of misconduct is important for the comprehensive identification of relevant cases, AAERs seem to be the best choice amongst the available, keeping the aforementioned limitations in mind.

[118] The fraction is similar to other studies relying on a realistic probability of fraud sampling.

[119] Karpoff et al. (2017) refer to the Center for Financial Reporting and Management as the primary source of AAERs, which also has been the underlying source for this study.

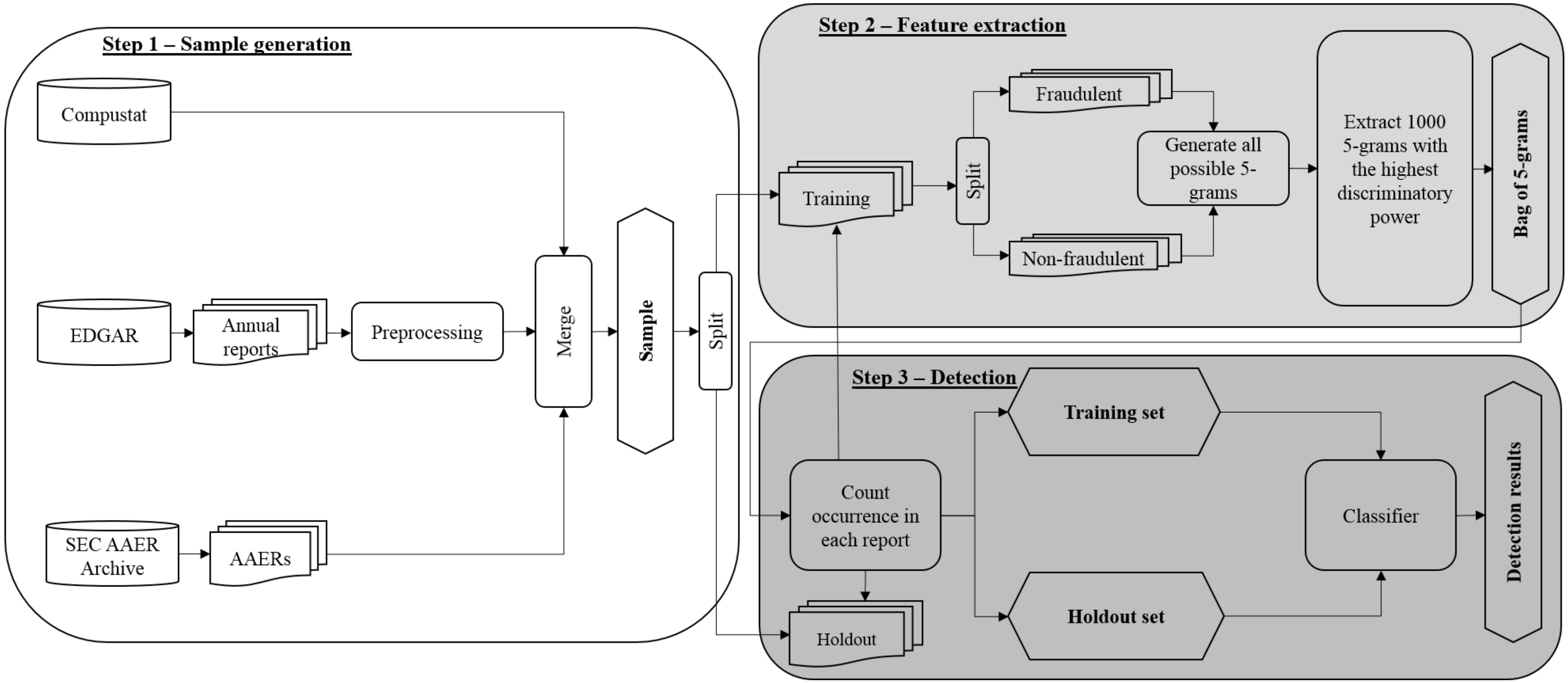


Figure 11 – Methodological overview

The three initial data sources provide the textual information as in the annual reports on Form 10-K, the additional financial information from Compustat and ensure the identification of fraudulent observations in the resulting sample through the examination of AAERs. In line with similar studies like Cecchini et al. (2010b), Goel et al. (2010) and Purda and Skillicorn (2015), a cross-sectional approach is applied. Concerning the financial industry, there is no universal way of dealing with firms from this branch. Although a number of studies exclude firms from the financial industry because of the SEC's special disclosure guides (e.g. Purda & Skillicorn, 2015) or unique topics discussed in the MD&A like market liquidity and capital structure (e.g. Hoberg & Lewis, 2017), this study follows Goel and Gangolly (2012) as firms from certain industries are not excluded. Detection performance might be negatively affected by this choice. However, the addition of a pair-matched approach in which observations are grouped by industry (besides size) controls for the potential impact. Table 1 shows the industry classification of the fraudulent observations.

	Fraudulent		**Non-fraudulent**	
Division	**Abs.**	**Perc.**	**Abs.**	**Perc.**
A	3	0.37%	265	0.31%
B	9	1.12%	3,298	3.88%
C	13	1.61%	929	1.09%
D	344	42.73%	31,607	37.20%
E	40	4.97%	8928	10.51%
F	34	4.22%	2929	3.45%
G	45	5.59%	4876	5.74%
H	118	14.66%	17,640	20.76%
I	199	24.72%	14,488	17.05%
Total	805	100.00%	84,960	100.00%

Division via standard industry classification

A	Agriculture, forestry, fishing
B	Mining
C	Construction
D	Manufacturing
E	Transportation, communications, electric, gas and sanitary services
F	Wholesale trade
G	Retail trade
H	Finance, insurance and real estate
I	Services

Table 3 – Industry classification

Years before 1996 are excluded because of the limited availability of annual reports in the EDGAR online database. Furthermore, firm years of 2011 onwards are removed, as

AAERs are released with a significant delay after the fraudulent action. The average time between opening an investigation and commencing an enforcement action was 21 months, with various cases lasting for several years.[120] This procedure is highly conservative and limits sample size considerably but avoids the inclusion of unidentified fraudulent firm-year observations in the sample. Reports without textual content are excluded from the examination.

The initial sample comprises 1,269 instances of fraud and 131,296 annual reports, which are further reduced to 805 fraudulent and 84,960 non-fraudulent observations after disregarding unidentifiable and unprocessable observations and merging with the Compustat universe. Unprocessable reports are mainly empty files and files including a header but no real content or corrupted files, which could not be processed in the automated approach. The terms "fraudulent" and "non-fraudulent" will be used hereafter in order to reflect the assignment of the observations based on the AAERs, limited to the scope of SEC enforcement actions and therefore not with all certainty resulting in the correct separation of truthful and fraudulent reports. Merging the annual reports with data from the Compustat universe leads to a considerable loss of observations, which is unfortunate but typical, as similar studies have shown (e.g. Griffin, 2003; Li & Ramesh, 2009).

Fraudulent	
Initial instances of fraud	1,269
With CIK	1,236
With processable annual reports from EDGAR	902
With total assets and SIC in Compustat	818
Matched*	805

Annual reports	
Initial annual reports from EDGAR	131,296
Processable annual reports	130,395
With total assets and SIC in Compustat	85,778
Non-fraudulent	84,960

*matched: total assets (95% confidence interval), sic (two digits), year (exact)

Table 4 – Sample size

In the next step, the firm-year matching for the alternative matched sampling approach is performed. In line with previous studies, pair-matching relies on total assets (95% confidence interval), industry (two-digit SIC codes) and year (exact) to assign the cases (e.g.

[120] SEC (2014, p. 54), Agency Financial Report – Fiscal Year 2014. Retrieved from https://www.sec.gov/about/secpar/secafr2014.pdf.

Beasley, 1996; Summers & Sweeney, 1998; Cecchini et al., 2010a). Congruent with Beasley (1996), a small number of cases could not be successfully matched and are therefore disregarded in the following.

The fraudulent cases are unevenly distributed over time with a peak in 2001, as can be seen in Figure 12. An explanation can be found in the enactment of the Sarbanes-Oxley Act from 2002 in the aftermath of the major fraud cases in the early 2000s (Goel & Gangolly, 2012, p. 79). The difference between all fraudulent cases and cases remaining in the examination does not seem to follow a time-dependent pattern.

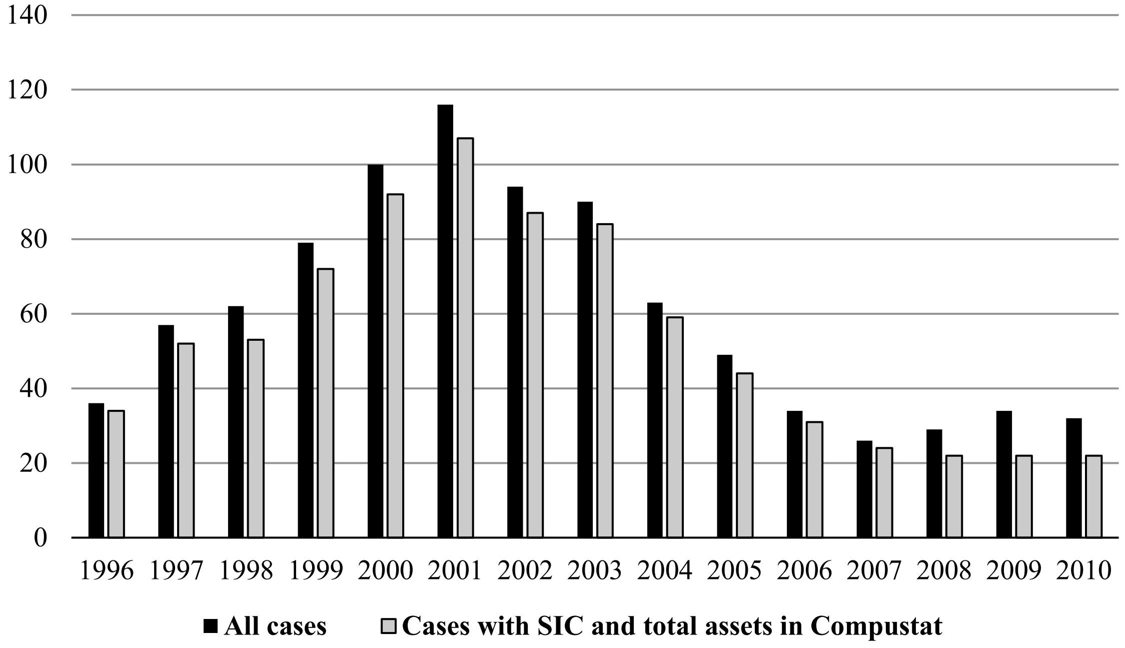


Figure 12 – Distribution of fraudulent cases

4.1.2 Textual Analysis

Before explaining how the qualitative features are extracted from the documents, this study's method in the greater scheme of text analytics is categorized. The amount of textual data produced and publicly available has increased immensely in recent years (Aggarwal & Zhai, 2012, pp. 2–4). Textual data must be distinguished from other forms of data like quantitative or relational data. It is characterized by its sparsity and high dimensionality (Aggarwal & Zhai, 2012, p. 3). When looking at a corpus, a collection of documents that one wants to analyse, dimensionality on a word-to-word basis is derived from each unique word in the corpus, even the words that only occur once (sparse). As this grows with corpus size, the document vector, which spans across all documents in the corpus and words in the document, grows rapidly. When strings of words, as opposed to single words, are taken into consideration, it accelerates even faster. Hence, the high dimensionality of its nature and the fast increasing amount of textual data, combined with technological advances in both hardware and software, have led to an emerging development of text analytic techniques (Aggarwal & Zhai, 2012).

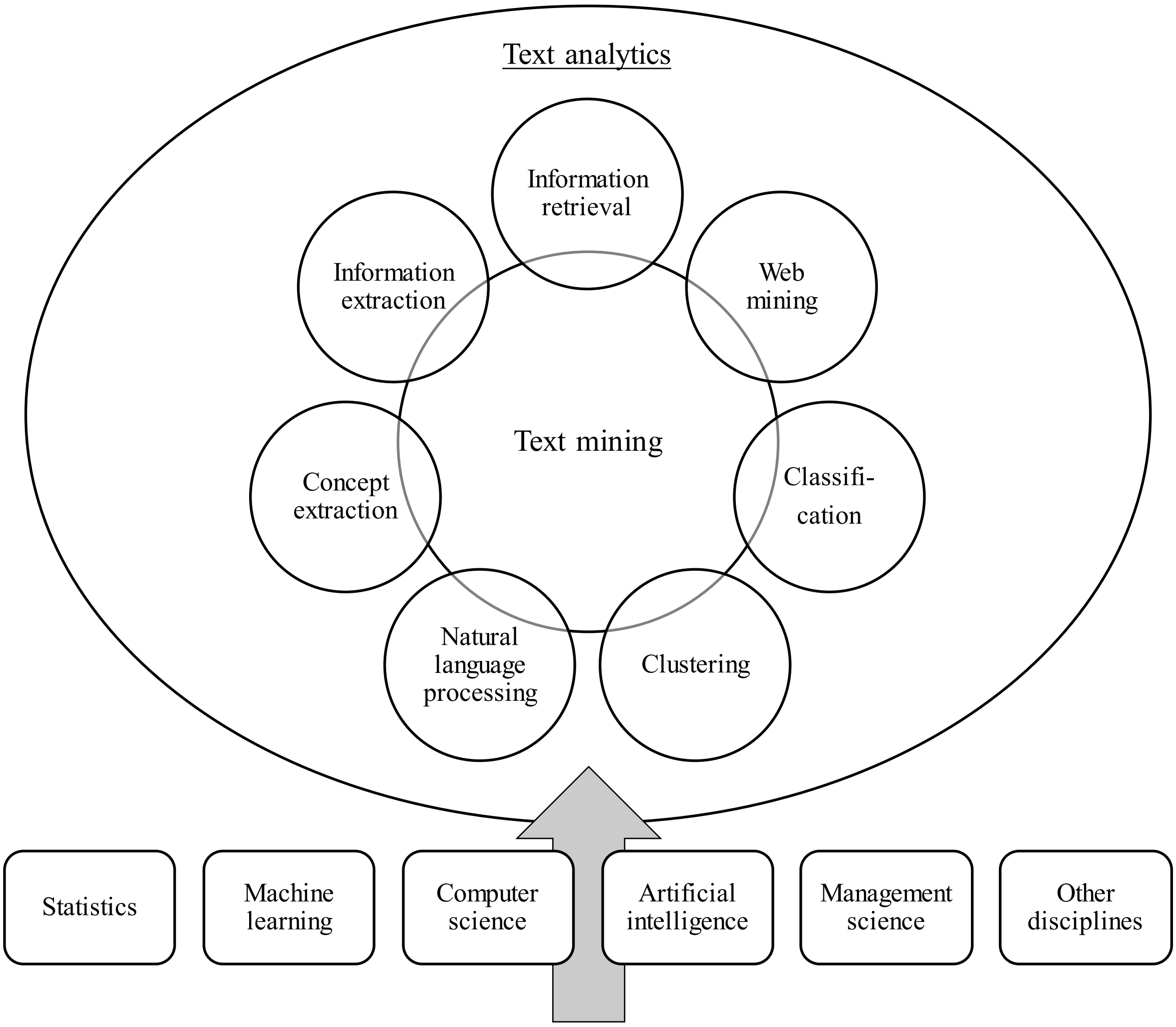


Figure 13 – Text mining[121]

Figure 13 is based on the text analytics overview of Miner et al. (2012). It depicts numerous related fields, which are of considerable importance in themselves but are connected through the basic concept of text mining. In their explanation of text mining, Aggarwal and Zhai (2012, pp. 4–8) mention the supportive nature of its purposes, like helping to digest and consume texts as well as analytical characteristics such as pattern recognition and the discovery of outliers. Miner et al. (2012) suggests that text mining can

[121] Referring to Miner et al. (2012, p. 40).

be seen as the practical application of text analysis techniques, ranging from rather descriptive techniques like information or concept extraction to more sophisticated approaches such as natural language processing, whereby machines are not only able to describe what was said but actually to understand the content and generate textual data like humans.

Machine learning and text mining or textual analytics are often mentioned in the same sentence. As illustrated in Figure 13, text mining is a highly interdisciplinary field that lives and develops with advances in other fields such as statistics and computer science With increasing interest in artificial intelligence and machine learning in recent years, the fields have become even more closely intertwined (Fisher et al., 2016, pp. 157–158). The nature of many text mining applications builds upon typical machine learning approaches of training and testing, for example when patterns from known texts are extracted and used to categorize unknown texts.

This interdisciplinary is also reflected in the text mining setting of this study, where an economic problem – the existence of fraudulently altered financial statements – is tackled by examining texts from annual reports in order to learn patterns to detect fraudulently altered statements of future periods. The detection of fraudulent statements can mostly be referred to as a document classification setting, whereby classification patterns are derived from the actual texts (Fisher et al., 2016, p. 164). For the purpose of the successful extraction of patterns, each document needs to be characterized based upon the results of the examination of its textual data. These examinations can be manifold, for example, as previously mentioned, focusing on re-occurring topics or word frequencies. The results of the examination describe each document, which can now be used to identify patterns based on predefined groups of documents, i.e. fraudulent and non-fraudulent ones. In this way, differences between the results of the examination of the textual data across groups are critical.

Cavnar and Trenkle (1994) have proposed an n-gram-based text categorization approach, where n-gram frequencies are exploited to classify texts to predefined categories. “An n-gram is an n-character slice of a longer string”, in their example referring to slices of characters from words (Cavnar & Trenkle, 1994, p. 162). Their approach is associated with supervised machine learning, as they trained a model in order to “learn” patterns of n-grams from training data to classify texts from a holdout sample in a setting where text labels are known (Feldman & Sanger, 2007, pp. 8–10).[122] The methodology of this study builds upon the very basic idea of their work, although here texts are analysed on a word level and longer strings are taken into consideration. Furthermore, the texts are preprocessed based on common preprocessing approaches for an n-gram text categorization (e.g. Goel & Gangolly, 2012; Vijayarani, Ilamathi, & Nithya, 2015).

4.1.3 Qualitative Features

In the following, the extraction of features from the qualitative and quantitative data of the annual reports will be discussed. The process covers steps 2 and 3 of Figure 11. In the literature, two basic streams on how linguistic predictors should be generated to detect fraud can be identified. A normative stream uses predefined predictors, based on informed reasoning and previous findings (e.g. Bell & Carcello, 2000; Humpherys et al., 2011). The second stream adopts adaptive and evolutionary techniques to identify discriminative patterns directly from the texts (e.g. Goel et al., 2010; Purda & Skillicorn, 2015). In order to avoid a normative design, this study follows the latter view that allows the updating of linguistic predictors and reacting to variations in the corporate narratives over time, which is exceptionally important for the general research design by which this study attempts to capture potential changes over the observed timeframe. Qualitative feature extraction is not limited to certain sections like MD&A, which is often subject to textual analysis for fraud detection (e.g. Dong et al., 2016). Choosing the MD&A as the only source of qualitative features comes with several problems as Loughran and McDonald (2016) argue. The first is the imperfect labelling of the MD&A section in a 10-K. Loughran and McDonald (2016) mention several tripwires in identifying MD&A sections, most obviously wrongfully labelled segments or the case of MD&A sections reported under exhibit 13. The second problem comes with the content that can be shifted between segments. Relevant content may be found in the footnotes or other segments, which would be excluded if only isolated parts are examined. For a study relying on an automated approach and a vast sample, manually looking for errors, irregularities, or peculiarities in certain documents is not possible. Therefore, an isolated examination of certain sections is not performed.

The following paragraphs will highlight the generation of the textual features, dividing the process into preprocessing of texts and feature extraction.

[122] The machine learning approach of this study is highlighted in section 4.2.

Preprocessing of Texts

All corporate narratives are consistently preprocessed following Grüning (2011). First, plain text is extracted, i.e. all layout features (bold printing, tables, or graphs) are disregarded. Words are stemmed using a dictionary-based stemmer (Kowalski & Maybury, 2000, p. 77) and the Automatically Generated Inflection Database (AGID) that contains the roots of about 280,000 inflected forms.[123] For example, "organisations" and "organisation's" are reduced to "organisation" (one of two British spelling variants) and "organizations" and "organization's" are reduced to "organization" (the unique American spelling variant); "reading" is stemmed to "read". Subsequently, British, American, and Canadian spelling variants are harmonized using the Variant Conversion Info (VarCon), a database that collects spelling variants for about 16,000 words.[124] All stop words are removed using the list of 319 stop words from the Information Retrieval Group of the University of Glasgow (Cowans, 2006, pp. 133–134).

Feature Extraction

Feature extraction is based upon the fundamental research goal of this study: the detection of future cases of financial statement fraud. The design is meant to train detection models based on known information and detect the cases of future periods. Observations for the model set-up are allocated to the training sample based on the respective period to which they are referring.[125] For each pre-treated annual report of the training sample, all unique n-grams of up to five words in length are collected from the harmonized texts. For instance, in a text, each word is considered an n-gram of one-word length. The first two words are an n-gram of two words length (bigram) and the second and third word from another bigram. The first three words are an example of a trigram; words two to four constitute a second trigram; and words three to five another one. Altogether, a text of x words contains $5x–10$ n-grams of up to five words length (x unigrams, $x–1$ bigrams, $x–2$ trigrams, $x–3$ four-grams, and $x–4$ five-grams).

The n-grams are regarded as unique independent of their word order (Grüning, 2011). For instance, the pre-treatment of the two original text passages "a financial statement analysis" and "the analysis of the financial statements" are harmonized to "financial statement analysis" and "analysis financial statement". They merely differ in their word order and are therefore considered identical. Subsequently, the term normalized n-gram is used. Using

[123] AGID is available at http://wordlist.aspell.net/other/.

[124] VarCon is available at http://wordlist.aspbell.net/varcon/.

[125] See section 3.3 for the basic research goals and Figure 26 for an explanatory breakdown of the subsamples.

normalized n-grams in this ways enables topics and terms to be identified in a text regardless of variations in grammar, enabling to identify more generalizable clues, which are not differing in slight variations. Without these normalizations, the terms ("the analysis of the financial statements" and "a financial statement analysis") are identified as separate n-grams. Assuming that both terms would be regarded as clues with high discriminatory power, both had to be considered in the qualitative feature vector, whereas the normalization results in a single feature (only "analysis financial statement" instead of the two initial terms). As studies so far (this one included) are restricted to limited bag-of-words/bag-of-n-grams sizes (number of clues, for example 100), considering both clues would block a spot in the vector for another clue.

The normalized n-grams of all fraudulent and non-fraudulent annual reports in the training sets are collected separately. This has been done for the realistic and the matched sampling approach individually. In the training sets, n-grams with the highest potential to distinguish between fraudulent and non-fraudulent observations are identified. The highest ranked n-grams are collected for each training set (for each subsample) to attain the respective features according to the reports in the subsample. The training observations are regarded as known cases up until a specific year, with observations from the following years representing the artificial future. After determining the top 1,000 qualitative features, their occurrence in each of the texts of the training and the holdout set is counted and scaled to text length. The result is a vector for each document that can be used to train the detection models and test them on the holdout set. For further explanations on feature vector size and n-gram collection, see sections 5.1.1 and 5.2.1, where the vector size implications of qualitative features, based on the collection of n-grams exploiting their inherent information value using an information gain ratio, are discussed.

4.1.4 Quantitative Features

The next section reviews the selection and generation of the quantitative predictors, also referred to as quantitative features. The variables are based on previous studies by Kinney and McDaniel (1989), Persons (1995), Fanning and Cogger (1998), Kaminski et al. (2004), Kirkos et al. (2007), Skousen et al. (2009) and Dechow et al. (2011). Thus, this study focuses on incorporating a large number of predictors, capturing fraud factors that have been identified in section 2.2 (especially 2.2.8) and fraudulent schemes as discussed in section 2.3, while maintaining a suitable sample size.

Persons (1995) has assumed that companies in financial distress are more likely to engage in fraudulent manipulations, creating her fraud detection model using 10 variables to measure financial leverage, profitability, asset composition, liquidity, capital turnover, size and overall financial position. Kaminski et al. (2004) have also measured financial distress, reusing eight of the variables from Persons (1995) and adding 13 new ones. They seek to capture the same dimensions but add different potential measurements. Fanning and Cogger (1998) have broadened the scope by adding measures for corporate governance, personal interrelations and auditor choice. They argue that management characteristics might provide relevant information for the prediction of fraudulent behaviour. However, they also assume that personal characteristics like potential gambling debts or criminal records may be discriminatory predictors even though they are hardly observable. In contrast to the previous studies, Dechow et al. (2011) have created their own set of predictors, focusing on different accrual based approaches, performance variables, nonfinancial measures, off-balance sheet activities and market-related incentives.

This study incorporates 16 of the variables from Kaminski et al. (2004), one from Kirkos et al. (2007) and complements the quantitative feature vector by adding two variables from Dechow et al. (2011). Some measures with very limited coverage in the Compustat database are disregarded. The final set of predictors can be found in Table 5.

No.	Name	Definition	Fraud factor
v1	accr	residuals from cross-sectional regressions	r
v2	btm	common equity / market value	p_ep
v3	logat	natural logarithm of total assets	o_os
v4	arta	receivables / total assets	o_ic
v5	invsal	inventory / sales	o_ic
v6	invta	inventory / total assets	o_ic
v7	nisal	net income / sales	p_fs
v8	nita	net income / total assets	p_ft
v9	opxsal	operating expenses / sales	p_fs
v10	opisal	operating income / sales	p_fs
v11	reta	retained earnings / total assets	p_ft
v12	salar	sales / receivables	p_fs
v13	salta	sales / total assets	p_fs
v14	into	costs of goods sold / inventory	o_ic
v15	tlta	total liabilities / total assets	p_ep
v16	cogssal	costs of goods sold / sales	o_ic
v17	de	total liabilities / total equity	p_ep
v18	gp	gross profit / sales	p_fs
v19	ietl	interest expenses / total liabilities	p_ep
Pressure:		financial stability (p_fs)	
		external pressure (p_ep)	
		personal financial need (p_pfn)	
		financial targets (p_ft)	
Opportunity:		industry considerations (o_ic)	
		monitoring (o_m)	
		organizational structure (o_os)	
Rationalization		rationalization (r)	

Table 5 – Variable definitions

Financial leverage (*tlta* and *de*) is assumed to be an indicator for financial fraud, as higher leverage is associated with a higher risk of covenant violations, which may instigate the management to engage in fraudulent behaviour in the case of potential debt covenant violations (e.g. Persons, 1995, p. 40). A similar argumentation can be brought forward for performance indicators (*reta* and *nita*). Covering diminishing performances by overstating revenues or understating expenses may be in the interest of the management to meet expectations. Alternatively, a window-dressing effect can be assumed (Kinney & McDaniel, 1989, p. 72). Furthermore, management compensation is often tied to stock prices, which may result in motivation for manipulations. The stock market performance is measured with the book-to-market ratio (*btm*).

Asset composition (*invta*, *nisal, opisal, invsal, arta* and *salar*) is of importance for fraud prediction purposes as receivables, sales and inventories are typically manipulated balance sheet positions, due owing to difficulties in auditing the positions caused by the option of subjective estimations (e.g. Persons, 1995, pp. 40–41; Kirkos et al., 2007, p. 998). In the light of manipulated sales numbers, there may also be an untruthful reporting of the associated costs of goods sold (*cogssal*, *into* and *gp*) as an overstated gross margin may yield incentives for the management (Dechow et al., 2011, p. 34). In this context, operating and interest expenses (*opxsal* and *ietl*) are also taken into consideration. Capital turnover (*salta*) captures the ability of the management to be competitive and generate sales with the invested capital (e.g. Persons, 1995, p. 41). Similar to the argumentation following the performance and the asset composition, the management might cover up their lack of competitiveness with fraudulently reported sales. From Dechow et al. (2011), the cross-sectional version of the modified Jones model (*accr)* is employed. They assume that the accrual component of earnings is the major subject to earnings management and potential manipulation. Accrual-based predictors have been revealed to yield good results (e.g. Dechow, Sloan, & Sweeney, 1995).

Besides the reliance on previous findings, this study also tries to bridge the gap to fraud theory. According to the Statement on Auditing Standards No. 99 – Consideration of Fraud in a Financial Statement Audit, the three factors of the fraud triangle (pressure, motivation, rationalization) can help as a guideline when trying to detect and deter fraud. The interdependences between them constitute the basis of the fraud triangle (Cressey, 1953) and their interwoven nature has often been proven subsequently (e.g. Skousen et al., 2009; Schuchter & Levi, 2016). Breaking down the three factors further into subordinate types in reference to the SAS 99 may help to identify proxies (Lou & Wang, 2009; Skousen et al., 2009). Therefore, capturing each of the factors (and of course as many subtypes as possible) by the predictors may help to improve detection performance. According to the findings of Huang, Lin, Chiu, and Yen (2017), the factors from the fraud triangle are not equally important or suitable in detecting fraud. Their results of a survey of experts suggest that pressure is the most important fraud factor, with opportunity coming in second place and rationalization depicting the least important factor.

Pressure can be separated into financial stability, external pressure, manager's personal financial situation, and financial targets. The factor can be mostly referred to predictors capturing performance and the financial situation at the company end. The manager's financial situation cannot be incorporated by exploiting the information available. Opportunity can be further segregated into industry-specific considerations, (ineffective) monitoring, and the organizational structure. Loebbecke, Eining, and Willingham (1989) already suggested that size as a proxy for organizational complexity represents a solid predictor for financial statement fraud, which may be due to the greater opportunity to conduct fraudulent schemes in larger corporations. Rationalization is hard to capture by variables from the company characteristics and its financials, but it may be possible to consider discretionary accruals as a proxy for management decision-making, which can be associated with factors of rationalization, like financial reporting rationalization. This study refers to Skousen et al. (2009) to assign the variables to the three factors of the fraud triangle, where possible.[126] To incorporate an extensive range of possible fraud predictors, variables are selected from as many factors and dimensions as possible and combined into the quantitative feature vector. The set of variables and the underlying calculations, as well as the factors of the fraud triangle they potentially capture, can be found in Table 5. In line with Huang et al. (2017) the predictors focus on pressure and opportunity to capture the most important factors in great detail. The descriptive statistics are reported in Table 6.

Variables from the company financials have proved able to predict fraudulent behaviour. However, when making use of fraud theories like the fraud triangle, these variables are rather poor at capturing the relevant factors (especially rationalization, see Skousen et al., 2009), which are said to be interwoven to a considerable degree. Going further and applying this to enhancements of the fraud triangle – or other fraud theories to a certain extent – the assignment would potentially be even more difficult. Therefore, adding predictors from other data sources is necessary, especially data on the managers or accountants who are in most cases responsible and/or involved in fraud schemes. Factors like capability from the fraud diamond, for example, could best be studied with personal information about the fraudster. Such personalized data are hard to come by but may significantly improve the detection systems (Gottschalk, 2018, pp. 38–41). Hence, enriching the pool of predictors to better capture all of the factors may considerably boost fraud detection likelihood. Therefore, this study incorporates predictors from the textual components of annual reports to overcome this problem to a certain extent.

[126] The assignment of predictors to certain factors of the fraud triangle is not exclusive. As a matter of simplicity, the most apparent factor has been chosen in table 5.

Variable	Mean			Stdev.		Fraudulent vs non-fraudulent	
	Fraudulent	Non-fraudulent	Difference in mean	Fraudulent	Non-fraudulent	p-value	t-statistics*
accr	-0.013	-0.013	0.001	0.178	0.205	0.943	-0.071
ietl	0.029	0.036	-0.007	0.024	0.032	0.000	7.473
salar	9.108	0.566	8.541	16.100	0.806	0.000	3.845
btm	0.549	11.374	-10.825	0.537	19.804	0.390	0.860
invsal	0.216	0.183	0.033	0.270	0.266	0.000	-3.687
into	0.126	0.106	0.020	0.151	0.148	0.001	-3.364
invta	0.116	-0.389	0.505	0.141	1.758	0.000	-4.312
arta	0.195	-0.276	0.471	0.164	1.605	0.374	0.890
nisal	0.625	0.275	0.351	0.334	0.666	0.000	-6.462
nita	0.375	1.186	-0.811	0.334	1.473	0.000	-6.227
opxsal	-0.153	0.725	-0.878	1.002	0.666	0.000	11.736
opisal	-0.029	0.200	-0.230	0.267	0.201	0.000	-10.931
salta	0.931	0.094	0.837	0.589	0.129	0.814	-0.235
cogssal	-0.005	-0.089	0.084	0.674	0.364	0.000	8.169
gp	0.936	0.930	0.006	0.722	0.839	0.000	-8.168
reta	-0.113	-0.908	0.795	1.096	3.234	0.000	-19.686
tlta	2.459	2.493	-0.034	4.302	4.725	0.000	7.685
de	0.537	0.613	-0.076	0.277	0.376	0.831	0.214
logat	20.636	19.534	1.102	2.010	2.202	0.000	-15.415

*The t-test is based on unequal variances.

accr: residuals from cross-sectional regressions, ietl: interest expenses / total liabilities, salar: sales / receivables, btm: common equity / market value, invsal: inventory / sales, into: costs of goods sold / inventory, invta: inventory / total assets, arta: receivables / total assets, nisal: net income / sales, nita: net income / total assets, opxsal: operating expenses / sales, opisal: operating income /sales, salta: sales / total assets, cogssal: costs of goods sold / sales, gp: gross profit / sales, reta: retained earnings / total assets, tlta: total liabilities / total assets, de: total liabilities / total equity, logat: natural logarithm of total assets

Table 6 – Descriptive statistics

4.2 The Machine Learning Methodology

In the following sections, the machine learning methodology of this study will be presented. After a rudimentary introduction to machine learning in general and for an application in the fraud detection context in particular, validation approaches and learning methods will be presented and the specific techniques relevant to this study explained and discussed. In the following, the four different classifiers that are utilized for the detection of financial statement fraud will be highlighted in detail. In this regard, for each classifier, the individual parameter optimization is carried out and the respective results that serve as the foundation of the in-depth analysis thereafter are presented.

4.2.1 Machine Learning in a Fraud Detection Context

Machine learning is an interdisciplinary field that combines aspects from statistics, information theory, game theory and optimization in a computer science environment (Shalev-Shwartz & Ben-David, 2014, pp. 19–21). Its primary goal rests in the training of algorithms embedded in computer programmes. Machine learning is a subfield of artificial intelligence, fulfilling tasks based on the fundamentals of human intelligence, as the recognition of patterns from extracted data. However, as Shalev-Shwartz and Ben-David (2014, pp. 24–25) note, in contrast to artificial intelligence, the focus of machine learning is not meant to imitate human intelligence and behaviour by machines but to utilize its abilities to fulfil specific tasks that might not be possible for human intelligence, such as those beyond human capabilities due to the amount or complexity of the underlying data (Goodfellow, Bengio, & Courville, 2016, pp. 1–8). Tasks that are suited for machine learning applications include classifications, regressions, transcriptions, machine translations, and anomaly detection, among others.

The methods are closely connected to statistics, but the purpose differs. Generally speaking, statistics are used to test hypotheses and hence check for assumed relationships, whereas machine learning provides an unformulated description of data. Some authors refer to machine learning as a form of applied math or statistics (Goodfellow et al., 2016, pp. 11–16). A basic example given by Shalev-Shwartz and Ben-David (2014, p. 25) depicts a physician who uses inferential statistics to test the influence of a particular indicator on a specific disease but then uses machine learning tools to extract patterns and predict diseases without predefining relationships and stating assumptions between indicators and diseases.

In this regard, machine learning has the character of a black box compared to statistics (Hastie, Tibshirani, & Friedman, 2016, pp. 351–352). Evaluating the outcome of machine learning applications also often diverges from statistics. The assessment of the outcome of machine learning techniques is usually based on the ability of the learner to absorb the provided set of data, for example, it is evaluated based on the achieved accuracy or error rate (Goodfellow et al., 2016, pp. 101–102).

In contrast, in statistics, the asymptotic theory assumes that sample sizes grow indefinitely and that the outcomes of tests are evaluated within this framework.[127] Another difference is based on the assumptions that statistical models require of their data, like distributional characteristics, linearity or independency assumptions. In machine learning, on a very general level, the learner figures out the representation of the underlying data without prescribing distributional and other characteristics beforehand.

With the multitude of statistical methods exploited for fraud detection tasks, an extensive body of literature has been developed, striving for good detection results. The approaches range from traditional and commonly applied logistic regressions (e.g. Beasley, 1996), discriminant analysis (e.g. Fanning & Cogger, 1998) or k-nearest neighbour classifiers (e.g. Liu et al., 2015), to rarely applied approaches in the accounting literature, like support vector machines (e.g. Cecchini et al., 2010a) or artificial neural networks (e.g. Green & Choi, 1997). More recently, machine learning algorithms have seen growing interest as more sophisticated approaches tend to yield better results (West & Bhattacharya, 2016).[128] Moreover, most studies do not only focus on one method but test several different tools (e.g. Fanning & Cogger, 1998; Humpherys et al., 2011; Abbasi et al., 2012; Liu et al., 2015). As discussed in design question 8, this study relies on a similar design and tests detection performance through exploiting several classifiers.

In this study, the term "detection" is commonly used when discussing the differentiation (classification) of fraudulent and truthful observations. Some studies alternatively refer to the term "prediction" (e.g. Skousen et al., 2009). Figure 14 depicts a fundamental concept in the domain of data mining and machine learning. Classification and prediction are related to each other and share similarities, although some fundamental differences in their forms of application exist (North, 2012, p. 9). Technically, in a classification setting, the outcome is a categorical class label, whereas a prediction models a continuous-valued function.

[127] This is not true for all models: for example, when using panel data, one dimension would be fixed as the others grow infinitely.

[128] For a comprehensive overview, see West and Bhattacharya (2016).

Classification is usually associated with supervised learning. Hence, the resulting class label can be interpreted as correct or incorrect because the true state of the new observation is actually known. The outcome of most studies in financial statement fraud detection is measured in accordance with the percentage of correctly classified observations, rather than falling into classification. Classification in an unsupervised setting results in a cluster analysis or association rules (Kotu & Deshpande, 2014, p. 167). Predictions, on the other hand, can generate an outcome for a new, unknown, and potentially future observation. For example, when using regression analysis to determine the factors that drive household income, the resulting regression function can be used to make predictions about the household income of an unknown observation, given its input data. The separation between both is important for the assessment of the outcome, which will be relevant in the upcoming sections and especially highlighted in 4.2.5.

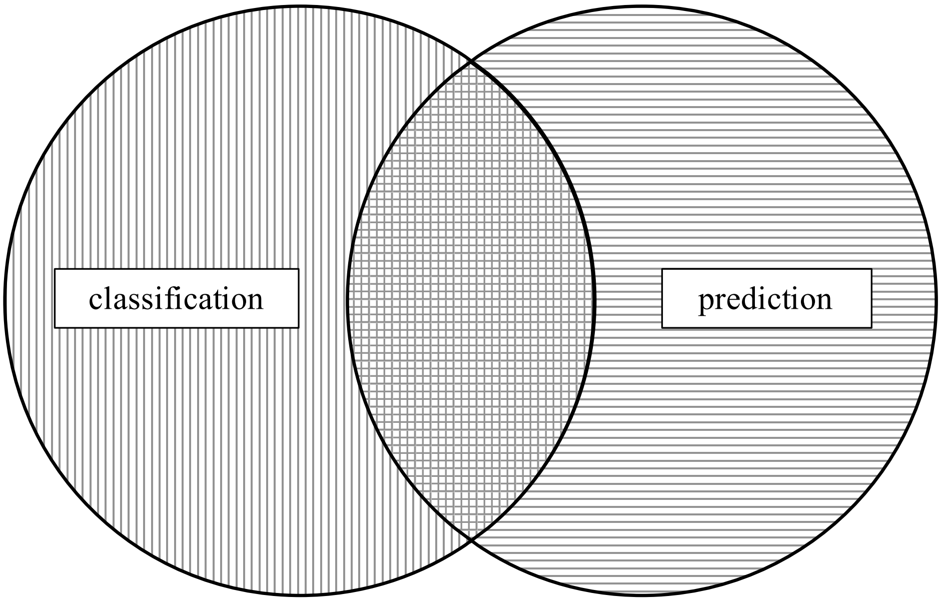


Figure 14 – Classification versus prediction

4.2.2 Validation

The need for validation approaches lies in the general procedure of machine learning tasks. If the entire data set is used for training, then performance can only be assessed on observations that the learner has already seen. The resultant risk is a potential lack of generalizability of the extracted patterns, as it is not tested on unseen data. Therefore, when comparing results such as error rates, one must differentiate between training error rate and test error rate, as the latter is usually of greater interest but may also be weaker than the one scored on already known data (James, Witten, Hastie, & Tibshirani, 2017, p. 37).[129]

Another problem that arises from the lack of validation in general or proper validation in particular, is the risk of undetected overfitting. A model overfits when the learner cannot capture the signal without capturing a major portion of the noise surrounding it, which would then again lead to a generalizability problem (James et al., 2017, p. 22). Overfitting also occurs, for example, when a model can fit almost any functional form due to the number of parameters and follows the data without abstracting generalizable rules. Even with validation approaches in place, overfitting can occur if the model performance is optimized and evaluated iteratively on a single validation set. In this case, the validation set would become a part of the training process. Hence, if not properly validated, overfitting cannot be tested for and ruled out (Theodoridis, 2015, pp. 91–93). Therefore, to assess the outcome of machine learning applications, a number of approaches have been developed to counter this and similar problems.

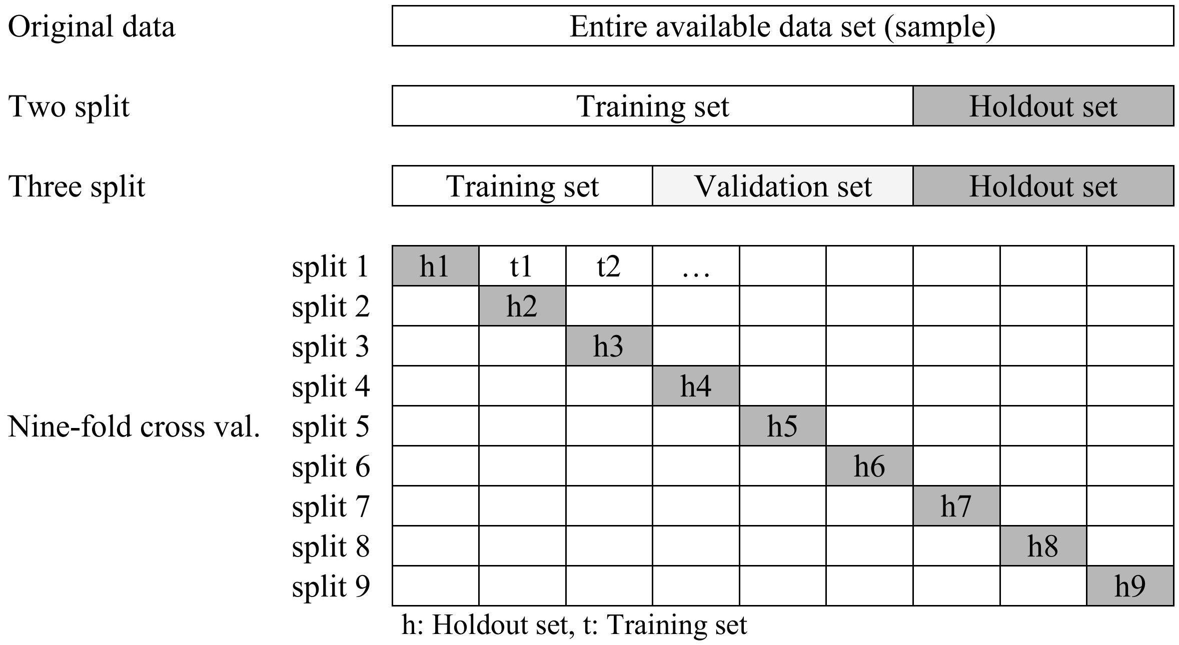


Figure 15 – Comparison of validation methods

Figure 15 depicts the most commonly used validation approaches, namely two split, three split and k-fold cross-validation (Shalev-Shwartz & Ben-David, 2014, pp. 146–150). Assume a given data set, for this example, called original data. In the two and three splits, the entire data set is split into several subsets, called training, validation, and holdout sets.

[129] Error rate is the fraction of wrongly classified observations.

The observations from the training set are used to obtain the model by training the learner. The holdout set is then the portion of observations used to assess the performance on unseen data. In the case of further model set-up steps or tuning steps, an additional subset is necessary to assess the performance of this iterative procedure before testing the final result on unseen data. The reasons have been discussed before. K-fold cross-validation partitions the original data into k subsets; in the explanatory split presented in Figure 15, k equals nine. The iterative procedure comprises k steps, each with training and validation of the model. For every step, the training happens on k-1 subsets and the validation on the remaining subset while iterating over the subsets, so that every subset is used for validation once (Theodoridis, 2015, pp. 92–93). One potential pitfall exists when drawing subsets from the entire dataset. The random drawing process may lead to distorted class distributions, no longer reflecting the original proportions. This problem can be overcome by stratification, under which the class distribution is kept according to the original data (sometimes the method is then called stratified k-fold cross-validation) (Witten, Frank, Hall, & Pal, 2017, pp. 167–168). The k-fold cross-validation has the advantage of reserving little data for validation purposes, which is especially important for scenarios where the number of observations is low and where holdout sets would distort the data (Shalev-Shwartz & Ben-David, 2014, p. 149). It is also possible to combine the different approaches, for example by using an initial two split validation approach, with model set-up and optimizing the training data using k-fold cross-validation before testing the final model on the unseen holdout subset.

In this study, a mixture of approaches is applied that is mainly relatable to the aforementioned combined validation, starting with 5-fold cross-validation for the model parameter set-up. The achieved parameter set-ups are then used for model building, which comprises feature extraction and model training on the training subset representing the known observations and assessing the performance on a holdout set, representing the unknown future cases, as time relevance is critical in the fraud detection context. This procedure is conducted several times on different subsamples in accordance with the respective sampling design for each design question, depicting the overall detection performance on several subsamples and for several points in time, avoiding the risk of overfitting onto a single subsample.[130]

[130] See 5.1.1 for the initial sampling design or 5.1.5 for the final sampling design.

4.2.3 Learning Methods

Two learning approaches form the basis of machine learning tasks, namely supervised and unsupervised learning, also referred to as "learning with or without a teacher" (Hastie et al., 2016, p. 485). The type of learning method that can be applied is primarily determined by the structure of the data (James et al., 2017, pp. 373–374). One can distinguish between response and predictor variables, which will be called labels and features in the following. In the scenario where every observation has a non-missing label, it is possible to utilize a supervised learning approach. In this case, the learner can relate the measurement of the features to the labels in order to predict the class affiliation for (other) observations. In simple terms, the student presents an answer to the teacher, who tells him or her if the answer is correct or not. If the data do not support the claim for non-missing labels, or known class affiliation is intentionally withheld, one has to fall back on an unsupervised learning approach. Without the labelling variable, building upon the relationship between the labels and the features is not possible and different evaluation methods are necessary. Some typical representatives are cluster analysis, association rules or self-organizing maps (Hastie et al., 2016, pp. 485–486).

The key difference between supervised and unsupervised learning lies in its adaptation. Supervised learning can be characterized as a density estimation problem, where one is interested in the properties of the conditional density of a label and a feature (joint probability density) (Hastie et al., 2016, p. 485). The success of the estimation can, for example, be measured by expected loss over the joint distribution. In contrast, unsupervised learning infers a probability density without the labelling variable, rendering the measurement of success less obvious. In other terms, supervised learning relies on the class label, is centred on the idea of finding relationships between labels and predictors and predicts the class affiliation correctly, whereas unsupervised learning aims to find and describe patterns in the features and the relationship between the data points themselves (Kotu & Deshpande, 2014, p. 167).

Between the two fundamental learning techniques exists a third option: semi-supervised learning. This represents a possible learning technique for tasks where not all examples are labelled (Barber, 2012, p. 306). In this environment, one could rely on supervised learning by eliminating the observations with missing labels or by falling back into unsupervised learning, ignoring the labels or selecting only unlabelled observations. Both possibilities would alter the initial sample to a certain extent, which may not be in the interest of the researcher. Semi-supervised learning can fix this problem by combining potentials from unsupervised learning like clustering with the ability of supervised learning to rely on correctly classified instances.[131] Common semi-supervised learning methods include self-training, generative models or graph-based algorithms (Chapelle et al., 2010, pp. 8–12). Unsupervised and semi-supervised learning is closely related to the learning structure of humans and animals and therefore may be better suited for machine learning tasks that aim to build upon human skills (LeCun, Bengio, & Hinton, 2015, p. 442). However, it is beyond the scope of this study to dive deeper into particular methods.

In the fraud detection context, the label is binary-coded and distinguishes between fraudulent and non-fraudulent reports. Observations with missing labels do not exist in the sample, as all observations that are not identified as fraudulent by the analysis of AAERs are regarded as non-fraudulent. Hence, this study is built upon a supervised learning approach, in which the main goal is centred on the classification performance and therefore the detection quality of fraudulent reports by features extracted from qualitative and quantitative data from annual reports.

4.2.4 Model Selection and Hyperparameter Optimization

After discussing the different validation approaches and the learning techniques in the previous sections, this section combines the implications for this study's setting in the final model building process. Before training the models and testing their detection performance on the holdout sets, as described in Figure 11, the classifiers need to be established. In the following sections, the four classifiers will be highlighted and the hyper-parameter set-up discussed.

The goal of hyperparameter tuning or optimization is to adjust the effective capacity of the model to match the complexity of the task (Hutter, Kotthoff, & Vanschoren, 2019, pp. 3–4). Usually, classifiers such as k-nearest neighbours (KNN) require the *a priori* determination of several parameters, also referred to as "hyperparameters" (Witten et al., 2017, p. 171). In the example of KNN, one would have to decide how to measure the distance to the respective neighbours and how many neighbours (k) are taken into consideration when classifying an observation (Hassanat, Abbadi, Altarawneh, & Alhasanat, 2014).

[131] If it is the intention to treat the sample as if it would not have any missing labels, semi-supervised learning is closer to supervised learning, due to the similar goal of correctly predicting labels (Chapelle, Schölkopf, & Zien, 2010, pp. 4–11).

Classification quality depends on the determination of these parameters. However, to assess the performance of different parameter combinations, an additional holdout data set is necessary. Choosing one of the data sets this study insists for later usage would cause the optimization to be centred on this data set, leading to an optimized detection result for this very subsample and potential overfitting onto the underlying observations (Witten et al., 2017, pp. 171–172). Avoiding overfitting on a single subsample is important in order to achieve robust results on other subsamples, which is even more relevant for the research design in this study, where changes to the underlying features over 15 years are assumed.[132]

To solve this problem, a random data set of observations from the entire timeframe is generated.[133] This data set is then used to train models using different hyper-parameter combinations and is validated with 5-fold cross-validation.[134] In detail, the parameter set-up is carried out for each of the feature vectors (quantitative only, qualitative only, quantitative, and qualitative combined). The three feature vectors are similar by design and thus purpose, however fundamentally different in terms of features and requirements for a good classifier. Optimizing the parameters using only one of the aforementioned feature vectors and exploiting the resulting hyper-parameters for the other feature vectors would very likely cause a bias. The feature vector, on which the parameters would have been optimized, would most likely outperform the other two feature vectors. Thus, optimizing the hyper-parameters for each of the feature vectors in a similar manner is necessary. Given that this implies an extensive optimization task, the parameters are only partly optimized. A more in-depth optimization approach for each subset lies beyond the scope of this study and will most likely not change the overall tendency of the results.

When dealing with unbalanced data sets, applying resampling methods on the initial data can lead to an unintended impact on class distribution. With the 5-fold cross-validation for parameter set-up purposes, a random divide of the entire data set into the subsets could lead to said bias. To keep the class distribution similar in the subsets, this study relies on a

[132] For clarification, the sample represents all observations covering a period of 15 years. Subsamples represent a part of the entire sample, in this study typically covering a continuous part of the timeframe. Each subsample can furthermore be split into sets, typically training and holdout sets. See Figure 25 for the initial sample, for example.

[133] All randomizing processes in this study are based on a local random seed of 1992 to ensure replicability and comparability. Seeds are used to initialize pseudo-random number generators (Kotu & Deshpande, 2014, p. 82).

[134] Witten, Frank, Hall, and Pal (2017, p. 172) mention the computational extensive nature of hyperparameter optimization using grids under a cross-validation approach. In these cases, they suggest the common procedure of falling back to a lower amount of splits. Under similar constraints, this study desist from using the more common 10-fold cross-validation and relies on a 5-fold cross-validation.

proportionate random stratified sampling algorithm. Proportionate random stratified sampling divides the entire data set into homogeneous subsets based on the class distribution of the initial data set (Kotu & Deshpande, 2014, p. 81).

For optimization purposes, traditional, easy-to-understand yet comprehensive optimization grids are utilized (Hutter, Lücke, & Schmidt-Thieme, 2015, p. 330). These typically two-dimensional grids illustrate the classification results for different combinations of parameters. The downside of the grid search rests in its computational effort, as all parameter combinations have to be tested and are seen as equally important, even if some parameters or their combinations may be of only minor relevance (Bergstra & Bengio, 2012, p. 302). This study does not exploit predefined optimization algorithms to ensure the replicability and understandability of the upcoming presentation of the results. Recent literature advocates the use of tuning algorithms or even random parameter tuning, which might outperform the manual extensive grid search but is especially useful for complex classifiers with many parameters to tune (Hutter et al., 2015, p. 330). The difference between a pair-matched sample with only quantitative features and a realistic probability of fraud sample with the combined feature vector regarding size and complexity is very large. The test for optimal parameter combinations is extensive, as it must be conducted for each feature vector and sampling approach. The upcoming sections deliver a basic understanding of the classifiers as well as their peculiarities, which had to be considered in this study's scenario. In this context, the relevant parameters are discussed and the results of the initial parameter tuning using the grid approach presented.

After the optimization process on the random subsample, the classifiers are ready to be deployed in the actual detection setting. It is important to mention that classifier optimization via hyper-parameter configuration is undertaken entirely beforehand to avoid overfitting and skewing the final detection results. Figure 16 depicts the final detection process, with its learning and testing phase.

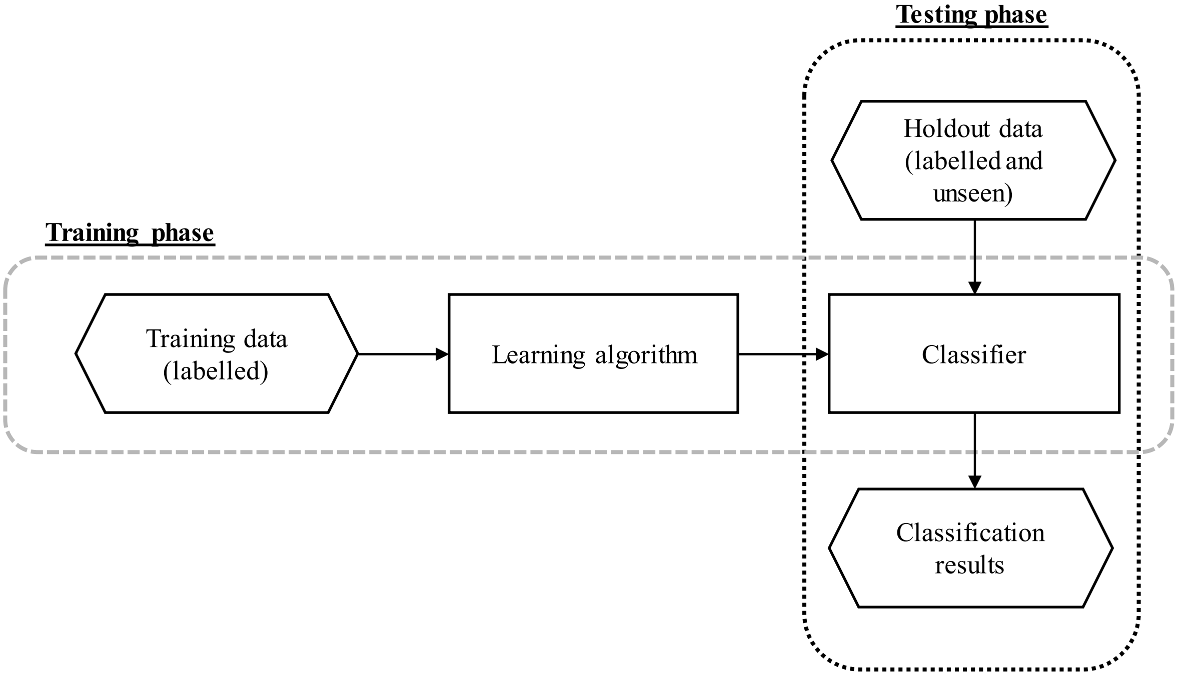


Figure 16 – Basic machine learning process

The observations of the entire sample are assigned to training and holdout sets, constructing the subsamples that cover the different timeframes. In the following, the qualitative features are extracted from the training set. This step is repeated for each subsample. The differentiation between the training and the holdout set is crucial for this kind of methodology, as the observations in the training set represent the universe of potential qualitative features based on known cases. The class affiliation (fraudulent/non-fraudulent) of the observations in the holdout set is assumed to be unknown at the time of the model training. Hence, the observations of the holdout set are not utilized for feature extraction. For example, for the initial sample as presented in Figure 25, the features for i1 are extracted from the 49,903 observations of the training set from 1996 until 2003. In the next step, the training set is used to allow the algorithms to learn patterns based on the extracted features. Finally, the trained models classify the observations from the holdout set and the quality of the classification can be assessed. In the example for subsample i1, the 15,375 observations of the corresponding holdout set are from the years 2004–2006. This procedure is carried out separately for all subsamples throughout the entire study.

4.2.5 Measuring Detection Performance

To measure the detection performance of the models, the area under the receiver operating characteristic (AUC) is utilized in this study. The AUC is a commonly applied metric to evaluate the performance of fraud detection models, especially in studies with unbalanced samples (e.g. Fawcett, 1997; Purda & Skillicorn, 2015; Brown et al., 2018). However, the reporting of results, especially in terms of the measurement of detection performance, has been carried out in diverse ways and many different performance metrics can be found in the literature. In addition to the AUC, studies have typically reported accuracy (fraction of correctly classified observations) and/or recall (fraction of correctly classified fraudulent observations), which might be sufficient to assess the classification results for balanced samples. However, with the severely unbalanced data sets, which are common in the fraud detection context (around 1%–2% of observations are fraudulent in this and in similar studies relying on a realistic sampling approach), accuracy alone would not be enough, as a high level of accuracy can be achieved without correctly classifying the important but underrepresented fraud class (Ben-Hur & Weston, 2010, p. 234). Reporting both values may solve part of the problem but still has the limitation that observations, which are wrongly classified as fraudulent (type 1 error) are not taken into consideration. For example, if one suggests that all observations that are classified as fraudulent should be investigated in detail, implying examination costs, the number of false positives is of interest as well.[135]

The receiver operating characteristic (ROC) was developed during World War II to assess the performance of radar-guided signal detection and to find an optimal trade-off between hits and false alarms (Pepe, 2000, pp. 308–309). Under the crime signal detection theory, the ROC has also been used extensively to evaluate the performance of models in a similar fashion and to determine ramifications with the best signal-to-noise ratios (Phillips et al., 2001, 296-297). In general, the ROC can be seen as the amalgamation of the true positive rate (*TPR:* fraction of correctly classified fraudulent cases, formula 29) and the false positive rate (*FPR:* fraction of non-fraudulent observations classified as fraudulent, formula 30) of varying discrimination thresholds.

$$TPR = \frac{TP}{TP + FN} \tag{29}$$

[135] See section 5.2.4 for an estimation and utilization of the costs of misclassification.

$$FPR = \frac{FP}{TN + FP} \tag{30}$$

The decision to classify an observation either as positive or negative is based upon the outcome X of the classification function obtained through training a classifier upon a set of observations. X_i represents the decision value of the i-th observation after training the classifier and applying the resultant classification vector to the data of the i-th observation. The threshold parameter T depicts the decision boundary, for which $X_i > T$ is classified to the positive group and $X_i < T$ to the negative group, respectively. X follows a probability density function $f_1(x)$ for true positives and $f_0(x)$ for false positives. Over varying thresholds of T, the TPR can be plotted against the FPR, resulting in the receiver-operating characteristic. The AUC is then, as the name suggests, the area under the curve of the ROC between 0 and 1.

$$TPR(T) = \int_{T}^{\infty} f_1(x)dx \tag{31}$$

$$FPR(T) = \int_{T}^{\infty} f_0(x)dx \tag{32}$$

$$AUC = \int\limits_{x=0}^{1} TPR\big(FPR^{-1}(x)\big)dx \tag{33}$$

To clarify, each discriminatory threshold is a possible result with a respective confusion matrix and associated metrics like accuracy or recall. This implies that when comparing single metrics like recall or accuracy to other studies, the threshold maximizing the respective metrics may be selected, rendering the comparison arbitrary. The AUC is calculated from all possible thresholds, improving the comparability of the results. Graphically, with the true positive rate on the y-axis and the false positive on the x-axis, the best prediction results can be found in the upper-left corner, where the true positive rate is high and the false positive rate is low. Technically, it ranges between 0 and 1, with higher values indicating better detection quality and a value of 0.5 suggesting a random classification process. The AUC is also frequently used in other disciplines, such as in medicine and pharmacy to evaluate diagnostic tests (e.g. Mandrekar, 2010; Scheff, Almon, Dubois, Jusko, & Androulakis, 2011) and in meteorology to assess forecasting quality (e.g. Mason & Graham, 2002). Translating the AUC values into generally accessible terms can become rather subjective due to the imprecision of language and contrasting perceptions regarding certain quality levels in different fields of application. According to Hosmer, Lemeshow, and Sturdivant (2013, pp. 173–181), values larger than 0.8 depict excellent detection quality, although quality should always be seen in comparison with similar applications.

However, the AUC is not flawless in its ability to assess classification outcomes. Lobo, Jiménez-Valverde, and Real (2008) have examined the disadvantages of using this single value performance metric. The AUC is a measure of discriminatory power, not model fit, because it does not take predicted probabilities into consideration. Indeed, a well-fitted model may have low discriminatory power and vice versa, although this is not a significant issue in this study's application of the AUC because it primarily focuses on the former. Another downside lies in the calculation of the AUC across the entire interval of the FPR. For most applications, the marginal areas that represent extremely low or high FPRs are rather uninteresting. The optimal thresholds usually lie in the upper-left corner, where the FPR is low and the TPR is high. However, the AUC is calculated across all thresholds. In this regard, the AUC also does not weight the classification errors according to any sort of theoretically or empirically derived costs that may be associated with type 1 and type 2 errors. When searching for a solution that minimizes the associated costs, only one threshold is of interest.[136] Figure 17 depicts an explanatory AUC of one of the detection models of 0.861 and references the AUCs of random classifiers (0.5) and an outstandingly well-performing classifier (0.99).

[136] On the other hand, determining the costs of different errors can be subjective or the relationship between the costs may change over time, making this process difficult to determine beforehand.

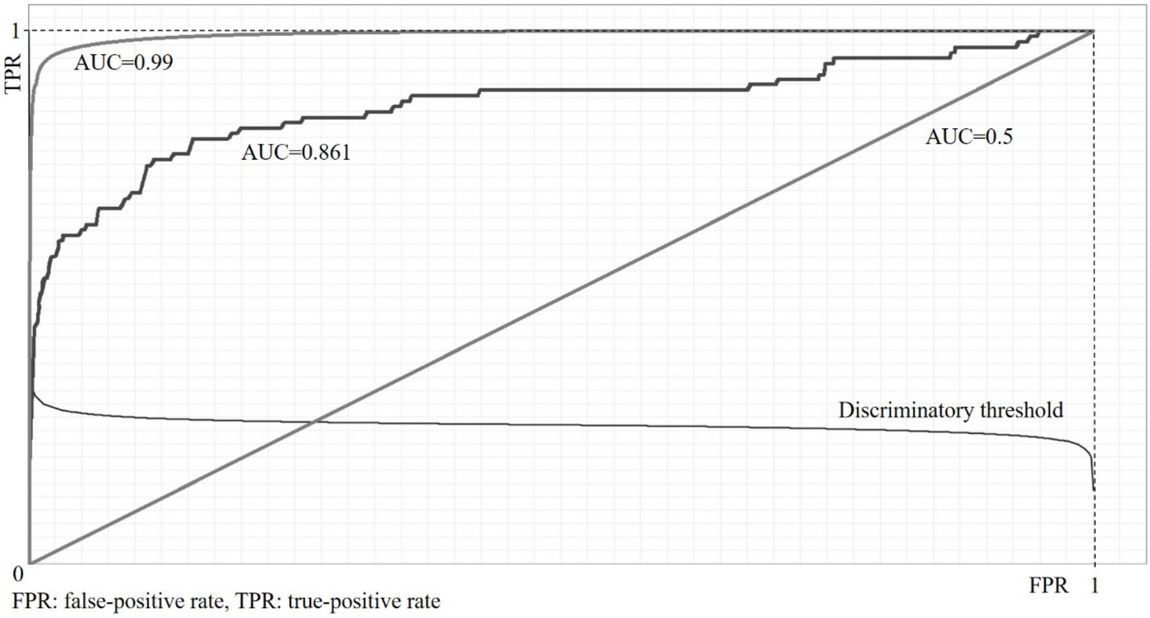


Figure 17 – Explanatory AUC

Despite the aforementioned commonly applied detection performance metrics, some studies utilize different or additional measures, like a costs-of-misclassification approach (e.g. Perols et al., 2017). Cost-sensitive results weigh the classification results in accordance with predefined costs associated with specific outcomes. For example, a fraudulent case that is not correctly classified (not detected) would be weighted with the implied costs that might arise through its further existence; on the other hand, a non-fraudulent observation that is wrongly classified would be weighted with the already mentioned examination costs. Determining the costs is vital to the results of the costs of the misclassification approach, rendering it a certain degree of freedom (subjectivity). This study strives to use a single metric applicable to balanced and unbalanced samples alike and that is free from subjectivity and comparable to other studies. However, to shed light on the potential economic benefits of the developed detection models of this study, an additional cost-sensitive result will be presented lastly.

4.2.6 Naïve Bayes

The first classifier is the Naïve Bayes classification, which is one of the oldest and most popular techniques in the area of text classification (McCallum & Nigam, 1998, p. 41). The Bayes theorem, published in 1763 by Thomas Bayes, represented an important and influential work in statistics and probability theory (Kotu & Deshpande, 2014, p. 113).[137] The basic idea behind a classification algorithm is the assignment of a class label to an observation with the help of variables describing each observation. Naïve Bayes is rooted in probabilistic theory, making use of evidence to predict the most likely outcome. The Naïve Bayes classifier assumes that regarding the label (class), the value of a feature (variable) is independent to the value of the other features (variables) in the data set (Barber, 2012, pp. 9–10). Hence, correlations between features are not taken into consideration. This is referred to as class conditional independence.

Let X be a set of variables (features), $X = \{X_1, X_2, \dots, X_n\}$ and Y the outcome, $Y = \{fraudulent, truthful\}$. X_i is an individual feature, for example, the (absolute) frequency of a single n-gram. $P(Y)$ denotes the probability of the possible outcomes from the underlying data, for example, the probability of the observations to be fraudulent; analogously $P(X)$ denotes the respective probability of the inputs of X. Given a specific input for X, $P(X|Y)$ is the probability that we would observe the input X given the class label Y, also known as class conditional probability. $P(Y|X)$ is then the probability of an outcome Y (to be fraudulent or non-fraudulent), given an input X. Under the Bayes theorem, the probability is calculated as follows:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} = \frac{P(Y)\prod_{i=1}^{n} P(X_i|Y)}{P(X)} \tag{2}$$

The classification boundary is simply determined by the probability of a new vector (observation) X^* belonging to either of the classes. For example, $Y = fraud$ if:

$$P(Y = fraud|X^*) > p(Y = truthful|X^*) \tag{3}$$

Translated to the fraud detection problem, at the end of the process depicted by Figure 11, each observation is labelled according to its class affiliation Y as either fraudulent ($Y = 1$) or non-fraudulent ($Y = 0$). For each observation j, a d-dimensional attribute (feature) vector X^j is created, with a maximum size of 1,019 features (1,000 features based on textual data and 19 based on numeric financial data). Attribute characteristics can either be binary, multi-state or continuous (Barber, 2012, p. 241). This study relies on continuous variables for the qualitative and quantitative features. For each n-gram, absolute and relative

[137] It was published posthumously as "Essay towards solving a problem in the doctrine of chances" in the Philosophical Transactions of the Royal Society of London (Bayes & Price, 1763).

frequencies of occurrence for the respective report (feature vector) are computed.[138] However, continuous variables require further transformation. There exist two possibilities for the transformation purpose: either by binning, which translates the variables into categorical variables or by using distributional functions (kernel functions) (John & Langley, 1995, p. 339). The categorical solution is rather unsuitable because of the sheer number of values that the relative frequencies can take. Therefore, the distributional modelling relying on the commonly adopted Gaussian distribution function (4) is chosen (Witten et al., 2017, pp. 100–103).

Gaussian distribution function

$$P(X|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{4}$$

with mean

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{5}$$

and variance

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\mu)^2. \tag{6}$$

With the Gaussian kernel, the parameters that need to be estimated from the training data are the mean and the standard deviation for each class and variable, as depicted by formulas 8 and 9 (John & Langley, 1995, pp. 339–340).

$$\delta() = 1\ if\ Y^j = y_k, else\ 0 \tag{7}$$

$$\hat{\mu}_{i,k}^2 = \frac{1}{\sum_j \delta(Y^j = y_k)}\sum_j X_i^j \delta(Y^j = y_k) \tag{8}$$

$$\hat{\sigma}_{i,k}^2 = \frac{1}{\sum_j \delta(Y^j = y_k)}\sum_j (X_i^j - \hat{\mu}_{i,k})^2 \delta(Y^j = y_k) \tag{9}$$

The advantages of this rather simplistic classifier are fast training and relatively small storage requirements (Barber, 2012). In contrast to other classifiers that rely on iterative

[138] Relative frequencies of occurrence are used, because binary coded occurrences of n-grams could result in a bias, as larger companies tend to have larger reports, resulting in a higher possibility of including a greater fraction of the total qualitative features. This also solves potential pitfalls like the 1-of-m coding, resulting in highly dependent features (Barber, 2012, p. 169).

procedures, the training of this probabilistic classifier solves closed-form expression, resulting in good scalability. Under the theorem of "simple first", Naïve Bayes is a good starting point for any classification scenario and is known for its ability to keep up with more sophisticated approaches, depending on the data set (Witten et al., 2017, p. 105). However, Naïve Bayes starts to lose its competitiveness on data sets with suboptimal feature characteristics, especially when compatible features are part of the data and the redundancy of similar features skews the results towards these features (Witten et al., 2017, p. 105). In the fraud detection scenario, particularly the qualitative features are assumed to occur in some sort of pattern and would in this case not be independent and might even be highly correlated.[139] This also applies to quantitative features. Hence, the Naïve Bayes classifier can be expected to perform more poorly than the other approaches.

4.2.7 K-Nearest Neighbour

From a broader angle, the learning phase for k-nearest neighbour (KNN) approaches can be described as creating a lookup table and classify unknown observations using similar observations (Kotu & Deshpande, 2014, p. 99). This approach is referred to as a "lazy learner". In more technical terms, k-nearest neighbour builds upon an n-dimensional space, constructed by the features of the underlying feature vector, with each observation representing one point in said space and being labelled according to the supervised learning strategy.

K-nearest neighbour is a non-parametric approach. Therefore, no assumptions of the distributional characteristics of the data are made (Peterson, 2009, p. 1883). It can be used for the classification of bi- or multinomial class labels. The algorithm uses a predefined number of nearest neighbours (k) of an observation to determine its class membership by majority vote. The nearest neighbours are selected by predefined similarity measures. Hence, the two parameters that are used to define the similarity of observations are the number of neighbours in conjunction with the respective distance measurement and need to be determined beforehand in order to achieve good classification results.

There exist a multitude of similarity measurements, like distance, correlation, or cosine similarity. The most commonly applied distance measurements (Manhattan and Euclidean distance) can further be generalized under the Minkowski distance formalized hereafter

[139] Choosing a different distributional kernel might also improve the results significantly, as shown by John and Langley (1995).

(Kotu & Deshpande, 2014, pp. 102–108). The distance d between two points $P(p_1, p_2, \dots, p_n)$ and $Q(q_1, q_2, \dots, q_n)$ in an n-dimensional space is depicted by equation 10.

$$d = \left(\sum_{i=1}^{n} |p_i - q_i|\right)^{\frac{1}{b}} \tag{10}$$

For $b = 1$ we attain the Manhattan distance, for $b = 2$ the Euclidean distance. For a new observation, the predicted label c^* is the majority vote over the class affiliation c_i of the i-th neighbour n_i.[140]

$$c^* = maximum\ class\ (c_1, c_2, \dots, c_k) \tag{11}$$

To account for the degree of similarity in the majority vote, each neighbour that is part of the voting decision in the scenario for $k > 1$ needs to be weighted (Kotu & Deshpande, 2014, p. 105). Higher weights resemble higher similarities and the sum across all weights must be one. For a new observation x^*, w_i is the weight of the i-th neighbour n_i with the total number of neighbours determined by k and the respective distance d. In this example, exponentially decaying weights following Kotu and Deshpande (2014, p. 205) are used.

$$w_i = \frac{e^{-d(x^*,n_i)}}{\sum_{i=1}^{k} e^{-d(x^*,n_i)}} \tag{12}$$

After calculating the weight, the class majority vote can be adjusted accordingly.

$$c^* = maximum\ class\ (w_1 * c_1, w_2 * c_2, \dots, w_k * c_k) \tag{13}$$

Besides distance measurements, this study also tests the performance of cosine similarity and correlation as measures of similarity in order to rely on a comprehensive set of approaches. Correlation assesses the linear relationship between two observations P and Q. It ranges between -1 and 1, with higher absolute values representing a stronger relationship in a negative or positive direction and a value of zero indicating no relationship. A correlation coefficient of 1 would simply relate to a linear relationship, not ruling out the possibility to have of having a higher degree relationship (e.g. quadric) (Kotu & Deshpande, 2014, p. 106). To calculate the Pearson correlation $\rho_{P,Q}$ between P and Q, the covariance between both observations needs to be calculated and normalized using the standard deviation of P and Q.

$$\rho_{P,Q} = \frac{cov(P,Q)}{\sigma_P * \sigma_Q} \tag{14}$$

[140] Besides the common majority vote, a number of alternative voting procedures have been proposed but will not be discussed hereafter (e.g. Coomans & Massart, 1982)

The covariance is formulated in equation 15.

$$cov(P, Q) = \frac{1}{n-1}\sum_{i=1}^{n}(p_i - \mu_p) * (q_i - \mu_q) \tag{15}$$

The k-highest correlating observations represent the k-nearest neighbours and the class of a new observation is determined following equation 13. It is important to note that this study is interested in the correlation between two observations and not the correlation between variables (features). To clarify in brief, each observation is described by a vector of fixed size formed across the features. For the calculation of the correlation between the two observations, the values from the feature vectors are taken.

Cosine similarity measures the angle between two vectors P and Q (Kotu & Deshpande, 2014, p. 107). An angle of 90° would describe orthogonal vectors and result in a cosine of 0. Parallel vectors with an angle of 0° or 180° would take the respective values of 1 and -1. The cosine similarity therefore is a measure of orientation, with higher values representing more similar vectors.

$$similarity = \frac{P \cdot Q}{\|P\| * \|Q\|} = \frac{\sum_{i=1}^{n} P_i * Q_i}{\sqrt{\sum_{i=1}^{n} P_i^2} * \sqrt{\sum_{i=1}^{n} Q_i^2}} \tag{16}$$

Cosine similarity is often used in document classification settings, where absolute word counts represent the vectors that describe each observation (document) (Oduntan, 2018).[141]

Determining the optimal number of neighbours requires some testing. The most straightforward approach with a single neighbour suffers from high variance and usually, high testing error, decreasing with higher values of k (Shalev-Shwartz & Ben-David, 2014, pp. 260–264). This trend usually reaches a minimum level before classification quality decreases again. K-nearest neighbour is a conceptually simple classification approach, yet with considerable computational effort due to the requirements to stored data, especially during the testing phase, where the whole table of observations has to be stored in order to determine the nearest neighbours of the unknown observations (Shalev-Shwartz & Ben-David, 2014, p. 264).

The parameters are optimized for each feature vector (qualitative, quantitative, and combined) and sampling approach (matched and realistic) via 5-fold cross-validation on a random data set. The parameters that have been tuned are the types of distances and the

[141] Thereby, cosine similarity does not have to be incorporated into a KNN classifier but is often used in a one-on-one comparison of documents (Oduntan, 2018, p. 60).

number of neighbours (k). Table 7 presents the results of the optimization process.[142] For the realistic sampling approach, the Manhattan distance with a large number of neighbours yields the best results for each of the feature vectors. For the matched sampling approach, the Manhattan distance also scored the best results, albeit in conjunction with a lower number of neighbours compared to the realistic sampling setting.

[142] See 4.2.5 for a detailed discussion of the underlying performance measures.

Matched sampling

Quantitative feature vector

k

	3	5	7	9	11	13	15	17
Euclidean	0.584	0.602	0.587	0.580	0.577	0.562	0.564	0.560
Correlation	0.522	0.534	0.556	0.547	0.555	0.553	0.554	0.540
Cosine	0.563	0.553	0.561	0.550	0.564	0.563	0.559	0.554
Manhattan	0.641	0.638	0.628	0.622	0.612	0.603	0.587	0.579

Qualitative feature vector

	3	5	7	9	11	13	15	17
Euclidean	0.693	0.716	0.721	0.716	0.713	0.711	0.695	0.690
Correlation	0.654	0.654	0.664	0.657	0.643	0.630	0.635	0.634
Cosine	0.744	0.754	0.737	0.727	0.711	0.699	0.696	0.691
Manhattan	0.823	0.817	0.806	0.799	0.790	0.782	0.775	0.774

Combined feature vector

	3	5	7	9	11	13	15	17
Euclidean	0.586	0.603	0.587	0.581	0.577	0.562	0.564	0.560
Correlation	0.527	0.525	0.550	0.530	0.555	0.545	0.551	0.540
Cosine	0.563	0.553	0.561	0.550	0.564	0.563	0.559	0.554
Manhattan	0.651	0.641	0.630	0.626	0.609	0.601	0.591	0.583

Realistic sampling

Quantitative feature vector

k

	3	5	7	9	11	13	15	17
Euclidean	0.615	0.630	0.638	0.646	0.654	0.656	0.660	0.671
Correlation	0.581	0.596	0.607	0.615	0.613	0.618	0.620	0.618
Cosine	0.581	0.600	0.614	0.626	0.624	0.627	0.634	0.634
Manhattan	0.654	0.666	0.679	0.691	0.693	0.700	0.707	0.710

Qualitative feature vector

	3	5	7	9	11	13	15	17
Euclidean	0.746	0.750	0.755	0.763	0.769	0.771	0.773	0.778
Correlation	0.770	0.773	0.771	0.770	0.776	0.773	0.775	0.775
Cosine	0.774	0.773	0.782	0.786	0.791	0.793	0.797	0.802
Manhattan	0.793	0.798	0.799	0.800	0.811	0.812	0.816	0.816

Combined feature vector

	3	5	7	9	11	13	15	17
Euclidean	0.615	0.630	0.638	0.646	0.654	0.656	0.660	0.671
Correlation	0.580	0.597	0.610	0.621	0.617	0.618	0.626	0.625
Cosine	0.581	0.600	0.614	0.626	0.624	0.627	0.634	0.634
Manhattan	0.659	0.673	0.683	0.697	0.696	0.707	0.718	0.710

The table reports AUC values of parameter combinations for different feature vectors and sampling approaches. The parameter setup has been carried out on a random, unused dataset via stratified 5-fold cross-validation (mean AUC values are reported). Darker grey areas indicate better detection performance. k: number of neighbours.

Table 7 – KNN parameter optimization

4.2.8 Artificial Neural Networks

The basic idea behind artificial neural networks (ANN) is taken from biology and roughly resembles the function of a neuron (Goodfellow et al., 2016, pp. 12–13). In biology, the neuron or nerve cell is the primary component of the nervous system that receives, processes and transmits electric and chemical signals (Zurada, 1992, pp. 26–30). A typical neuron consists of the cell body with the nucleus, which is surrounded by dendrites and a single axon with synapses (Jain, Mao, & Mohiuddin, 1996, p. 33). A neuron is connected to another neuron via the synapses. Thus, the synapses of the neuron are connected to the dendrites of another neuron, forming a complex communication network.

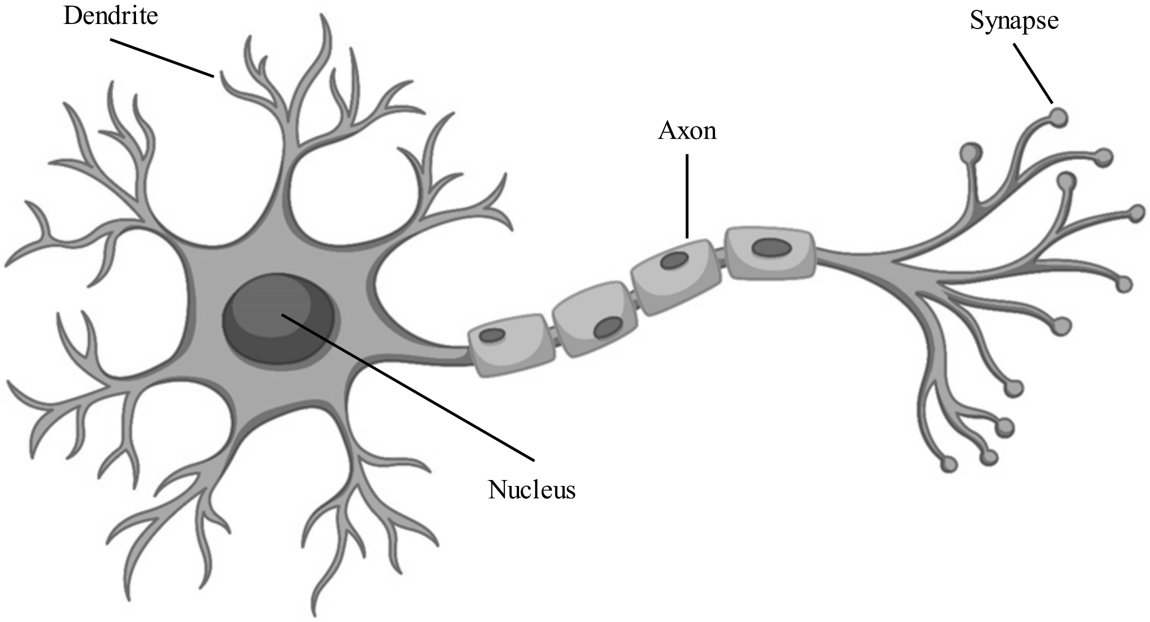


Figure 18 – Schematic anatomy of a neuron[143]

An artificial neural network is modelled in reference to this biological prototype. In its basic structure, it consists of neurons, also called nodes or units, which are arranged consecutively in multiple layers (Zurada, 1992, pp. 37–42). In terms of graph theory, a rudimentary artificial neural network resembles a directed graph with nodes being called neurons, while the links between them are edges (Shalev-Shwartz & Ben-David, 2014, p. 268). The number of layers may vary: in its most simplistic form, an artificial neural network has two layers, an input, and an output layer. This type of ANN is called perceptron and is comparable to a linear classifier (Kotu & Deshpande, 2014, pp. 124–125). With the

[143] Modified from the original, designed by “brgfx / Freepik” in accordance with free-use terms.

addition of further layers in between the input and output layer, called hidden layers, modelling non-linear relationships is possible. In feed-forward networks, the input travels one-directionally through the net without any loops between layers (Shalev-Shwartz & Ben-David, 2014, p. 269). Recurrent or feedback networks, in contrast, have feedback connections to initiate loops (Jain et al., 1996, pp. 34–35).

The similarities can be seen when comparing the biological structure of a neuron as exemplified in Figure 18 to the basic design of a neuron (for example as part of a hidden layer) in an artificial neural network as in Figure 19. The neuron receives inputs, which are processed via an activation function and further transmitted in the network (Zurada, 1992, pp. 32–34).

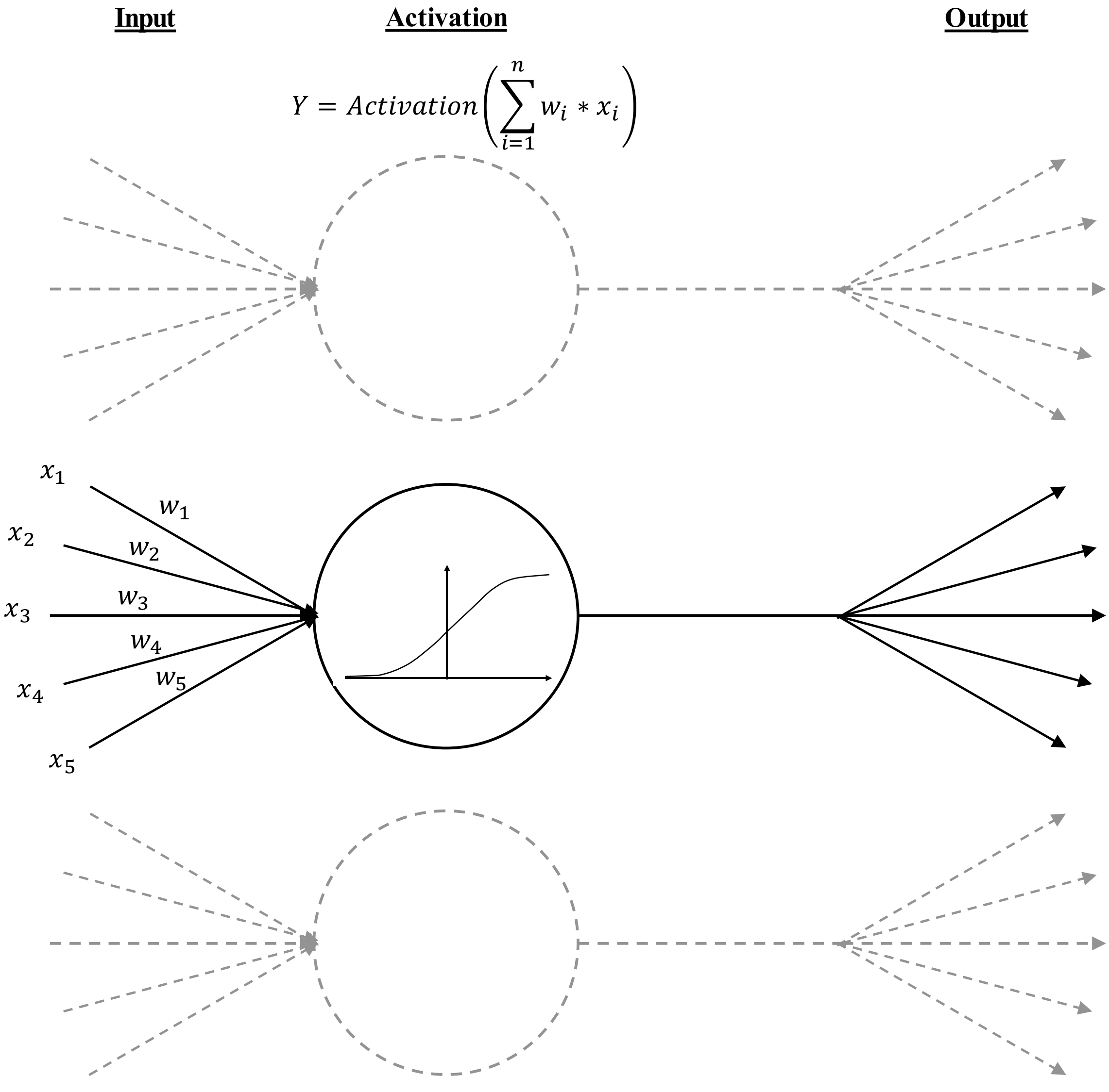


Figure 19 – Neuron of an artificial neural network

Each neuron of the following layer is connected to each neuron of the previous layer. The value that is transferred from the neurons of the previous layer to a specific neuron in the following layer is the weighted sum of the inputs of the neuron (Y), where x_i is the input from the i-th neuron of the previous layer and w_i is the respective weight.[144] To process the weighted sum of the inputs to an output, an activation function is used (Jain et al., 1996, p. 35). In biological terms, if a particular activation potential is achieved, the neuron fires over its axon to connected neurons. The mathematical equivalent is the output of the activation function. The weighted sum of the inputs defines the value that is processed by the activation function and the output determines the value, which is then the input for connecting neurons, again weighted accordingly.

$$Y = \sum_{i=1}^{n} w_i * x_i + bias \tag{17}$$

As the activation function is used to transform the input of a neuron into an output, determining the type of function is crucial (Karlik & Olgac, 2011). The (uni-polar) sigmoid function, as expressed in equation 18, is a common activation function that is typically applied in binary classification scenarios, like that presented in this study (Karlik & Olgac, 2011, pp. 112–113).[145] Concerning the calculations, sigmoid functions tend to require more computational effort than, for example, the restricted linear unit function (ReLu), which is also commonly used. In general, the number of different applicable activation functions is manifold (Witten et al., 2017, p. 425).

The activation functions may also vary between the layers. Therefore, it is possible to have a ReLu function in the hidden layer and a sigmoid function in the output layer. The simple ReLu function has an output of 0 if the respective input is less than 0 and the true output is otherwise (19).

$$f(x) = \frac{1}{1 + e^{-x}} \tag{18}$$

$$f(x) = \max(x, 0) \tag{19}$$

The size of the input layer (the number of neurons in the input layer) is determined by the size of the input vector. The input layer usually consists of F or $F + 1$ neurons, where F is

[144] The relevance of a bias term will be discussed later.

[145] The uni-polar sigmoid activation function takes values between 0 and 1, describing the output of a typical binary coded classification problem. The bi-polar sigmoid function takes values between -1 and 1 (Karlik & Olgac, 2011, pp. 112–113).

the number of features in the input vector (Zurada, 1992, p. 143). For this study, for the most part, three different-sized feature vectors are utilized.[146] The quantitative feature vector consists of 19 features, the qualitative vector of 1,000 features, and the combined one of 1,019 features. As a general rule of thumb, higher dimensional feature vectors and therefore larger input layers are said to be better suited to extract classification patters than smaller ones (Zurada, 1992, pp. 96–97). The number of features and therefore the input layer size directly influences the training time of the ANN. As a general rule of thumb, the length of the computation (time) L does increase by order of L^2 when adding new features, which would translate to quadruplicating the training time if the input layer size is doubled (Kavzoglu, 1999, p. 676).

On the other hand, the output layer in the binomial classification scenario (non-fraudulent versus fraudulent) consists of two neurons that translate the inputs from the hidden layer neurons and assign each observation to one of the two groups. Usually, the number of classes imposes the number of output neurons. However, in multi-class scenarios, the output neurons can be pruned through distributed representation (Zurada, 1992, pp. 215–216).[147] For example, four classes (0,1,2,3) could be represented via two neurons (00,01,10,11). However, pruning may lead to longer training cycles and less robust results (Zurada, 1992, p. 216).

It has already been discussed that the input of a neuron represents the weighted sum of inputs of the neurons from a previous layer. Determining the weights is an essential part of the actual learning process. The learning process is carried out through a technique called backpropagation. Backpropagation is an iterative technique that updates the weights of the connections between the neurons. In this case, the network topology and the underlying activation function are already given. Backpropagation again resembles the biological process of signal transmission between neurons (Kotu & Deshpande, 2014, pp. 127–128). The strength of the connections between the neurons is optimized by adjusting the weights of them based on their relevance for the resulting output, which can be done by gradient descent algorithms (Witten et al., 2017, pp. 266–268). Therefore, the labelled observations from the training set are used to determine the error (err), where x is the input, $f(x)$ is the output from the sigmoid function and Y^* is the true output.[148]

[146] For answering enhancing question 2, the layer sizes of the quantitative and the combined vector are increased. See 5.2.3 for this special case.

[147] This is only possible if no class decoding is required (Zurada, 1992, p. 216).

[148] Despite error-correction rules, other learning algorithms like Boltzmann learning or the Hebbian rule (among others) have developed (Jain, Mao, & Mohiuddin, 1996, pp. 36–37).

$$err = Y^* - f(x) \tag{20}$$

Often the quadratic loss function is used instead of the absolute error loss function, as the latter is not differentiable at 0, leading to problems in the optimization procedure (Witten et al., 2017, pp. 177–178).

$$err = (Y^* - f(x))^2 \tag{21}$$

By adjusting the weights continuously, the error is minimized. This correction rate of the weights is crucial to determine an optimized model without overcorrecting, which could lead to missing potential minima through larger steps (Kotu & Deshpande, 2014, p. 129). The old weights w' are updated by a fraction λ of the error, which is called the learning rate, to calculate the updated weight w (Kotu & Deshpande, 2014, pp. 129–130). The learning rate can take values between 0 and 1.

$$w = w' + \lambda * err \tag{22}$$

Under the gradient descent algorithm, the derivative of the squared error loss function with regard to the weights can be taken from equation 23.[149]

$$\frac{\partial err}{\partial w_{i,j}} = \big(Y - f(x)\big)f'(x)w_i f'(x_i)a_i \tag{23}$$

Where x represents the weighted sum of the inputs, $f(x)$ is the output of the sigmoid activation function, f'(x) denotes the derivative of the sigmoid function and a_i is an input vector. The weight of the connection between the j-th input unit and the i-th hidden unit is denoted by w_{ij}. The weight from the i-th hidden unit to the output unit is denoted by w_i.[150] This derivative for the error function is calculated for every observation in the training set. The result after each iteration is then used to sum up the changes associated with the specific weights w_{ij}, multiplying the learning rate λ and subtracting the outcome from the current value of w_{ij} (Kotu & Deshpande, 2014, p. 129).

The learning rate does not have to be static throughout the optimization process but can be adjusted gradually to move from higher values to smaller values later (Kotu & Deshpande, 2014, p. 129). It is conducted by adding a penalty to the error function, which is calculated as the square sum across all weights, thus also solving the problem of irrelevant connections that do not add to the error reduction.

[149] See Witten, Frank, Hall, and Pal (2017, pp. 267–268) for the derivation of equation 23.

[150] In the case of a single output unit, which is feasible in a scenario with two classes.

The explanatory artificial neural network in Figure 20 illustrates a four-layer approach with 68 neurons. It is divided into an input layer with 20 neurons, two hidden layers with 23 neurons each and an output layer with two neurons. The connections between the nodes (in the figure illustrated by grey lines) represent the weights. The darker the line, the greater the weight of the connection. Each neuron of a layer is connected to each neuron of the subsequent layer. The input nodes represent the 19 quantitative features as well as one additional node to capture bias. The bias neuron is similar to an absolute (constant) term of a regression function (Zurada, 1992, p. 165). It can be added to help shift the activation function to the left or to the right, which cannot be done solely by the weighting term. Changing the weight only adjusts the shape (steepness) of the activation function. When looking closely at Figure 20, the additional bias neuron can be identified, which is not connected to a previous layer. The subjective part of the topology of the network is constituted by the number of hidden layers and the respective size of each layer and is depending on the complexity of the task but sometimes regarded as an invariant concerning the parameter tuning (Zurada, 1992, p. 569). Hastie et al. (2016, p. 400) suggest to use rather too many than too few neurons in the hidden layer but also mention that topology is often a matter of experimentation or background knowledge.

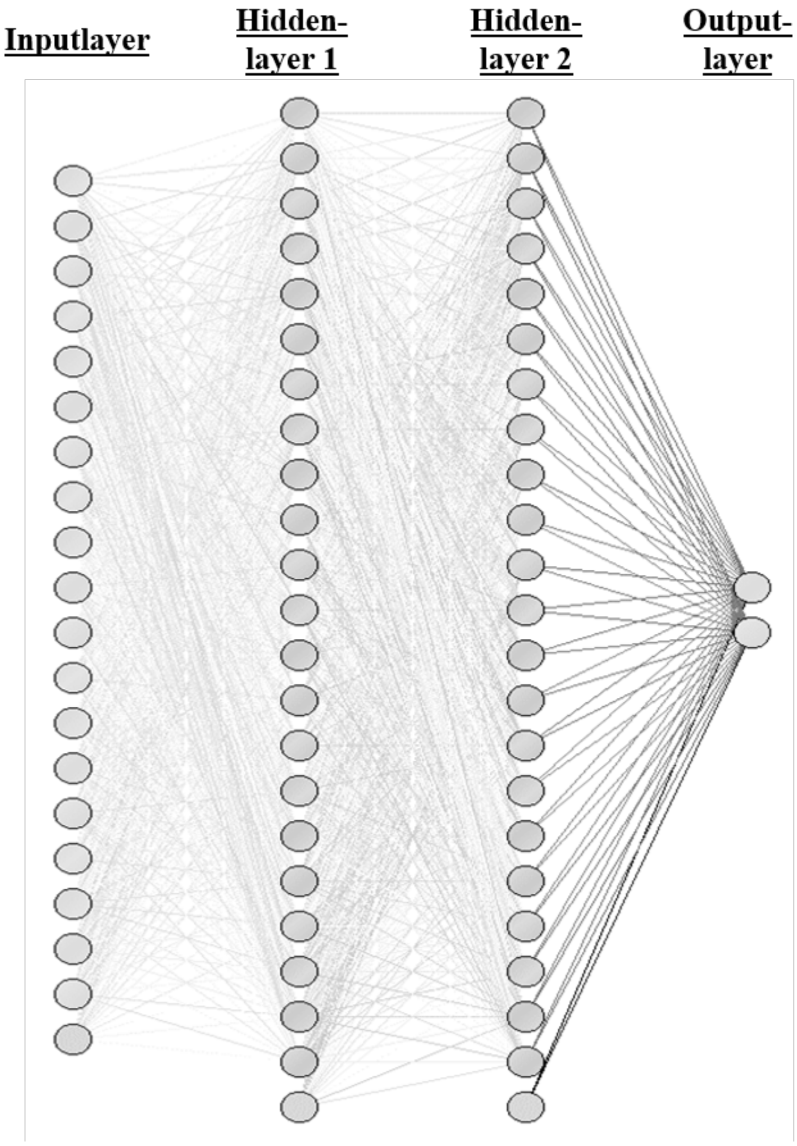


Figure 20 – Example of an artificial neural network

The parameter set-up for artificial neural networks will briefly be discussed hereafter. As with the other classifiers, the parameters have been tuned on a random data set using 5-fold cross-validation. The biggest problem when setting up artificial neural networks is the sheer number of possibilities when constructing the network. This study relies on general hints from the literature and tests a reasonable number of network sizes in combination with different parameter values. In comparison to the other classifiers, the ANN has the highest potential for further optimization.

The first parameter of interest is the learning rate and related parameters, which influence the changes to learning rate during the training process. As the learning rate determines the sensitivity of the changes to weights, slowing down learning rate changes reduces the impact of outliers in later iterations (Kotu & Deshpande, 2014, p. 131). Moreover, weight decay adds to generalization and limits overfitting (Krogh & Hertz, 1992). Hastie et al. (2016, pp. 398–400) also reveal the positive impact of small weight decays on the error rate and suggest to use this technique as a regularization to avoid overfitting. In addition to weight decay, the momentum parameter influences the learning rate by adding a fraction of the prior update to the current iteration, smoothening the optimization process and reducing the likelihood of becoming stuck in a local maximum (Kotu & Deshpande, 2014). Learning rate and momentum are interdependent and should be tuned in conjunction (Smith, 2018). At first glance, momentum and weight decay might work against each other; however, they serve different purposes and work very well together (Smith, 2018). Momentum adjusts the step heights continuously, whereas weight decay leads to different values throughout the training process. In this study, different learning rates and values for momentum have been tested while keeping the weight decay constant.

Matched sampling						Realistic sampling					
Quantitative feature vector						**Quantitative feature vector**					
		Learning rate						Learning rate			
		0.1	0.3	0.5	0.7			0.1	0.3	0.5	0.7
m.	0.1	0.554	0.562	0.562	0.564	m.	0.1	0.695	0.711	0.712	0.711
	0.3	0.557	0.566	0.561	0.563		0.3	0.702	0.712	0.711	0.713
	0.5	0.559	0.562	0.562	0.563		0.5	0.705	0.711	0.713	0.717
	0.7	0.562	0.568	0.569	0.558		0.7	0.712	0.713	0.719	0.722
Qualitative feature vector						**Qualitative feature vector**					
		Learning rate						Learning rate			
		0.1	0.3	0.5	0.7			0.1	0.3	0.5	0.7
m.	0.1	0.846	0.851	0.451	0.567	m.	0.1	0.649	0.717	0.720	0.723
	0.3	0.845	0.597	0.551	0.507		0.3	0.650	0.717	0.723	0.734
	0.5	0.849	0.425	0.461	0.478		0.5	0.686	0.720	0.735	0.749
	0.7	0.800	0.469	0.200	0.400		0.7	0.719	0.730	0.753	0.732
Combined feature vector						**Combined feature vector**					
		Learning rate						Learning rate			
		0.1	0.3	0.5	0.7			0.1	0.3	0.5	0.7
m.	0.1	0.838	0.842	0.851	0.859	m.	0.1	0.715	0.755	0.754	0.768
	0.3	0.844	0.846	0.847	0.851		0.3	0.731	0.753	0.762	0.780
	0.5	0.845	0.847	0.851	0.859		0.5	0.740	0.759	0.776	0.787
	0.7	0.844	0.852	0.852	0.836		0.7	0.754	0.775	0.771	0.775

The table reports AUC values of parameter combinations for different feature vectors and sampling approaches. The parameter setup has been carried out on a random, unused dataset via stratified 5-fold cross-validation (mean AUC values are reported). Darker grey areas indicate better detection performance. m: momentum.

Table 8 – ANN parameter optimization

The results of the tuning test for learning and momentum rates are presented in Table 8. For the realistic sampling approach, higher values for learning rates and momentum seem to achieve the best results for each of the feature vectors, while the results for the matched sampling approach seem to follow no overarching pattern.

The network architecture is the core of every ANN and theoretically allows for endless possible configurations. The two most important parameters regarding the architecture are the number of hidden layers and their respective sizes (Wilamowski, 2009, pp. 57–59). A very simplistic way to start is by using one hidden layer and increasing the number if the ANN is not able to produce reliable and satisfying results (Jain et al., 1996, pp. 38–39). The universal approximation theorem states that a single hidden layer with a finite amount of neurons can approximate arbitrary continuous functions (Cybenko, 1989, p. 312). The theorem was first proved by Cybenko (1989) for sigmoid activation functions. The assumption was later relaxed by Hornik (1991), who showed that feed-forward neural networks with a single hidden layer represented a potential universal approximator.[151] Further, Wilamowski (2009, p. 60) argues that for generalizability purposes an artificial neural network should have as few neurons as possible. Building upon this knowledge, the number of hidden layers was set to one, leaving only the hidden layer size open to tuning. The number of neurons in the hidden layer should not be too small, as this would lead to a loss of flexibility and the ANN may not be able to capture peculiarities, especially in the case of non-linearities in the data (Hastie et al., 2016, p. 400). The hidden layer size should furthermore be adjusted according to the size of the dataset and the number of input features. Both are varying considerably in this study with the dataset size being considerably influenced by the realistic and the matched sampling approaches, and the feature input size being influenced by the three different feature vectors. Therefore, the detection quality is tested for variations of the peculiarities for four different hidden layer sizes, ranging from 100 to 800 neurons. The differences in detection performance between the results for larger networks and the ones presented in Table 8 are hardly noticeable in most cases. Table 9 shows the single highest AUC values for different network sizes and feature vectors for the matched sampling process.[152] As suggested by Zurada (1992, pp. 569–570), the network size was an invariant in the other parts of the parameter tuning procedure.

[151] In practice, width (number of neurons in the hidden layer) cannot always substitute for depth (number of hidden layers) very well (Lu, Pu, Wang, Hu, & Wang, 2017).

[152] The parameter optimization in this scenario is carried out exactly as the one presented in table 8, with learning rate and momentum variations. To reduce the number of tables, only the highest AUC values, which are the most relevant, are shown in table 9.

Layer size	Quant.	Qual.	Qual. & quant.
100*	0.569	0.851	0.859
200	0.567	0.856	0.864
400	0.565	0.847	0.861
800	0.571	0.851	0.857

Highest AUC values for different-sized networks

*100 hidden units as presented in Table 8

Table 9 – ANN performance with different hidden layer sizes

The results reveal that larger networks do not seemingly perform any better than smaller ones. Furthermore, to compare the highest AUC values, paired t-tests were conducted.

Sizes	Quant.	Qual.	Qual. & quant.
200>	0.895	0.817	0.395
400>	0.996	0.996	0.004
800>	1.000	1.000	0.395

p-values for paired t-tests comparing results from the baseline network with 100 neurons to larger networks[153]

Table 10 – Comparison of network sizes

The tests compare the results regarding the learning rate and the momentum of the baseline network with 100 neurons to the larger ones. The results can be taken from Table 10, which reports the respective p-values. From the nine additional tests (three compared network sizes for three feature vectors) for the matched sampling approach, only in one case is a larger network significantly better. For the realistic sampling approach, a similar procedure is undertaken. However, to keep computational effort reasonable, only three in contrast to nine additional combinations are tested.[154] In none of the cases does the larger network achieve significantly better results.

The network size has little impact on the performance, regardless of the feature vector and the sampling approach. This, at first glance counterintuitive result, has been proven to be true on different types of settings (e.g. Kavzoglu, 1999). In accordance with Occam's principle, Wang, Venkatesh, and Judd (1994) demonstrate that if the network is large enough

[153] The p-values are one-tailed for evaluating the alternate hypothesis of better performance of larger network sizes. The ">" sign indicates the respective direction. A similar notation will be chosen in the reporting of the results in chapter 5.

[154] At this point, it may be important to provide some context on the time requirements. For the realistic sampling approach, applying a 5-fold cross-validation on the random data set used for parameter optimization (size similar to one of the training sets of the initial sample), with qualitative features, 100 neurons in the hidden layer and 16 parameter combinations, it took six days and 15 hours to attain a single output (matrix) as reported in Table 8.

to learn from the data, further increases to the size of the ANN play little role in the generalizability performance and therefore should be avoided.[155]

To complement the aforementioned tuning parameters, some more general settings must be considered. The starting values of the weights are randomized near zero. Weights of zero would lead to zero derivatives, which in return would imply that the algorithm does not change accordingly in each step. Weights that are too big lead to poor performance due to larger differences in step size (Hastie et al., 2016, p. 398). Furthermore, normalizing the inputs on a scale from -1 to 1 leads to the eliminations of outlier inputs and translates to the domain of the sigmoid activation function (Kotu & Deshpande, 2014, p. 132). Furthermore, the models are trained with an early stopping rule (error epsilon of 1.0^{-5}) to prevent the ANN from simply memorizing the data.[156] Finally, the number of training iterations must be determined. The number of iterations determines the amount of each training cycle undertaken. It is largely contingent on the learning algorithm of the ANN (Wilamowski, 2009, p. 57). Error backpropagation, which is used in this study, is a rather inefficient method that requires more iterations to achieve satisfactory error rates (Wilamowski, 2009, p. 57).

The number of iterations is tested during the parameter optimization process. The AUC barely changes after around 1,000 iterations. An additional test was carried out for 5,000 iterations to see if the results would considerably change after longer training, but no significant improvement was noticeable. Figure 21 depicts the results measured by the AUC for different numbers of iterations.

[155] Occam's principle or Occam's razor is considered a fundamental tenant of science (see Domingos, 1999, for general effectuations and the adaption to knowledge discovery and machine learning in particular). The principle was postulated by William of Occam around the beginning of the 14th century and criticizes the trend of formulating theories that are more complex without generating much or any improvement over already existing, simpler ones. In simple terms, it comes down to the lesson that when one needs to consider different theories that are all making the same predictions, the simplest one is the best. Two interpretations, also known as the first and second Occam's razor, are relevant for this study. The first Occam's razor states that "given two models with the same generalization error, the simpler one should be preferred because simplicity is desirable in itself" and the second one that "given two models with the same training-set error, the simpler one should be preferred because it is likely to have lower generalizability error" (Domingos, 1999, p. 410). Given both principles and the results of the parameter set-up, the simpler structured network will be preferred if performance is similar. Occam's razor is not only applicable to neural networks but to every classifier that is applied in this study to a certain extent.

[156] When measuring the performance of ANNs, epsilons of ~1.0^-5 are considered very small errors (Wilamowski, 2009).

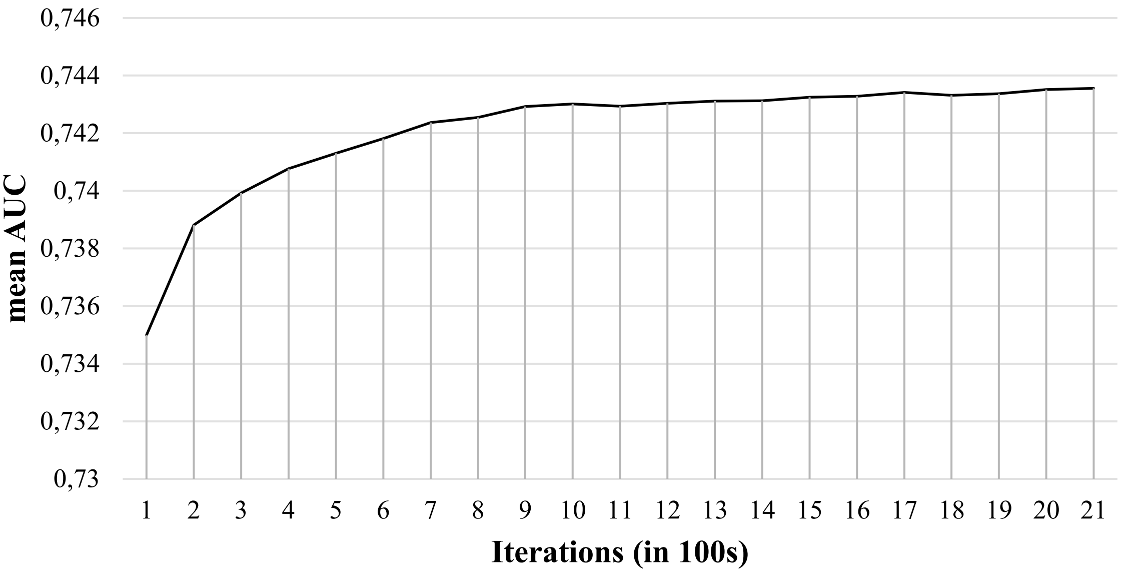


Figure 21 – Iterations and detection performance

Like every other approach, artificial neural networks have their advantages and disadvantages in general as well as regarding their suitability in this study's classification scenario. Besides this unusual and seemingly remarkable approach, artificial neural networks are at their core a non-linear classifier akin to a classification or a regression model with two stages (Hastie et al., 2016, p. 392). Finding the optimal combination of possible parameter set-ups with respect to typology, learning rate, activation functions, and iterations, among others, takes considerable time, especially compared to the other methods used in this study. Hastie et al. (2016, p. 397) have stated that training an artificial neural network is an art, given the number of possible parameter combinations and the overall architecture. Finding the optimal parameter set-up – if it even exists – is beyond the scope of this work. Depending on the set-up of an artificial neural network, computing effort during the training phase can be much greater compared to the other classifiers incorporated, but trained networks are very fast during the testing phase. Incremental error correction may cause the final model to fall into a local optimum. This is a general problem with the gradient descent algorithm, which can only find local minima when several minima are present and may not converge to the global minimum (Witten et al., 2017, pp. 265–266).

4.2.9 Support Vector Machines

The fourth classifier applied comprises support vector machines (SVMs). Like the naïve Bayes classifier and the artificial neural network, support vector machines are a commonly used tool for machine learning tasks. Support vector machines reflect a younger development than the other techniques, although they adopt a basic idea that dates back to the early stages of pattern recognition.

In 1936, Fisher proposed a first algorithm for pattern recognition, inspiring an emerging subfield of statistics. He used the Iris sample (one that is very popular today), which contains data on the measurements of flowers, to demonstrate its ability to discriminate different plants based on a set of attributes (Fisher, 1936). Fifty observations for each of the three plants Iris setosa, Iris versicolor and Iris virginica were collected and their sepal length and width, as well as their petal length and width, were measured.[157] His idea of assessing linear functions to discriminate populations based on measured attributes was subsequently developed further to also take non-linear relationships into consideration, but it remains a fundamental approach even for modern techniques like support vector machines (Cortes & Vapnik, 1995). While working at AT&T Bell Labs, Cortes and Vapnik developed algorithms for pattern recognition and published one of the first formulated approaches of support vector machines, with the intention of solving binary classification problems (Kotu & Deshpande, 2014, p. 134).

Within the basic concept of support vector machines and as the name suggests, data points are represented in vector form. The set of attributes (features) defines the position of each data point in an n-dimensional space. The term “support” in the name “support vector machines” thus refers to the data points that determine the boundary between the groups (Cortes & Vapnik, 1995, p. 274). This boundary is called the hyperplane. To ensure the boundary not being biased (too close) towards one or another group, the average geometric distance (margin) is maximized (Kotu & Deshpande, 2014, p. 135). As illustrated in Figure 22, the two groups represented by the black and grey dots are linearly separated by the fitted hyperplane, which is based on the marked support vectors (dots in the circle).

[157] The sample was collected and published by Anderson (1935) but became even more popular through the work of Fisher (1936).

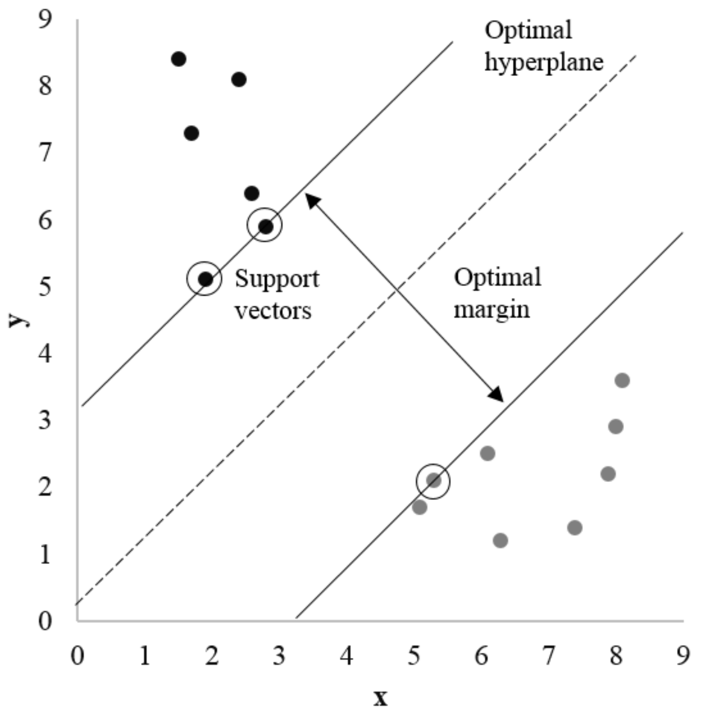


Figure 22 – SVMs in a two-dimensional space

In the example depicted in Figure 22, the classes are perfectly separable by a linear function.[158] However, finding the best possible hyperplane among all of the potential options is important for subsequent classification performance. The optimal hyperplane can be expressed as

$$x = \beta + \sum_i \alpha_i c_i a(i) \cdot a \tag{24}$$

where c_i is the class label (in this case distinguishing the class affiliation through the values 1 and -1) of the input data point $a(i)$ (Witten et al., 2017, p. 254). Support vectors are part of the training data and are indicated through i. The dot product of a support vector $a(i)$ and another vector a is a measure of geometrical similarity between both data points. The parameters α_i and β need to be determined through an optimization algorithm. They represent the "shape" of the hyperplane, in this form similar to the slope and absolute term of a linear function. With the help of the well-established field of quadratic constraint optimization, the optimization problem can be solved (Fletcher, 2010, pp. 165–214; Kotu & Deshpande, 2014, p. 140).

[158] The convex hulls (polygons) formed when connecting the data points of each class do not overlap, as they are linear separable. Hence the optimal hyperplane is the perpendicular bisector of the shortest line connecting the hulls (Witten, Frank, Hall, & Pal, 2017, p. 253).

So far, the basic idea of support vector machines has been presented in a linearly separable environment but as stated earlier, SVMs are also capable of determining non-linear boundaries.

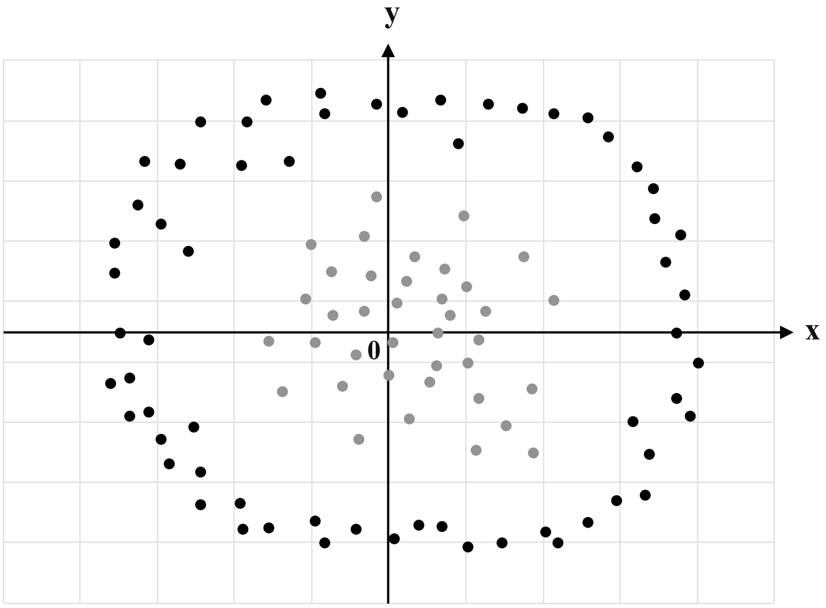


Figure 23 – Data before transformation

In comparison to Figure 22, Figure 23 depicts (at first sight) a non-linear separable data set. With the help of the so-called “kernel-trick”, the initial structure of the data is transformed by an *a priori* specified kernel function (Shalev-Shwartz & Ben-David, 2014, pp. 217–222). The basic idea is to map the data points to a higher dimensional space and thereby make them linearly separable (Cortes & Vapnik, 1995, p. 274). After the transformation process, the gap between the hyperplane and the data points is maximized, like in the case without kernel transformation. In this schematic example, a kernel function, as described in equation 25, provides a data transformation for the two groups, resulting in a solution separable by a linear function again, which schematically can be seen in Figure 24.

$$z = \sqrt{x^2 + y^2} \tag{25}$$

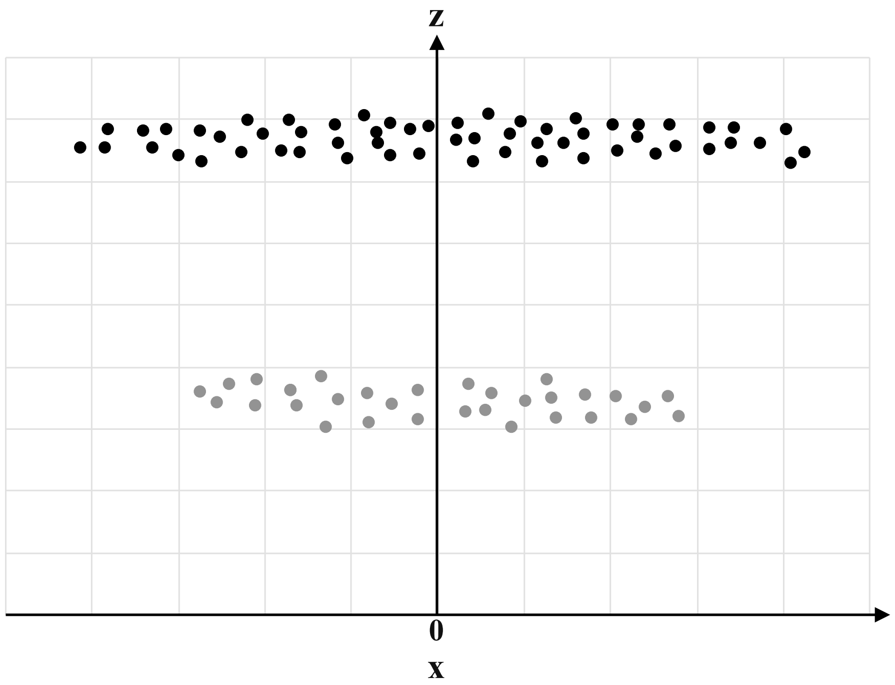


Figure 24 – Data after transformation

In most classification scenarios, both groups will not be linearly separable owing to the presence of data points lying inside the decision boundary or in the opposite group, resulting in misclassification. When determining the shape of the hyperplane, misclassified instances are taken into consideration when assessing the hyperplane parameters. For each classification error, a penalty is calculated (Kotu & Deshpande, 2014, p. 136). The penalty value (often abbreviated with a C) must be determined beforehand and assessed by cross-validation (James et al., 2017, p. 358). The parameter reflects the bias-variance tradeoff. Small thresholds of C result in smaller margins and thus hyperplanes leading to fewer violations, whereas large values of C result in softer margins with more potential violations.[159] Therefore, larger values of C result in a greater number of support vectors, with more bias but less variance. With the characteristic of a permeable boundary, the support vector machine is also referred to as a soft margin classifier (James et al., 2017, p. 345).

Support vector machines tend not to overfit as easily as other techniques (Witten et al., 2017, p. 255). The hyperplane is defined by a fraction of the complete data set, making sure to generalize the underlying discriminate. Adding or deleting data points only changes the

[159] C can also be regarded as costs of misclassification. Moreover, high values of C are associated with smaller margins and vice versa for lower values of C (Hastie, Tibshirani, & Friedman, 2016, p. 420). The technical implementation in this study's case builds upon the initial explanation for the parameter C.

underlying hyperplane, when support vectors are concerned (Kotu & Deshpande, 2014, p. 147). Furthermore, the robustness to smaller changes in the underlying data results in faster remodelling when additional instances are added. A disadvantage of support vector machines is the computational complexity, depending on the implementation (Kotu & Deshpande, 2014, p. 148; Witten et al., 2017, p. 255). Vector operations in a high-dimensional space require computational effort during the training and testing phase. Computational costs increase considerably with higher-order SVMs (Kotu & Deshpande, 2014, p. 147).

To ensure good classification results, choosing the "best" kernel function and misclassification penalty is critical. Moreover, the parameters of the kernel functions need to be considered as well. As is commonplace in machine learning tasks, the parameter set-up of SVMs is contingent on the underlying data and needs to be determined for the individual problem beforehand (Ben-Hur & Weston, 2010, pp. 230–234). Hastie et al. (2016, pp. 421–422) highlight the importance of cross-validated determination for C for each kernel function individually. Hence, for the complex classification problem, several kernels and values for C are tested here on a random data set using 5-fold cross-validation. To limit the parameters, this study relies on generalizable hints and practical guides from the literature.

Kernel function	**Mathematical expression**	**Additional parameters**	
Dot product	$f(x,y) = x \cdot y$		(26)
Radial basis function	$f(x,y) = \exp(-\gamma\|x-y\|^2)$	γ	(27)
Polynomial function	$f(x,y) = (x \cdot y + 1)^n$	n	(28)

Table 11 – Kernel functions

The kernel functions exploited in this study range from polynomials of degree 1-4, to radial and dot kernels. Witten et al. (2017, pp. 255–256) suggest that small values of n suffice for polynomials, while Ben-Hur and Weston (2010, p. 226) and James et al. (2017, pp. 352–354) note that especially for high-dimensional data sets (with many features), one should start with a linear kernel as a benchmark before beginning to exploit non-linear kernels. Hsu, Chang, and Lin (2003) have empirically tested different kernels on a range of data sets and have extracted baseline rules to initialize the parameter optimization process, recommending that one start with radial basis functions. The optimization parameter of interest for the radial basis function, as depicted in Table 9, is γ, which accounts for flexibility and complexity. Low values of γ result in low flexibility, meaning that the function is unlikely to incorporate complex structures, whereas high values stimulate the function to reach out to individual data points, enabling it to capture complex data structures. Balancing γ is another important task to avoid running into overfitting (Hsu et al., 2003, pp. 4–6). Table 11 presents the exploited kernel functions with the respective additional parameters that need to be established. As Hsu et al. (2003) suggest, several grids with C, γ and C, n are tabulated to find suitable combinations of the parameters.

Table 12 presents the results of the parameter optimization process. This study does not undertake optimization through additional iterations in optimal areas of the grid with smaller steps of parameter values. The results in Table 12 suggest that a polynomial kernel function of degree two is best suited for both feature vectors containing qualitative features. The results hold for the realistic and matched sampling approach, with greater values for C. For the quantitative feature vector, a high-complexity radial basis function (high values of C and γ) scores the best results.

Matched sampling

Polynomial kernel function

		Quantitative feature vector				Qualitative feature vector				Combined feature vector			
		n				n				n			
		1	2	3	4	1	2	3	4	1	2	3	4
C	-1	0.555	0.564	0.545	0.524	0.847	0.780	0.602	0.500	0.845	0.784	0.572	0.500
	0	0.555	0.564	0.545	0.519	0.847	0.780	0.602	0.500	0.845	0.784	0.572	0.500
	5	0.556	0.594	0.557	0.590	0.786	0.862	0.745	0.528	0.792	0.859	0.692	0.511
	10	0.556	0.548	0.557	0.590	0.786	0.862	0.745	0.528	0.792	0.859	0.692	0.511

Dot kernel function

		Quantitative feature vector	Qualitative feature vector	Combined feature vector
C	-1	0.556	0.858	0.861
	0	0.556	0.858	0.861
	5	0.554	0.787	0.797
	10	0.553	0.787	0.797

Radial kernel function

		Quantitative feature vector				Qualitative feature vector				Combined feature vector			
		γ				γ				γ			
		0	0.5	2	4	0	0.5	2	4	0	0.5	2	4
C	-1	0.500	0.689	0.689	0.689	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.502
	0	0.702	0.703	0.712	0.712	0.581	0.581	0.581	0.594	0.500	0.500	0.500	0.502
	5	0.712	0.711	0.711	0.706	0.602	0.602	0.602	0.605	0.500	0.500	0.500	0.502
	10	0.706	0.706	0.708	0.708	0.602	0.602	0.602	0.605	0.500	0.500	0.500	0.502

Realistic sampling

Polynomial kernel function

		Quantitative feature vector				Qualitative feature vector				Combined feature vector			
		n				n				n			
		1	2	3	4	1	2	3	4	1	2	3	4
C	-1	0.588	0.598	0.635	0.632	0.700	0.824	0.573	0.500	0.776	0.829	0.607	0.500
	0	0.588	0.598	0.635	0.632	0.700	0.824	0.573	0.500	0.776	0.829	0.607	0.500
	5	0.540	0.606	0.643	0.680	0.727	0.859	0.619	0.500	0.745	0.829	0.625	0.500
	10	0.511	0.600	0.643	0.680	0.726	0.873	0.619	0.500	0.748	0.829	0.625	0.500

Dot kernel function

		Quantitative feature vector	Qualitative feature vector	Combined feature vector
C	-1	0.632	0.750	0.775
	0	0.632	0.750	0.775
	5	0.540	0.656	0.696
	10	0.539	0.656	0.696

Radial kernel function

		Quantitative feature vector				Qualitative feature vector				Combined feature vector			
		γ				γ				γ			
		0	0.5	2	4	0	0.5	2	4	0	0.5	2	4
C	-1	0.500	0.650	0.722	0.735	0.498	0.798	0.802	0.709	0.500	0.509	0.504	0.501
	0	0.500	0.650	0.722	0.735	0.500	0.804	0.802	0.712	0.500	0.509	0.504	0.501
	5	0.500	0.695	0.725	0.734	0.500	0.804	0.802	0.702	0.500	0.510	0.504	0.502
	10	0.500	0.700	0.720	0.733	0.500	0.863	0.802	0.702	0.500	0.510	0.504	0.502

The table reports AUC values of parameter combinations for different feature vectors and sampling approaches. The parameter setup has been carried out on a random, unused dataset via stratified 5-fold cross-validation (mean AUC values are reported). Darker grey areas indicate better detection performance. n: polynomial degree; C: penalty value; γ: flexibility parameter for radial kernel function.

Table 12 – SVM parameter optimization

5 Results

The results are presented next, structured in accordance with the research goal outlined in section 3.3 and the related design and enhancing questions. Having presented the results covering the eight design and three enhancing questions, these will be compared with similar studies.

5.1 Results for Design Questions

The design questions cover tasks related to the setup of the fraud detection model. They help to understand how fraud can be captured best using the available data from annual reports and examine the performance of the final model in several dimensions like feature vector performance, sampling design implications, and classifier performance.

5.1.1 Size of the Qualitative Feature Vector

The results presented hereafter are organized in accordance with the design and enhancing questions in Figure 10. In the first step, the feature vector size is examined. Thus, this study tests the detection performance of different-sized qualitative features vectors to shed light on the influence of the number of clues from textual parts on the detection results. Presumably, larger vectors, which capture more clues based on the textual information, score better detection results (Cecchini et al., 2010a, p. 172). This is not only true for fraud detection but for pattern recognition in general (Zurada, 1992, pp. 13–17). From the perspective of crime detection theory, capturing a typically weak signal is difficult and requires a detector that can rely on numerous clues while being able to distinguish between signal and noise.[160] Therefore, the tendency can turn at some point, where adding additional features might actually reduce performance because the decisiveness of the vector is blurred by the sheer number of different clues, although it is hardly possible to determine at which point or even whether the effect will occur in the scope of this study's analysis. Accordingly, H1 was formulated as follows: larger qualitative feature vectors result in better detection performance.

The general procedure starts at the end of the feature extraction process, as described in section 4.1 and Figure 11. All possible n-grams of the fraudulent and non-fraudulent reports

[160] See section 2.4.1 for additional information on crime signal detection theory.

are generated from the observations of the training set (the procedure is carried out for every training set). To determine the size of the qualitative feature vector, which translates to the number of textual clues, an assessment of different vector sizes needs to be performed. Therefore, the discriminatory power of each of the qualitative features (each 5-gram) has to be assessed and the features ranked by their individual discriminatory power. In this study, the information gain ratio is used as the measurement of discriminatory power. The information gain ratio assigns a value to each feature that represents the feature's ability to reduce the level of entropy in the classification setting, with higher values representing better discriminatory power (e.g. Lee & Lee, 2006, p. 158; Goel et al., 2010, p. 33). The concept of information gain (IG) or its derivative, the information gain ratio (IGR), has found its way into the fraud detection literature and is used for feature selection purposes, among others (e.g. Cecchini et al., 2010a; Abbasi et al., 2012; Purda & Skillicorn, 2015).The number of features used in the detection models of similar studies varies considerably, such as 200 for Purda and Skillicorn (2015) and 10-1,200 for Cecchini et al. (2010a). Similarly to Cecchini et al. (2010a), this study tests the results for different amounts of features to determine the optimal vector size from a feature perspective, as a dynamic design with less static elements is preferable. The IGR is calculated for each feature of the training set, ranking the features from highest to lowest IGR.

The information gain for a single feature τ on a set of observations (data set) S is a measurement of the reduction of entropy, which is an indicator of impurity or disorder (Quinlan, 1986, p. 90).

$$IG(S,\tau) = E(S) - I(S,\tau) \tag{34}$$

The binary classification problem is characterized by fraudulent and non-fraudulent observations, with p_0 as the portion of non-fraudulent observations and p_1 as the portion of fraudulent cases. Entropy is calculated for the whole subset of observations as follows:

$$E(S) = -p_0 \log_2 p_0 - p_1 \log_2 p_1 \tag{35}$$

To assess performance on a feature level, the weighted average across the groups i needs to be calculated for each feature (τ).

$$I(S,\tau) = \sum_i \frac{|S_i|}{|S|} * E(S_i) \tag{36}$$

The basic concept of information gain is biased towards features that can take a lot of different values, which can be fixed using the information gain ratio instead (Quinlan, 1986, pp. 101–102). In the extreme case, a feature has a different value for each observation. The information gain ratio is calculated by dividing the intrinsic information (IntI) from the information gain.

$$IGR(S,\tau) = \frac{IG(S,\tau)}{IntI(S,\tau)} \tag{37}$$

The intrinsic information $IntI$ deals with the size of the subsets S_i that are created from splitting the data set S according to the values of a feature a.

$$IntI(S,\tau) = -\sum_i \frac{|S_i|}{|S|} \log_2\left(\frac{|S_i|}{|S|}\right) \tag{38}$$

In the next step, the feature vectors are created, starting with the top 100 and increasing the vector size in increments of 100 features. To test the influence of different vector sizes in line with the general research design of detecting cases from an unknown future, a random data set covering a 10-year timeframe is generated, divided in training and holdout sets. To avoid overfitting onto a fixed holdout set, five random subsets are drawn from the holdout set to assess the average performance. During the random sampling process, the class distribution is kept in line with the initial holdout set. The classifier applied here was an SVM.[161] The results are reported in Table 13 and indicate an upward trend, as expected when increasing the vector size. However, the trend seems to diminish and ultimately stop. For the matched sampling approach, the initial performance and the following increase is greater than for the realistic sampling approach. The results support H1 and are in line with the limited evidence of previous studies.

	Number of features									
	100	**200**	**300**	**400**	**500**	**600**	**700**	**800**	**900**	**1000**
m	0.653	0.688	0.721	0.735	0.750	0.753	0.760	0.764	0.770	0.770
r	0.757	0.768	0.768	0.767	0.782	0.784	0.790	0.792	0.793	0.792

The table reports AUC values for different sized qualitative feature vectors. m: matched sampling approach, r: realistic sampling approach.

Table 13 – Detection results for different-sized qualitative feature vectors

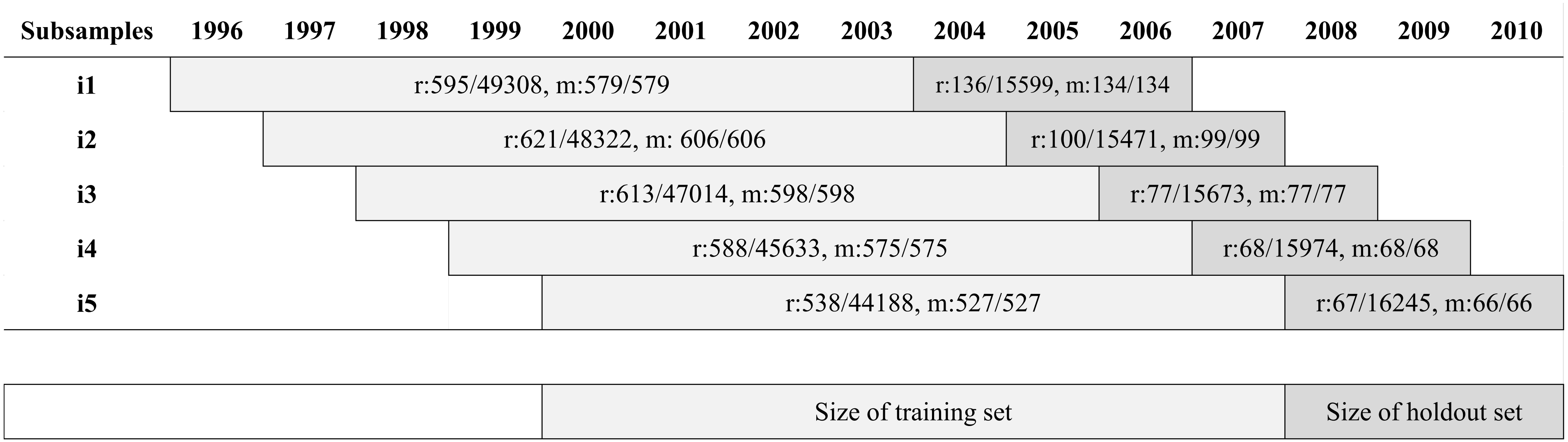
Cecchini et al. (2010a) have reported detection results for 11 different-sized vectors in a matched sampling setting (10, 20, 30, 40, 50, 100, 200, 300, 400, 500 and 1,200 features). Their results generally suggest that larger vectors are better suited to detecting anomalies,

[161] The parameter combinations of the machine learning techniques are assessed using a 5-fold cross-validation on a separate data set. This is done for all classifiers in accordance with the discussions in sections 4.2.7-4.2.9. Based on the hyperparameter optimization, SVMs are utilized as a classifier with good runtime and detection performance. A detailed test of the classifier is carried out in section 5.1.8.

with a vector comprising 500 features from narratives scoring the best results in the fraud detection context and the biggest vector scoring the best results in the bankruptcy detection context. A similar saturation effect is found in this study. For the following parts of the investigation, the qualitative feature vectors rely on the aforementioned results and comprise 1,000 features each, both for the realistic and matched sampling approach.

5.1.2 Intertemporal Stability of Qualitative Features

The following section examines the intertemporal stability of the qualitative features, before assessing the influence of time gaps on the detection results. The training sets represent the known universe of cases, which are used to extract the qualitative features and train the detection models. The holdout sets are constituted of observations from subsequent years and therefore contain unknown observations, which lie in an artificial future. The starting point of the analysis is the initial sample illustrated in Figure 25, which captures five rolling subsamples (i1-i5), each made up of an eight-year subset for feature extraction and model training. The detection models are tested on independent holdout sets from the subsequent three years of each training set.



r stands for the realistic sampling approach
m stands for the pair-matched sampling approach
The number of observations in each sample (i1-i5) is shown according to the sampling approach (realistic/matched).
This separation leads to a total of 20 subsets, with five training and five holdout sets for each sampling method.

Figure 25 – Initial sample

The basic feature generation process, presented in steps 2 and 3 of Figure 11, results in five qualitative feature vectors based on the narratives of the underlying annual reports for each training timeframe of the subsample i1-i5. Based on the previous results, the qualitative feature vectors comprise the 1,000 5-grams with the highest discriminatory power. To assess the intertemporal stability of the vectors, their composition across all five training sets is compared. Therefore, the 1,000 qualitative features of each subsample are compared to each other. Only exact matches, which concern the entire string, are regarded as consensus. The procedure is carried out both for the realistic and the matched sampling approach, as the feature extraction is conducted separately for both sampling designs due to differences in sample generation and the intention to control for size and industry effects. Each training set overlaps with the following set in six years and with the one after that in five years and so forth.

Subsamples	**i2**	**i3**	**i4**	**i5**
i1	18%/10%	31%/16%	42%/22%	54%/29%
i2		21%/9%	33%/17%	46%/23%
i3			21%/10%	36%/18%
i4				24%/10%

Time difference	2 years	4 years	6 years	8 years

Percentage of new features in the feature vectors extracted from the training sets for the matched/realistic approach.

i1: 1996-2003 i3:1998-2005 i5: 2000-2007
i2: 1997-2004 i4: 1999-2006

Figure 26 – Intertemporal stability of qualitative features

Figure 26 documents the changes in the composition of the qualitative feature vectors and is reinforced by the results of Brown et al. (2018), who have suggested comparable changes in the topics discussed in annual reports over a similar timeframe.[162] The results reveal a significant difference between realistic probability sampling and the pair-matched sampling approach, with the latter denoting that a greater portion of features changes for every subsample and time gap. For the largest time gap between the vectors of i1 and i5, more than

[162] The changes are even bigger when reducing the sample length, as can be seen in Figure 32 for the final sampling approach.

50% of the features are replaced, whereas in the realistic probability approach, less than 30% change from the same time gap. When exploring the magnitude of the changes from year to year, the one-year gap indicates the biggest change. The differences might be explained by the underlying sampling procedures: whereas the realistic sampling approach relies on every available report and therefore results in a greater variety of companies, pair-matching is carried out by identifying similar companies in terms of size and industry. This results in the absence of size- and industry-related predictors, which have already been identified as strong and consistent predictors regardless of the timeframe. Overall, these results support H2 and the assumption that predictors from the narratives change considerably over time.

5.1.3 Detection Results with Time Gaps

Having demonstrated the changes to the feature vectors over time, the influence on the detection results is tested. Therefore, the detection performance for all possible time gaps between the trained models and future holdout sets is assessed. According to H3 and based on the findings from the previous section, time gaps should cause a loss in detection quality due to the intertemporal instability of qualitative features. The test is carried out on the initial subsamples i1-i5. All possible time gaps between training sets and holdout sets are taken into consideration. The time gaps range from one to four years.[163] Figure 27 reports the respective AUC values.

Holdout **Training**	**2004-2006 (i1)**	**2005-2007 (i2)**	**2006-2008 (i3)**	**2007-2009 (i4)**	**2008-2010 (i5)**
1996-2003 (i1)	0.861	0.697	0.656	0.623	0.581
1997-2004 (i2)		0.807	0.710	0.643	0.529
1998-2005 (i3)			0.750	0.627	0.591
1999-2006 (i4)				0.746	0.598

Time gaps	No time gap	1 year	2 years	3 years	4 years

AUC values for detection performance with time gaps between training and holdout sets.

Figure 27 – Detection results with time gaps

[163] The classifier was an SVM.

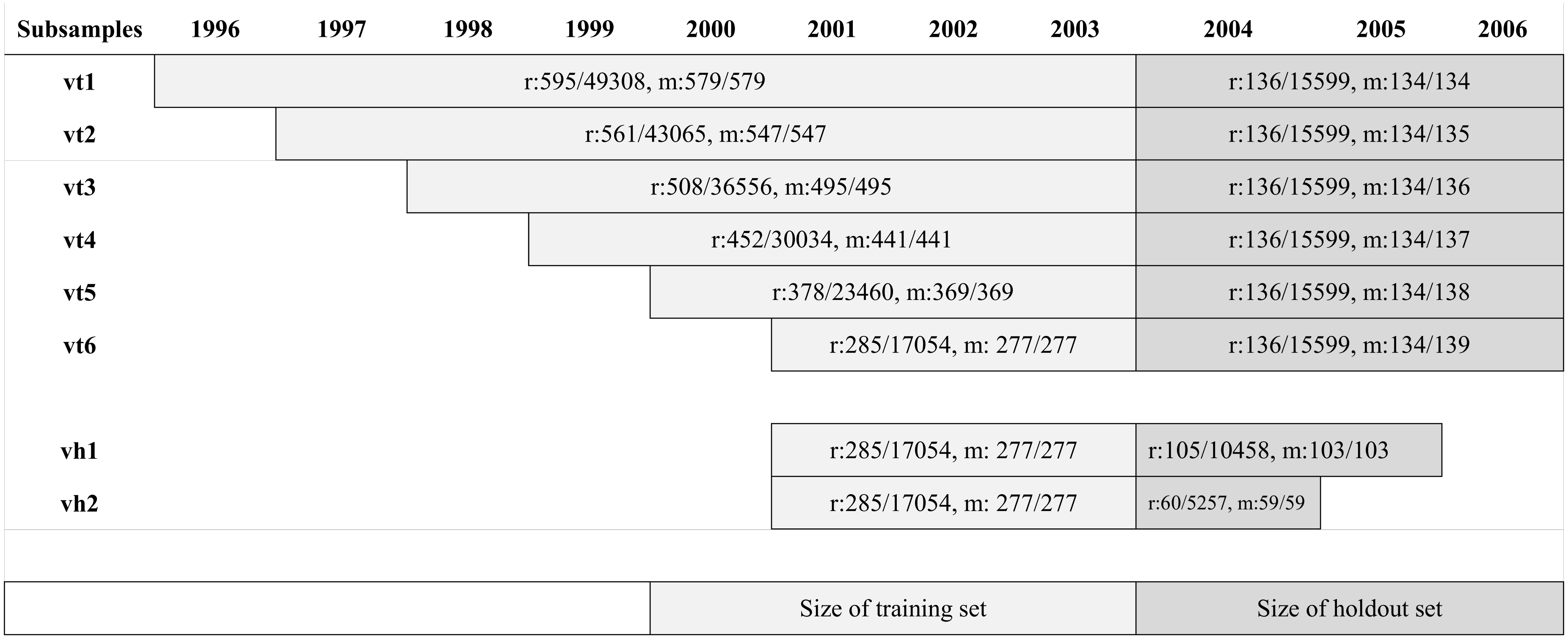
The results indicate a considerable drop in the AUC values for the one-year gap for all models. Afterwards, detection performance becomes more stable, with smaller declines in the AUC for the following gaps. To better understand the initial drop in the first year, the trend in Figure 25 should be mentioned, where the biggest change in the composition of the feature vectors can be found for the first year gap, regardless of the starting point (subsample). This initial drop hints at the relevance of qualitative features capturing time-dependent textual patterns, as bigger time gaps result in comparatively smaller drops in predictive power. The results reveal the importance of the topicality and actuality of features when constructing detection models, which rely on bag-of-words and support H3.

5.1.4 Model Size Implications

With the aforementioned results in mind, in the next step an understanding of the influence of sample sizes for model building (especially with regard to training and holdout set sizes) on the detection results as mentioned in design question 4 will be examined. Given that the detection models of all studies presented so far encompass a rather arbitrarily chosen number of years for training and detection purposes, this study seeks to shed light on the consequences of changes to the sampling design on the results. The basic assumptions are again related to changes to the narratives over time and the indications that were received via the prior results. Larger training sets would cover more years with more cases from which the detection models are trained. However, larger training sizes may not *per se* yield better detection quality. As they grow in length, the time difference between early cases of the training set and later cases of the holdout set cover a significant time span. For example, the training of a detection model with data from 10-year-old cases might lead to suboptimal results due to changes in predictor characteristics. Combined with the abovementioned changes to the narratives in general and the assumption that fraud is a less generalizable concept and is constantly changing, this may become a problem when relevant predictors are replaced over time or when the characteristics of fraudulent/non-fraudulent observations change. It may furthermore be suggested that under a fixed vector size, which for most studies is the typical research design approach, the number of features per clue decreases, resulting in a suboptimal representation. Moreover, in a real-world application, testing the annual reports of the current year would translate into a design where the detection performance of the models developed is tested on a one-year holdout set comprising the observations from the subsequent period. Keeping the aforementioned discussion and the explorative nature of the question in mind, making an assumption about the outcome is rather difficult.

The analysis consists of two parts in which the sizes of the training and the holdout sets are altered separately. In the first step, the size of the training set is reduced from the initial eight years in one-year intervals to three years. In the next step, the size of the previously determined training length is fixed and the holdout set is altered from the initial three years to a one-year length. The procedure can be seen in Figure 28. To avoid overfitting onto the holdout set, we draw five random subsets for each iteration and report the average AUC values. During the random sampling process, the class distribution is kept in line with the initial holdout set. Figure 28 presents the underlying subsamples for the examination of design question 4. To ensure a comprehensive insight, the test is conducted across all feature vectors for both sampling approaches.[164]

[164] Like for the previous tests, the classifier was an SVM.



r stands for realistic sampling approach
m stands for matched sampling approach

Figure 28 – Varying training and holdout sizes

The results depicted in Table 14 reveal consistent detection quality measured by the AUC regardless of the length of the training set. The results appear to be robust for the different feature vectors and for both sampling approaches.

	Subsample	Length	Qual.	Quant.	Qual. & quant.	Mean
Realistic	**vt1**	8	0.831	0.656	0.861	0.783
	vt2	7	0.823	0.656	0.852	0.777
	vt3	6	0.826	0.655	0.853	0.778
	vt4	5	0.820	0.654	0.849	0.774
	vt5	4	0.818	0.655	0.844	0.772
	vt6	3	0.838	0.651	0.868	0.786
Matched	**vt1**	8	0.740	0.617	0.768	0.708
	vt2	7	0.743	0.617	0.770	0.710
	vt3	6	0.737	0.625	0.787	0.716
	vt4	5	0.742	0.621	0.761	0.708
	vt5	4	0.776	0.643	0.842	0.754
	vt6	3	0.786	0.602	0.799	0.729
	Mean		0.790	0.638	0.821	

The table reports AUC values for varying subset sizes of the training set.

Table 14 – Detection results for varying training lengths

However, when decreasing the size and therefore the number of years of the holdout set in the second step, detection performance increases. This holds for both the realistic and the matched sampling approach, as indicated in Table 15. When comparing the different feature vectors, those with qualitative features show a considerable increase in the AUC, whereas the quantitative feature vector in the matched sampling approach hardly changes. The only decrease in detection performance is found for the quantitative feature vector under the realistic sampling approach. The mean AUC values across the three feature vectors reveal an upward trend, regardless of the one declining anomaly, which sets the final sampling approach for answering the remaining design questions at a training set length of three years and a holdout set length of one year. Regarding real-world applicability, a one-year holdout set represents the cases of the current year that need to be tested and hence offers greater insights for practitioners.

The tests for this design question are carried out for one point in time, meaning that the artificial future always starts with the year 2004. Therefore, the results are not controlled for time-dependent changes that might occur when choosing a different timeframe. Testing at different points in time can reveal disparate implications for the model set-up. However, the examination offers some sort of rationale to determine the subsample sizes, without the rather arbitrary decisions of previous studies and with the 3/1 split (three years of training, one year of testing), representing a design that captures the practical relatability and understandability.

	Subsample	Length	Qual.	Quant.	Qual. & quant.	Mean
Realistic	**vt6**	3	0.838	0.651	0.868	0.786
	vh1	2	0.888	0.597	0.891	0.792
	vh2	1	0.901	0.597	0.907	0.802
Matched	**vt6**	3	0.786	0.602	0.799	0.729
	vh1	2	0.782	0.590	0.796	0.723
	vh2	1	0.842	0.589	0.847	0.760
	Mean		0.840	0.604	0.851	

The table reports AUC values for varying lengths of the holdout set.

Table 15 – Detection results for varying holdout lengths

5.1.5 Detection Results over Time

The final detection approach, which builds upon the previous findings and the outcomes of design questions 1 to 4 along with their respective hypotheses, will be discussed in the following sections. Figure 29 presents the rolling subsamples over the timeframe from 1996 until 2010 that result from design question 4. Each of the 12 subsamples consists of a three-year-long subset, representing the timeframe in which the qualitative and quantitative features are extracted and the models are trained. The subsequent year of each training set is used as a holdout set for detection purposes. The number of observations in each subset f1 to f2 varies considerably, especially regarding the number of fraudulent cases, which additionally influences the relative frequency of occurrence of fraud in each subset. In the holdout sets, the relative frequency ranges between around 0.4% (22 fraudulent observations) and 1.8% (111 fraudulent observations).

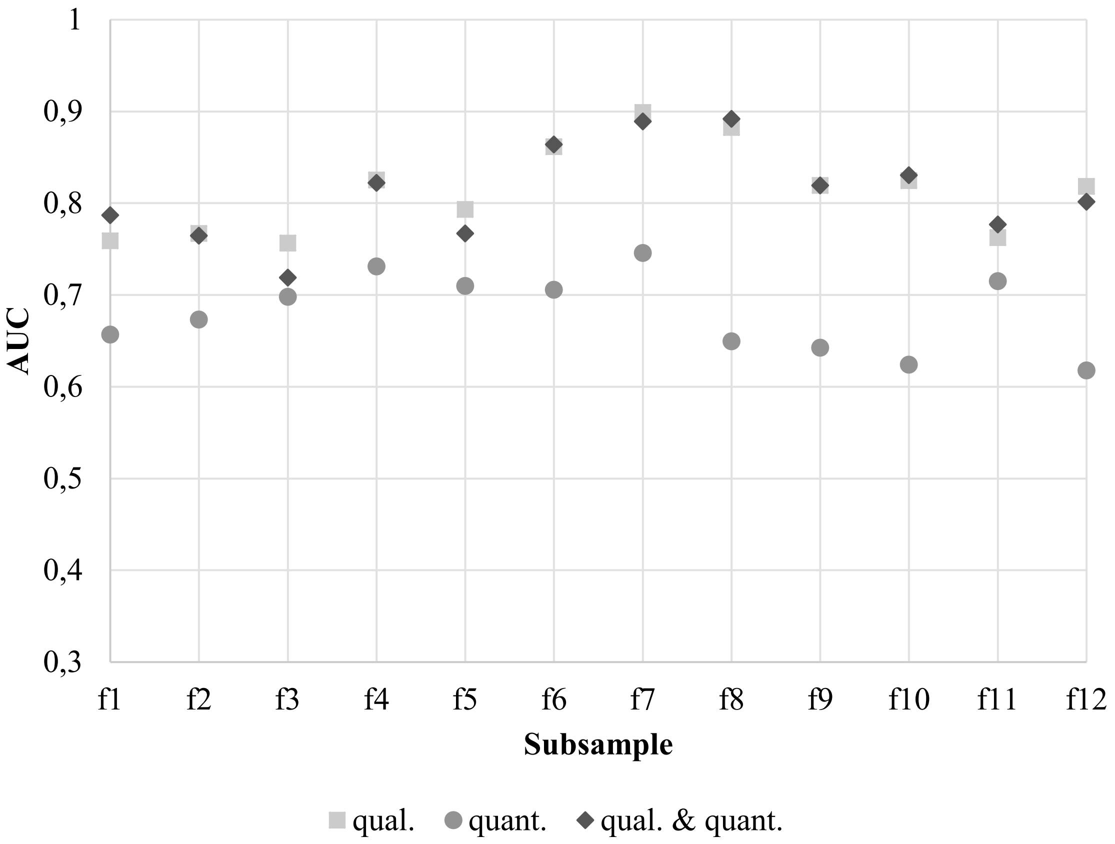
Sample	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
f1	r:143/19274, m:140/140			r:74/6574, m:72/72											
f2		r:183/19605, m:178/178			r:93/6407, m:92/92										
f3			r:223/19502, m:218/218			r:111/6011, m:107/107									
f4				r:278/18992, m:271/271			r:89/5712, m:87/87								
f5					r:293/18130, m:286/286			r:85/5254, m:83/83							
f6						r:285/17060, m:277/277			r:60/5257, m:59/59						
f7							r:234/16306, m:229/229			r:45/5201, m:44/44					
f8								r:190/15795, m:186/186			r:31/5141, m:31/31				
f9									r:136/15599, m:134/134			r:24/5129, m:24/24			
f10										r:100/15471, m:99/99			r:22/5403, m:22/22		
f11											r:77/15673, m:77/77			r:22/5442, m:22/22	
f12												r:68/15974, m:68/68			r:23/5400, m:22/22
								Size of training set				Size of holdout set			

r stands for the realistic sampling approach
m stands for the pair-matched sampling approach

Figure 29 – Final sample

A small number of positives (actual fraudulent observations) may negatively influence the expressiveness of the AUC, which depicts the ratios of true positive rates and false positives rates over the discriminatory thresholds (Hanczar et al., 2010, p. 829). Smaller absolute and relative frequencies of occurrence of one group may lead to larger steps per TPR/FPR ratio, distorting the overall AUC. However, as the sampling represents the closest approximation of the true population of fraudulent and non-fraudulent observations and the feature vectors and classifiers are compared on common ground, this problem should hardly affect the overall conclusiveness of the results.

Figures 30 and 31 depict the highest AUC values for each of the subsamples f1-f12 across all feature vectors and classifiers for the matched and realistic sampling approaches. An AUC of 0.5 represents the detection performance of a random classifier, indicating that every feature vector is able to perform above the minimum margin across all subsamples. So far, the detection results of previous studies seem to be rather static and commonly have only been tested for certain selected points in time. However, given that this study strives to build a reliable and robust detection model capable of detecting fraudulent cases at any given time, the results are tested on the 12 different subsamples over the rolling 15-year timeframe and for both sampling approaches.

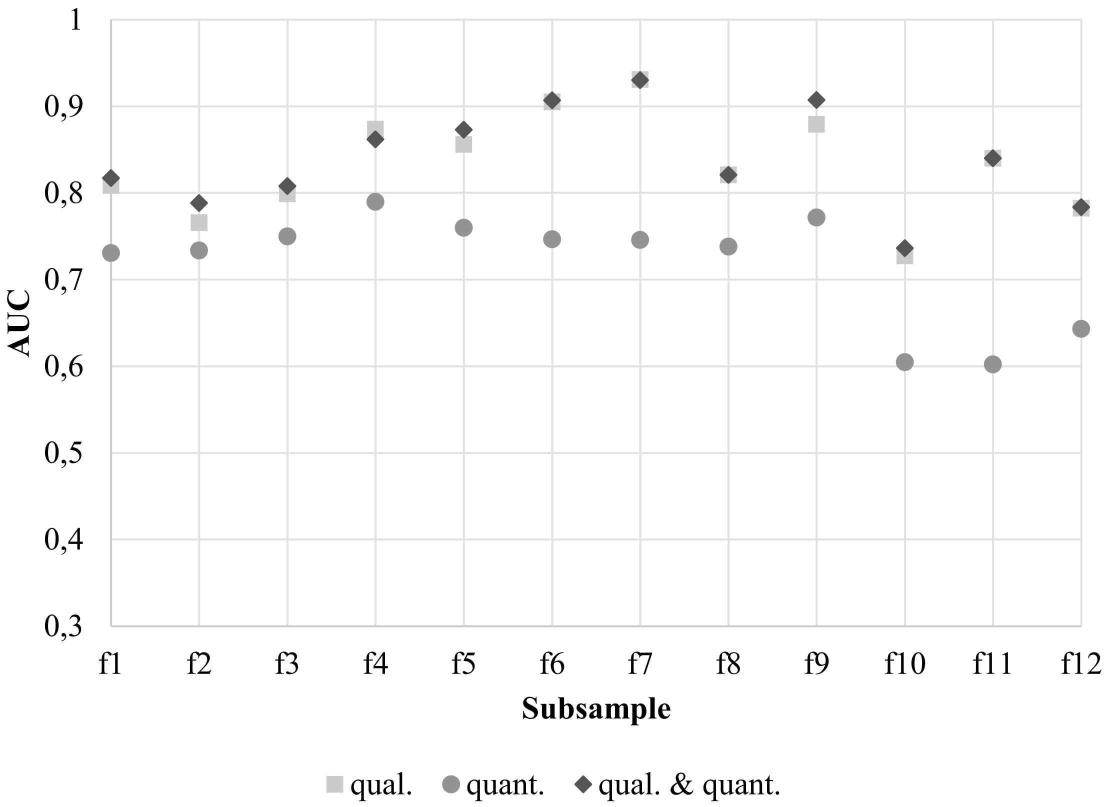


Subsample	Training	Holdout	Subsample	Training	Holdout
f1	1996–1998	1999	f7	2002–2004	2005
f2	1997–1999	2000	f8	2003–2005	2006
f3	1998–2000	2001	f9	2004–2006	2007
f4	1999–2001	2002	f10	2005–2007	2008
f5	2000–2002	2003	f11	2006–2008	2009
f6	2001–2003	2004	f12	2007–2009	2010

The training includes the model generation process with individual feature extraction and model training. The holdout set comprises unknown observations from the subsequent period.

Figure 30 – Results for pair-matched sampling approach

The results reveal a consistent detection quality for all feature vectors across the matched and realistic sampling approaches. Throughout the analysis, the results for both sampling approaches share a similar trend, with the best performances being achieved on the holdout sets during the middle of the timeframe, while earlier and later years are seemingly more difficult for the models. Assuming that an AUC of 0.8 represents very good detection performance, the qualitative and combined feature vector score an average AUC of above 0.81 (stdev. 0.05) in the matched sampling approach, while the quantitative feature vector reaches only an average value of about 0.68 (stdev. 0.04) across the entire timeframe.



Subsample	Training	Holdout	Subsample	Training	Holdout
f1	1996–1998	1999	f7	2002–2004	2005
f2	1997–1999	2000	f8	2003–2005	2006
f3	1998–2000	2001	f9	2004–2006	2007
f4	1999–2001	2002	f10	2005–2007	2008
f5	2000–2002	2003	f11	2006–2008	2009
f6	2001–2003	2004	f12	2007–2009	2010

The training includes the model generation process with individual feature extraction and model training. The holdout set comprises unknown observations from the subsequent period.

Figure 31 – Results for the realistic sampling approach

For the realistic sampling approach, a similar result can be observed. Both the qualitative and combined feature vectors score average AUC values of above 0.83 (stdev. 0.06), representing a very good detection performance. The quantitative vector also scores good results and an average AUC of 0.72 (stdev. 0.06).

Overall, the results indicate that the models are able to score very good detection performances across the overall timeframe and at no point in time drop under an unreasonable margin. The drop in detection performance resulting from time gaps between feature extraction, model training, and fraud detection (as discussed in design question 3) can be mitigated through the use of a design that includes regular model updates. The aforementioned trend in detection performance may be seen as a regular variation due to different cases, rather than as a potential deteriorating performance for later periods.

5.1.6 Feature Vector Performance

The previous section has presented the baseline outcome of the different feature vectors for the entire timeframe and for the matched and realistic sampling approaches. This section will examine and highlight individual performance on a less aggregated level. Thus, hypotheses H4a and H4b regarding the feature vector performance are tested. Paired t-tests are conducted to compare the detection outcomes based on the 12 subsamples and the classifiers utilized (e.g. Abbasi et al., 2012, p. 1311 for the comparison of different feature sets or p. 1313 for different classifier peculiarities).

First, the performance of the quantitative feature vector is compared to those incorporating qualitative features in order to test hypothesis H4a. Quantitative features are assumed to be less suitable for detecting financial statement fraud. Empirical evidence from other studies has already hinted at their baseline poor detection performance relative to approaches that exploit other data types (e.g. Dong et al., 2016). Previous results have implied the potential confirmation of this hypothesis, albeit under a limited test bed (e.g. Figure 30, average AUC for quant. 0.781, qual. 0.832, qual. & quant. 0.839). Table 16 reports the p-values from paired t-tests for the entire set of results (n=48; 4 classifiers, 12 subsamples). [165]

Realistic	
	Quant.
Qual. & quant.>	<0.001
Qual.>	<0.001
Matched	
	Quant.
Qual. & quant.>	<0.001
Qual.>	<0.001

p-values for paired t-tests comparing results of the quantitative feature vector to the qual. and qual. & quant. feature vectors

Table 16 – Results of t-tests for quantitative feature vectors

[165] The p-values are one-tailed for evaluating the alternate hypothesis of better performance of feature vectors constructed from qualitative features. The “>” sign indicates the respective direction. A similar notation will be chosen in the reporting of the results in subsequent sections.

The results support hypothesis H4a and show significantly better detection performance for the feature vectors that rely on extracted patterns from the narratives. The results also hold when comparing detection outcomes on the level of individual classifiers, as can be seen in Table 17.

Realistic		**Matched**	
SVM		**SVM**	
	Quant.		**Quant.**
Qual.>	<0.001	**Qual.>**	<0.001
Qual. & quant.>	<0.001	**Qual. & quant.>**	<0.001
ANN		**ANN**	
	Quant.		**Quant.**
Qual.>	0.003	**Qual.>**	<0.001
Qual. & quant.>	<0.001	**Qual. & quant.>**	<0.001
KNN		**KNN**	
	Quant.		**Quant.**
Qual.>	<0.001	**Qual.>**	<0.001
Qual. & quant.>	0.089	**Qual. & quant.>**	0.039
NB		**NB**	
	Quant.		**Quant.**
Qual.>	0.020	**Qual.>**	0.011
Qual. & quant.>	<0.001	**Qual. & quant.>**	0.001

p-values for paired t-tests comparing detailed results of the quantitative feature vector to the qual. and qual. & quant. feature vectors for individual classifier

Table 17 – Results of t-tests for quantitative feature vectors per classifier

The p-values in Table 17 report significantly better detection performance for the text-based feature vectors on the 10% level for every possible combination and for both sampling approaches.

In the next step, hypothesis H4b, which is concerned with detection performance improvements with a combination of qualitative and quantitative features, is tackled. Table 18 reports the p-values of two-sided t-tests (n=12), comparing the results of the combined feature vector (qual. & quant.) to those scored by the baseline feature vectors constructed from only one data type for each individual classifier across the 12 subsamples.

Realistic			**Matched**		
SVM			**SVM**		
	Qual.	**Quant.**		**Qual.**	**Quant.**
Qual. & quant.>	0.010	<0.001	**Qual. & quant.>**	0.297	<0.001
ANN			**ANN**		
	Qual.	**Quant.**		**Qual.**	**Quant.**
Qual. & quant.>	0.018	<0.001	**Qual. & quant.>**	0.027	<0.001
KNN			**KNN**		
	Qual.	**Quant.**		**Qual.**	**Quant.**
Qual. & quant.>	1.000	0.089	**Qual. & quant.>**	0.998	0.039
NB			**NB**		
	Qual.	**Quant.**		**Qual.**	**Quant.**
Qual. & quant.>	0.793	<0.001	**Qual. & quant.>**	0.953	0.001

p-values for paired t-tests comparing results of combined feature vectors to the baseline feature vectors for individual classifier

Table 18 – Results of t-tests for combined feature vectors

On average, for the previously reported results, the combined feature vector scores an AUC of 0.839, which is slightly better than the qualitative vector (AUC of 0.832) and considerably better than the quantitative vector (AUC of 0.718) (see Figures 30 and 31 for a similar result using the matched sampling approach). Moreover, the combined feature vector scores the highest result for the subsample ranging from 2002 until 2005, with an AUC of 0.939. However, the comparison between the combined feature vector and the baseline vectors on the level of the individual classifier depicts rather mixed results. Despite the fact that the combined vector scores superior results against the quantitative vector of the matched and the realistic sampling approaches (10% level), the results of the qualitative feature vector are less conclusive. For the more sophisticated classifier like SVMs and ANNs, the combined vector scores significantly better results across the 12 subsamples than for KNN or NB. Especially for the realistic probability of fraud sampling, which is closest to a real-world application, the combined feature vector seems to be slightly superior. Due to the mixed results, hypothesis H4b cannot be confirmed in its entirety but rather with a constraint regarding the classifier choice and the sampling approach.

The rather small jump in additional detection performance after the combination of both data types may be explained by the relative individual discriminatory power of the quantitative features in the combined feature vector, which will be highlighted further in section 5.2.1.

5.1.7 Matched and Realistic Sampling

In this section, the results of the matched and realistic sampling approaches are discussed. In contrast to the majority of previous research, this study has relied on both sampling approaches to produce robust and comparable evidence regarding the effectiveness of fraud detection models. In terms of hypothesis H5, the results should be lower under pair-matching, due to limitations regarding industry- and size-related predictors and the overall more complicated setting related to the less decisive characteristics of the observations. Despite these constraints, a sound detection model should still be able to distinguish reliably between fraudulent and non-fraudulent observations over the respective timeframe. Therefore, the additional sampling approach also serves the purpose of testing robustness to a certain extent. The difference is tested using paired t-tests.

	SVMs		
	Qual.	**Quant.**	**Qual. & quant.**
Realistic>	0.078	0.132	0.061
	ANN		
	Qual.	**Quant.**	**Qual. & quant.**
Realistic>	0.894	0.007	0.691
	KNN		
	Qual.	**Quant.**	**Qual. & quant.**
Realistic>	0.085	<0.001	<0.001
	NB		
	Qual.	**Quant.**	**Qual. & quant.**
Realistic>	0.026	0.004	0.002

p-values for paired t-tests comparing results of the realistic sampling approach to the matched sampling approach across feature vectors and classifier

Table 19 – Results of t-tests for matched and realistic sampling approaches

The detection performance across all feature vectors and classifiers (n=144: 12 subsamples, three feature vector, four classifiers) is significantly better (p<0.001) for the realistic sampling approach compared to the matched sampling approach. Table 19 presents the p-values resulting from t-tests comparing isolated feature vector and classifier results between the matched and realistic sampling approaches (n=12). In nine out of 12 cases, the results in the realistic sampling context are significantly (10% level) better. With regard to the results in 5.1.5 and the detection performance over time, the realistic feature vector achieves significantly better results for each of the three feature vectors at the 10% level for the highest single-scoring outcomes, overall confirming hypothesis 5.

Besides the expected inferior outcome, the matched sampling approach on average results in a 5% drop in detection performance. For a potent classifier like an SVM, this drop does not reduce the results under an AUC of 0.8, which still depicts very good detection performance. A similar trend can be postulated for the other classifiers as well, indicating that the detection performance holds for the matched sampling approach, albeit leading to the expected small but significant decline.

The results are assumed to be in line with findings from the literature, although evidence is scarce and hardly comparable. Purda and Skillicorn (2015) have conducted pair-matching besides a rather realistic sampling (less strictly balanced) approach and have reported AUC values for both, based on a logit model, which was used to estimate the coefficients of independent variables capturing different textual peculiarities. Similarly to this study, detection performance under pair-matching was lower, indicating difficulties in distinguishing between fraudulent and non-fraudulent observations in an artificial setting.

5.1.8 Classifier Performance

So far, each of the four distinct classifiers has been used to detect fraudulent observations in the matched and realistic sampling approaches for each of the 12 rolling subsamples covering the 15-year timeframe. This section will highlight the performance of the individual classifiers in the context of three different feature vectors in order to determine which is best suited for each task. The classifiers represent different levels of complexity and computational effort. It is the goal of this study to develop an effective fraud detection approach that can attain reliable and robust results and that can be run (including training, validation and testing) in reasonable time on contemporary and universal hardware.[166] According to hypothesis H6a, more sophisticated classifiers like ANNs and SVMs should generally score better results than KNN and NB. Moreover, H6b postulates that differences regarding the performance of each classifier for individual feature vectors due to underlying data peculiarities can be assumed. Table 20 presents the detection performance measured by the AUC for each classifier in the realistic and matched sampling approaches and for every feature vector. The results depict the average detection performance across the 12 subsamples.

[166] The major part of the analysis has been run on an Intel i5-4210m @ 2.6 GHz with 8 GB DDR3 RAM and an Intel Xeon Silver 4116 @ 2.1 GHz with 100GB DDR3 RAM. No GPU-centered architecture has been utilized.

Sampling	Classifier	Qual.	Quant.	Qual. & quant.	Mean
Realistic	SVN	0.826	0.680	0.834	0.780
	ANN	0.759	0.651	0.781	0.731
	KNN	0.792	0.705	0.717	0.738
	NB	0.676	0.632	0.663	0.657
Matched	SVM	0.802	0.653	0.805	0.753
	ANN	0.785	0.565	0.792	0.714
	KNN	0.749	0.640	0.660	0.683
	NB	0.636	0.561	0.591	0.596

The table reports average AUC values for each classifier across all feature vectors.

Table 20 – Average AUC values for classifier and feature vectors

Based on Table 20, support vector machines score the best results, followed by artificial neural networks and k-nearest neighbour. Naïve Bayes seems to be less potent in the fraud detection setting of this study, although it is still more capable of distinguishing between fraudulent and non-fraudulent observations than a random classifier (AUC=0.5). To further clarify the outcome, paired t-tests were conducted to compare the performance of the different classifiers against one another. Table 21 presents the p-values for paired t-tests for the entire non-aggregated results (n=36; 12 subsamples, three feature vectors).

Realistic		**ANN**	**KNN**	**NB**
	SVM>	<0.001	<0.001	<0.001
		KNN	**NB**	
	ANN>	0.684	<0.001	
		NB		
	KNN>	<0.001		
Matched		**ANN**	**KNN**	**NB**
	SVM>	<0.001	<0.001	<0.001
		KNN	**NB**	
	ANN>	0.043	<0.001	
		NB		
	KNN>	<0.001		

p-values for paired t-tests comparing results of the realistic sampling approach to the matched sampling approach for different classifiers

Table 21 – Results of t-tests for all feature vectors per classifier

In addition to confirming the superior performance of SVMs and the inferior performance of NB, KNN and ANNs are found to occupy the middle ground. For the pair-matched sampling, KNN and ANNs score very similar results, while ANNs are slightly ahead in the case of the matched sampling approach. With regard to the superior performance of SVMs and the case that ANNs are on par with KNN in the realistic approach but potentially ahead (5% significance level) in terms of matched sampling, H6a can be partially confirmed.

However, the results vary when only highlighting particularly isolated feature vectors. Although SVMs outperform every other classifier in the matched sampling approach and for feature vectors incorporating qualitative features in the realistic sampling approach, they are not superior to KNN when only quantitative features are used. Table 22 presents the p-values to compare KNN against the remaining classifiers for the quantitative feature vector.

	SVM	**ANN**	**NB**
KNN>	0.014	0.018	0.001

p-values for paired t-tests comparing results of the KNN classifier to the remaining classifiers for the quantitative feature vector

Table 22 – Results of t-tests for quantitative feature vectors and KNN

Clearly, KNN performs significantly better here (5% significance level). Thus, when restricted to quantitative data, KNN in a realistic probability of fraud environment may outperform more sophisticated classifiers like SVMs or ANNs. Regarding hypothesis H6b, the results reveal that SVMs score best for scenarios in which textual information is utilized for fraud detection purposes, whereas financial metrics are seemingly better under KNN, supporting the hypothesis of classifier dependent feature vector performance. Overall, when seeking to compare the results of this section to similar studies, a certain degree of overlap regarding the relative performance of SVMs, ANNs and NB can be found with Abbasi et al. (2012). Indeed, these authors have found varying performances of classifiers across different feature vectors compositions, with different types of support vector machines achieving relatively good results in each of them.

As a side note, the two more sophisticated approaches using the support vector machines and the artificial neural networks represent state-of-the-art approaches in the machine learning domain. Under the prerequisites of reasonable training, validation and testing effort on current hardware, the artificial neural network may underperform and not reach its full potential. This may additionally be influenced by the limited hyperparameter optimization procedure. The artificial neural network cannot reach the detection performance of the SVMs, although it nearly reaches very good results (average AUCs above 0.8). Despite having the highest detection performance, SVMs are also comparably fast, meaning that they should be the go-to approach based on this study’s results. This may also account for the extensive use of SVMs in the fraud detection literature, especially when qualitative features are incorporated.

5.2 Results for Enhancing Questions

The enhancing questions comprise of tasks to improve the results further and increase comprehensibility. However, before discussing the related outcomes, the section starts with an evaluation of the qualitative and quantitative features, which helps to understand the previous results and opens the possibility to relate the models to comparable studies.

5.2.1 Characteristics of Feature Vectors

Before diving into enhancing questions 1-3, it may be helpful to examine the qualitative features across the final subsample (f1-f12) in greater detail. Firstly to back up the findings of the intertemporal stability of qualitative features discussed in design question 2 and secondly to shed additional light on the characteristics of the extracted patterns based on 5-grams, the foundation of the previous results. Therefore, the changes in qualitative feature vector composition will be examined and the reoccurring features identified, before exploring the distribution of IGR values that represent the discriminatory power of individual features over time. The sampling approaches, in this regard the relative performance of quantitative features, will also be discussed. Finally, the qualitative feature vectors of this study will be compared to those from other studies, analyzing similarities and differences in the extracted patterns from the textual parts of corporate disclosure.

To support the previous findings, the composition (5-grams) of the qualitative feature vectors of different subsamples is compared. The time gaps range from single years to a maximum of 11 years when comparing the qualitative features of the first and last vectors from the respective subsamples. As already explained, the feature vectors are always constructed from training sets, capturing a timeframe of three years and consisting of 1,000 qualitative features. Figure 32 presents the percentage of new features in the vectors of the subsamples.

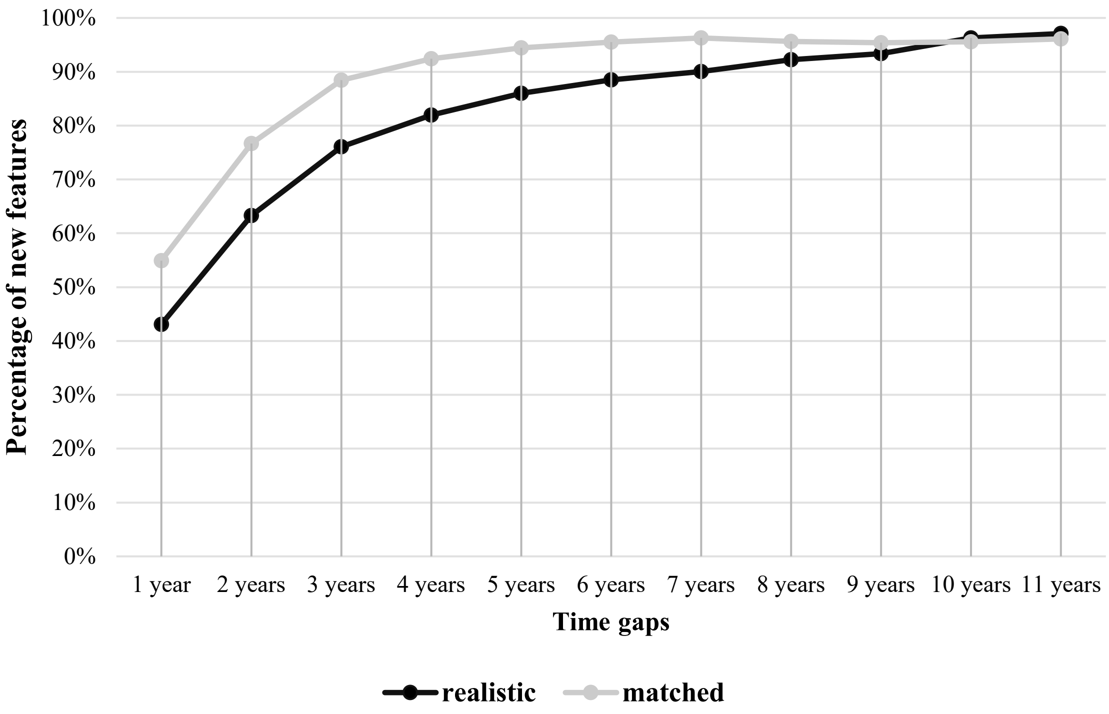


Figure 32 – Temporal stability of features for final sample

The results confirm the previous findings and present similar differences between the realistic probability and pair-matched sampling approaches, with the latter revealing a greater proportion of features that change for every subsample and time gap. However, the extent to which the features change is significantly larger here than compared to the eight-year-long subsamples. For both sampling approaches, around half of the features (43% for realistic and 55% for matched) are replaced after one year, reaching a maximum of almost 96% after six years in the case of the matched sampling approach and 10 years for the realistic sampling approach. When examining the magnitude of the changes from year to year, the one-year gap indicates the largest change, followed by a diminishing trend, in line with previous results.

In the next part, the differences between the qualitative feature vectors are examined in detail to further examine the differences between the realistic and matched sampling approaches. The IGR represents the discriminatory power of each qualitative feature and is calculated for every feature in the entire sample. Figures 33 and 34 present the distribution of the discriminatory power, divided into 10 bins, each consisting of 100 features, ranked from highest to lowest for each of the final subsamples f1-f12.

The differences between the sampling approaches are obvious. The realistic sampling approach seems to rely on a few qualitative features with high individual discriminatory power, whereas the biggest part of the vector possesses comparatively low individual power. In contrast, features from the matched sample seem to be rather stable concerning their individual potential of reducing entropy. For both sampling approaches, the trend is consistent over time, as represented by subsamples f1-f12. The phenomenon might be explainable by the abovementioned linguistic variety that may be required in the case of the matched sampling to distinguish between fraudulent and non-fraudulent observations, without being able to rely on size- or industry-related clues as in the case of the realistic approach. Furthermore, this may explain the differences in detection performance being lower for the matched sampling approach due to the aforementioned difficulties in classifying observations owing to a lack of truly decisive features.

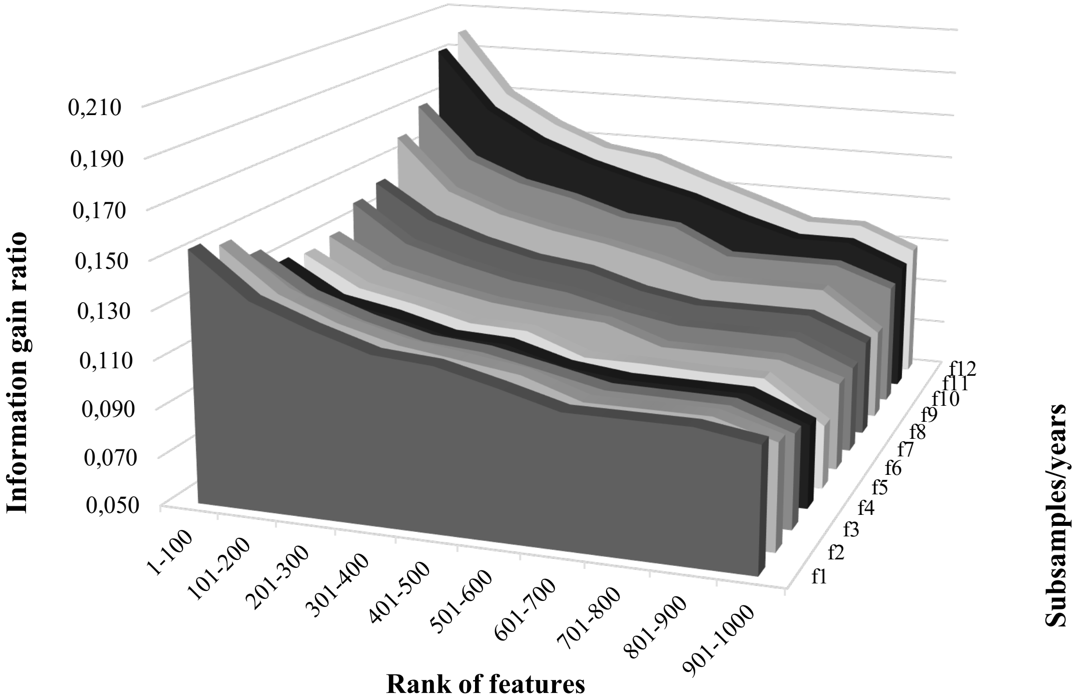


Figure 33 – Discriminatory power of qualitative features for matched sampling

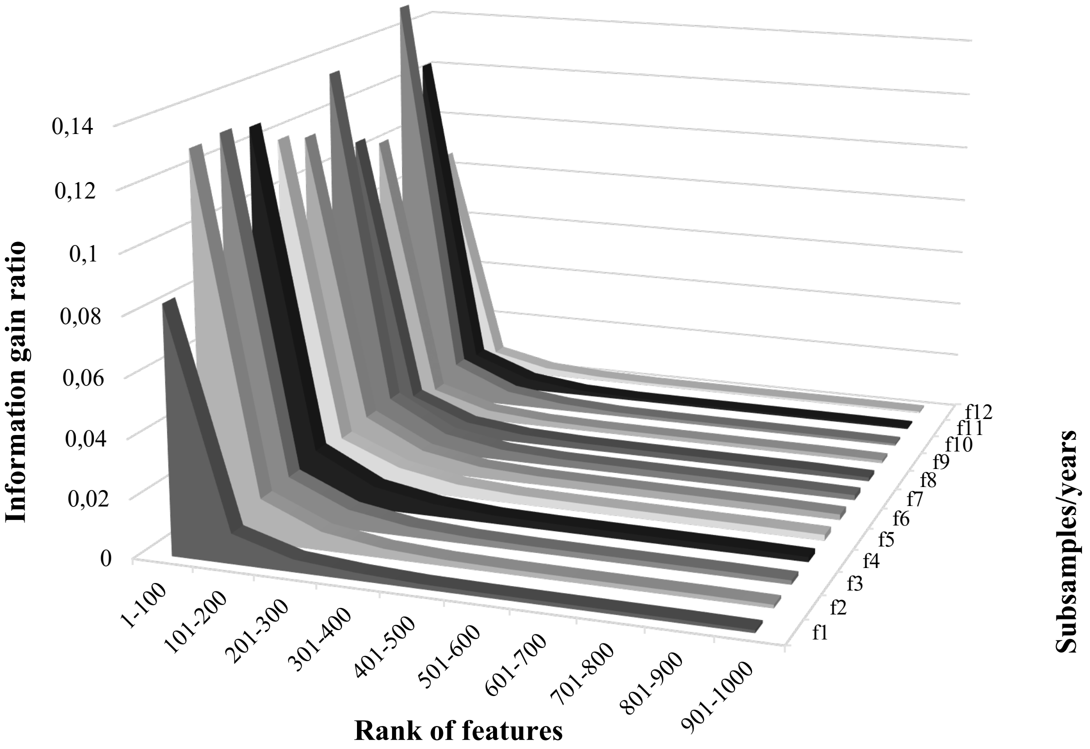


Figure 34 – Discriminatory power of qualitative features for realistic sampling

Before evaluating the consistency of the textual predictors, it is necessary to highlight the value of the quantitative features compared to the qualitative features in order to shed additional light on the previous results, especially regarding feature vector performance and therefore the overall relevance of quantitative features in the combined feature vector. Table 23 displays the average ranks of the quantitative features in the combined feature vector across the 12 subsamples.

The combined vector comprises 1,000 5-grams and 19 quantitative financial metrics. For the realistic sampling approach, the highest individual ranking features are the size proxy, followed by a variable capturing capital structure and performance. Despite their size, the first eight features are relatively stable, as are the bottom five features. The quantitative features seem to be rather unimportant compared to the qualitative features, with only one being located in the top 200 bracket, which might explain the rather small jump in detection performance when combining both types and the relatively bad detection performance for only quantitative features.

	Realistic			Matched	
Feature	Mean	Stdev.	Feature	Mean	Stdev.
logat	184	47	reta	72	74
tlta	251	7	tlta	99	29
reta	257	7	cogssal	184	39
opxsal	258	7	gp	185	39
opisal	268	8	opxsal	186	39
cogssal	283	6	opisal	187	39
gp	284	6	arta	264	71
arta	293	6	ietl	537	249
salta	349	23	nisal	538	58
ietl	423	275	salta	539	166
btm	453	50	logat	871	99
salar	819	230	accr	993	8
nisal	937	40	salar	996	7
de	988	40	btm	998	6
nita	1,011	2	invsal	999	7
invta	1,013	4	into	1,003	7
accr	1,014	1	nita	1,004	5
invsal	1,015	2	de	1,006	6
into	1,016	3	invta	1,007	8

Table 23 – Average rank of quantitative features

For the matched sampling, it would appear that even fewer quantitative features are relevant. Interestingly, despite the fact that size is one of the pair-matching variables, it scores better than eight other features in the analysis of individual discriminatory power. Comparing the results to similar studies is rather difficult, as related feature selection techniques – or feature selection techniques at all – have rarely been utilized in the case of quantitative features so far. Chen (2016) has tested two feature selection algorithms (with similar IGR measures) for the purpose of financial statement fraud detection and listed the top ranking quantitative features. The results suggest that out of 30 financial and non-financial predictors, only four for the first and respectively six for the second selection algorithm were selected by the algorithms to be incorporated into the detection model (due to overlapping only eight distinct features). In general, this result confirms the rather uncertain detection power of baseline quantitative features. When ignoring the two features based on cash-flow metrics, which have not been adopted in this study, four out of the six remaining ones from Chen (2016) are similarly or identically incorporated in this study. Kotsiantis, Koumanakos, Tzelepis, and Tampakas (2006) have relied on a relief score to assess the individual discriminative power of their quantitative features. According to their results, from the 25 tested features (22 individual predictors, three related to previous periods) only eight have been selected by their feature selection approach to be part of the final detection model. Six of the eight variables are also part of this study's quantitative feature vector, whereas the remaining two are referring to cash-flow related data. Overall, the comparison across studies confirms the severe differences in individual discriminatory power of quantitative features that can be taken from table 23. Unfortunately, to my knowledge, no study exists that also compares the differences across qualitative and quantitative features on an individual feature basis.

In the next step, the consistency of the features in subsamples f1-f12 is examined and presented in Table 24. Overall, the 12 feature vectors comprise 5,265 unique features in the case of the realistic sampling approach and 6,408 unique features in the pair-matching approach, again hinting at the linguistic variety that is required to distinguish the more similar observations in the matched sampling approach.[167]

	Frequency	
Number of occurrences	**Realistic**	**Matched**
1	2,514	3,434
2	1,128	1,522
3	688	788
4	355	376
5	223	161
6	130	72
7	84	34
8	60	15
9	56	4
10	21	0
11	3	1
12	3	1
Total	5,265	6,408

Table 24 – Frequency of qualitative features

Only three features occur in all of the 12 feature vectors for the realistic sampling approach (acquisition complete, gross margin and integration). For the matched sampling, only one feature (conduct operation) is consistently found.[168] The matched sampling vectors are less persistent than the realistic ones, with a lower number of features with an occurrence greater than four and more features occurring only one to four times. Of the realistic vectors, 357 features occur more than six times, while of the vectors constructed under the matched

[167] The feature vector size was determined in 5.1.1 and fixed to 1,000 qualitative features.

[168] The 5-grams also incorporate 1-grams to 4-grams.

sampling, only 127 can be found in half of the samples. The results may hint at the fact that generalizable qualitative features can hardly grasp financial statement fraud and that a greater number of clues is necessary. To further elaborate on the results, differences in the discriminatory power of the respective features with different occurrences are assessed by calculating the average rank in the feature vectors (as of the IGR) for each number of occurrence. It may be the case that a few qualitative features that often occur and that also have a high rank in the feature vector, represent the generalizable qualitative clues. However, the results do not support this claim. The average rank barely changes for different numbers of occurrences of 5-grams for both sampling approaches.

Finally, it may be interesting to examine whether similar studies that rely on textual analysis for fraud detection purposes demonstrate some overlap in their qualitative features. Therefore, a comparison of the publicly available qualitative features of other studies has been undertaken. Three similar studies have reported the composition (or at least an excerpt) of their qualitative feature vectors. The features are compared on a relatable basis, as preprocessing such as the elimination of stop words or stemming of texts differs across studies as well as in terms of the length of the qualitative features. For example, if one of this study's features is regarded as "additional capital require", a study reporting only "require" or "required" would be counted as an overlap.

Purda and Skillicorn (2015) utilize 200 single words (also containing the symbols $, %, & and the single letter s, most certainly referring to the possessive "s"), of which 170 (85%) can also be found in this study's overall qualitative feature vector. Goel et al. (2010) have reported two lists, each containing the 25 highest ranked single words by information gain. The overlap covered about 80%, most differences being associated with features that capture country or city names like Venezuela or Manhattan, which were not part of this study's highest-ranking features. Chen et al. (2017) have reported a list of 240 qualitative features based on Chinese characters while also delivering an idiomatic English translation. In contrast to the two aforementioned studies and the present investigation, they relied on a Chinese/Taiwanese rather than the commonly used US setting. Interestingly, the overlap was considerably high, standing at about 79%.

In sum, it can be ascertained that qualitative features are rather unstable over time, supporting the previous findings and again hinting at the necessity of regular updates, as depicted by the small number of reoccurring features. Moreover, the overlap of qualitative features across different studies, even for examinations in different languages, is surprising.

5.2.2 First Instance of Fraud Detection

Finally, the three enhancing questions should help to improve the reliability of the results, detection performance, and assessment of the models' economic benefits, rendering them more relatable as a result. The initial question deals with a first instance of fraud detection approach.

In the financial statement fraud detection scenario, it is important to control for cases that span across multiple firm years and therefore observations in the sample. If not, the detection model may be biased towards repeatedly occurring observations of the same firms, meaning that the model identifies the respective firms rather than the general concepts that distinguish between fraudulent and non-fraudulent reports. Especially when dividing cases into training and holdout sets, those that reach across subsamples may lead to biased results. This is mainly caused by the time gap between the release of the misstated reports and the issued AAERs, as the latter are issued during or at the end of an investigation, leading to the possibility of consecutively misstated reports. For fraud detection models, the goal should be to identify misstatements at their first occurrence. Therefore, this study controls for first instance misreporting, similar to Brown et al. (2018). It may be expected that the detection proves slightly worse than the original approach. However, the models should still be able to score results at a reasonable level.

For each case spanning across multiple years, the initial year is identified and the consecutive years are separated from the sample. This procedure has been carried out for the realistic and matched sampling approaches for subsamples r1/m1. For the pair-matching process, the deletion of cases with multiple instances of fraud (multiple firms years of the same company) leads to a slight shift in both groups, altering the relative group sizes from 50/50 to 40/60 (fraud/non-fraud), which is not rebalanced afterwards. The detection results are tested for the realistic, and the pair-matched approach. In accordance with the previous results, SVMs have been chosen as the default classifier. The results are presented in Table 25.

Sampling	Design	Qual.	Quant.	Quant. & qual.
Realistic	**Original**	0.809	0.687	0.817
	First instance	0.720	0.639	0.730
Matched	**Original**	0.755	0.657	0.771
	First instance	0.668	0.599	0.686

The table reports AUC values for the test of first instance of fraud detection.

Table 25 – First instance of fraud detection

According to expectations, the results reveal that detection performance decreases for the three feature vectors in both sampling approaches. The decline in the case of feature vectors constructed from qualitative features may be mitigated for both sampling approaches when qualitative feature selection is carried out separately for the first instance of fraud detection, which has not been done but rather the qualitative features from the original versions of the subsamples have been utilized for the test. This results in potentially party suboptimal feature vector compositions, especially for the matched sampling approach (for which greater linguistic variety can be assumed necessary), as shown previously. Overall, the results seem to hold for the first instance of fraud detection and detection performance above the random classifier can be achieved, in the case of the realistic sampling even on a satisfactory level with AUCs above 0.7.

5.2.3 Quantitative Feature Vector Enhancements

The basic performance of the quantitative feature vector is intrinsically bad, especially when compared to the results of the qualitative feature vectors. Abbasi et al. (2012) and Dong et al. (2016) have shown that a baseline detection model that relies on annual financial ratios yields rather unsatisfying results, consistent with the results of the present study.

To improve the detection quality of the quantitative feature vector, the features must be better contextualized to increase the detection potential. In addition to the use of vertical financial statement analysis, as partially covered through the financial ratios so far, horizontal analysis for every observation might provide greater context for the classifier (Gee, 2015, p. 320; Wells, 2017, pp. 354–355). Dechow et al. (2011) have shown that their set of variables (F-score) demonstrates a particular trend prior to misstatement events, which may be captured by additional quantitative features from previous years. Abbasi et al. (2012) have presented similar results when boosting their baseline annual financial ratios through additional information from previous quarterly reports. In this regard, it can be assumed that the augmented quantitative feature vector will score better detection results.

Observation	**Fraud**	**5-gram 1-1000**	**Quant. 1 (t)**	**Quant. 2 (t)**	**…**	**Quant. 19 (t)**	**…**
A(2010)	0	5-grams (t)	**0.3**	0.0		1.1	
A(2011)	0	5-grams (t)	**0.9**	0.1		0.8	
A(2012)	1	5-grams (t)	**0.5**	0.15		0.7	
B(2011)	0	5-grams (t)	1.0	0.3		0.55	
Observation	**Fraud**	**5-gram 1-1000**	**Quant. 1 (t)**	**Quant. 1 (t-1)**	**Quant. 1 (t-2)**	**Quant. 1 (t-3)**	**…**
A(2010)	0	5-grams (t)	**0.3**	0.2	0.4	0.1	
A(2011)	0	5-grams (t)	**0.9**	**0.3**	0.2	0.4	
A(2012)	1	5-grams (t)	**0.5**	**0.9**	**0.3**	0.2	
B(2011)	0	5-grams (t)	1.0	0.5	0.8	0.2	

Table 26 – Feature vector design comparison

Table 26 depicts the structure of the feature vectors being compared. The upper half of the table illustrates the combined feature vector utilized so far, which comprises 1,000 qualitative and 19 quantitative features. The highlighted cells represent the incorporation of time effects in the variables, which are only based on the current measurement for every observation. In the bottom half, the boosted quantitative feature vector structure is presented, which still consists of 1,000 qualitative features but now also has 76 quantitative features. As for each quantitative feature, three previous values from prior years are taken into consideration for each observation. Therefore, the highlighted cells now represent the augmented version, which additionally incorporates the changes in the measurements in a horizontal way.

The test has been conducted by utilizing both the quantitative and the combined feature vector to examine the influence first on an isolated level using only quantitative features and then by adding them to the previously assessed best-performing feature vector. Both tests have relied on the matched and the realistic sampling approach. Support vector machines have been chosen as the classifier in accordance with previous results. The test has been undertaken for the entire 12 subsamples.[169]

Hypothesis 7 suggests that the results will improve when the additional feature enhancements are incorporated. Table 27 presents the average AUCs across the 12

[169] To ensure comparability, parameter combinations of the SVM have been assessed in accordance with the procedure of the original feature vectors. The results can be seen in Appendix D.

subsamples. Surprisingly, the only feature vector that scored better on average was the combined one in the realistic sampling approach.

	Original		Enhanced	
	Quant.	Quant. & qual.	Quant.	Quant. & qual.
Realistic	0.680	0.834	0.655	0.836
Matched	0.653	0.805	0.640	0.795

The table reports AUC values for the comparison of original and enhanced quantitative feature vectors.

Table 27 – Feature vector enhancement

Paired t-tests for the results over the 12 subsamples for enhanced versus original feature vectors have been computed. In none of the comparisons is the enhanced feature vector significantly better at the 5% level, leading to the rejection of hypothesis 7. An explanation might be that the additional features convolute the vectors, resulting in less decisive patterns. Although the enhanced feature vector scores the best result of this study comprising qualitative and quantitative features with an AUC of 0.939, across the entire timeframe the results are not significantly better than for the original design.

5.2.4 Cost-Sensitive Results

In the final step, a cost matrix is introduced to assign weights to the classified observations based on artificial costs. The cost matrix should increase the understandability and relatability of the results and serve as an additional metric to assess detection performance and to compare the different feature vectors on a practical basis.

In the binary classification scenario, there exist two possible misclassified outcomes: an actual non-fraudulent observation being classified as fraudulent (false positive) and an actual fraudulent case being classified as non-fraudulent (false negative). The costs are assumed to be asymmetric, as false positives are less costly than false negatives (e.g. Hansen, McDonald, Messier, & Bell, 1996; Abbasi et al., 2012). Moreover, the costs vary for different stakeholder groups, like regulators or investors (e.g. Beneish, 1999). An investor who invests in a fraudulent company incurs losses that are attributable to the reduction in share prices after the fraud is unveiled (fraud classified as non-fraud). If the investor had not been investing in a company due to its wrongful classification as fraudulent, he or she would have incurred opportunity costs, which are likely to be smaller than for the first error. For investors, Cox and Weirich (2002) have estimated the impact on market capitalization for cases of fraud between 1992 and 1999. After being released to the public (based on *Wall*

Street Journal articles that were confirmed by SEC releases), the average negative impact on market capitalization was 23.2%. According to Beneish (1999), the loss accumulates on average 40% on a risk-adjusted basis for the quarter following the discovery of the fraud. Taking both studies into consideration and assuming that the change in market capitalization for a non-fraudulent company would be insignificantly small (an appreciation of around 1-2%), a cost ratio of 1:10 – 1:40 can be assumed for investors.[170]

From a regulator's perspective, balancing the costs of wrongfully accusing a firm of fraudulent manipulations and thus investigating the case, to costs of not protecting the market participants, is similarly delicate. However, both perspectives share in common that the costs of wrongfully classifying an actual fraudulent observation are considerably higher than the other way around. The literature has suggested considerably different ratios, which can be seen in Table 28.

Studies	**Cost ratios**						
Persons (1995)	1:1	1:5	1:10	1:20	1:30		
Beneish (1999)	1:1		1:10	1:20	1:30	1:40	1:60	1:100
Abbasi et al. (2012)			1:10	1:20				
Perols et al. (2017)					1:30			

Cost ratio: costs per false positive/costs per false negative

Table 28 – Cost ratios of previous studies

The ratios from the literature also differ with regard to specific stakeholder groups. Beneish (1999) considers costs for regulators to be around 1:20 – 1:30 (cost of a false positive to costs of a false negative), while costs for investors range between 1:1 – 1:100. In contrast, Abbasi et al. (2012) suggest that the cost ratio is 1:10 for regulators and 1:20 for investors. Both studies have estimated the costs based on empirical justifications, illustrating the difficulty in determining the cost ratio for this study.[171]

In the next step, the costs of misclassified observations for this study will be determined based on empirical evidence. To determine the audit costs, the results from the Financial Executives Research Foundation (FERF) are taken into consideration. In their 2017 Audit Fee Survey Report, the median audit fee for public companies in 2016 was around $523,000, according to an examination of SEC filings; moreover, this was rather stable in previous

[170] Assuming the equity appreciation rates from Beneish (1997).

[171] For example, Abbasi et al. (2012) derive the median loss for a financial statement fraud case from the ACFE's Report to the Nation in 2010, which was $4.1 million; however, the current estimated loss as published in 2018 was only $0.8 million. Taking these changes into consideration results in severe differences in the underlying misclassification costs.

years, with only slight increases.[172] The losses of cases of financial statement fraud have been determined based on the ACFES's reports to the nation 2008–2018. On average, across the reports, the losses of financial statement fraud were estimated at approximately $1.6 million (min: $800,000; max: $4.1 million). Taking the empirical justification and hints from previous literature into consideration, the cost ratios may be assumed to range between 1:2 and 1:10, without factoring hardly quantifiable negative effects like market participants' loss of trust. For this study, the two cost ratios 1:3 and 1:10 with the associated costs based on the regulators perspective will be used.

The above discussion indicates the difficulties in assessing reasonable ratios. Given that the cost ratios significantly influence the results, the findings must be interpreted with caution. Another problem with the costs of misclassification is associated with the underlying sampling approach. In a realistic sample, the costs of misclassification may largely be driven by the huge amount of non-fraudulent observations. Correctly classifying one additional fraudulent case may be associated with a larger amount of wrongfully classified non-fraudulent observations, especially compared to the matched sampling approach. Therefore, a fixed cost ratio will likely result in varying outcomes for different sampling approaches, rendering them difficult to compare.

To obtain cost-sensitive results, different threshold values for each model have been examined, ranging from 0.1 to 0.9 in increments of 0.1.[173] Each of the resulting confusion matrices has been used to compute the total amount of misclassification costs. The classification errors have been measured based on the empirically estimated costs, as previously discussed: $523,000 per wrongfully classified non-fraudulent observation and $1.6 million per wrongfully classified fraudulent case, resulting in a cost ratio of about 1:3. In addition, a cost ratio of 1:10, based on $523,000 for false positives and $5.23 million for

[172] The report of the FERF is available at https://www.workiva.com/resources/2017-audit-fee-survey-report.

[173] At this point, it is important to recall the foundation of the AUC as presented in section 4.2.5, which represents the results across varying discriminatory thresholds, each representing an individual confusion matrix. The AUC values for each model represent the final results of this study. The AUCs are obtained by testing the trained model on an unseen holdout set. Nine different thresholds (0.1-0.9) have been used to compute individual cost matrices for each AUC. The threshold resulting in the lowest costs of misclassification relative to the two baseline strategies (naïve and surveillance) is chosen to reflect the highest possible cost-saving potential. It is important to clarify that the following results are solely based on the aforementioned procedure and therefore do only depict a cost saving potential, which does not necessarily reflect the true cost savings. This is due to the fact, that for the assessment of true cost savings, the optimal threshold had to be determined beforehand and not on the holdout set. However, as the outcomes in the form of AUC values are correctly validated, the highest cost-saving potential rather serves as a translation of an abstract concept into an economically relatable measure of the detection performance, rather than finding one best confusion matrix for each case, which is questionable anyway, due to the subjective way of identifying correct cost ratios.

false negatives, has been created, to additionally control for a larger negative impact of misclassified fraudulent cases.[174] To incorporate the costs into the classification results, the number of false positives has been multiplied by $523,000 and the number of false negatives by $1.6 million (or $5.23 million for the 1:10 ratio). The total costs of each classification result have been compared to a naïve strategy, in which all observations are regarded as non-fraudulent (costs of model compared to naïve strategy, e.g. Persons, 1995; Beneish, 1999; Abbasi et al., 2012). In addition, they are also compared to a surveillance strategy, in which all observations are investigated (costs of model compared to surveillance strategy, e.g. Abbasi et al., 2012). With regard to both strategies, a successful detection model should achieve considerably lower total costs.

To generate comprehensive results and to be able to compare the outcome to the previous tests of the hypothesis concerning the feature vector performance, the results have been calculated across the three feature vectors and for both sampling approaches. Based on previous findings, the chosen classifier was an SVM.[175] In contrast to the aforementioned studies like Abbasi et al. (2012), this study relies on a design using 12 rolling subsamples, to test the detection power over the 15-year timeframe. Therefore, the average costs of misclassification have been computed for the 12 results for each feature vector and sampling approach. [176] Table 29 presents the results.

For the realistic sampling approach, the surveillance strategy results in an enormous amount of total costs due to a large number of observations, which need to be investigated, whereas comparatively small total costs are estimated for the naïve strategy in which no investigations are conducted, due to the relatively small number of fraudulent cases. This is true for both cost ratios investigated in this study but might change when immensely high costs are assumed for false negatives relative to false positives. To sum up, achieving superior results to the strategy in which nothing is investigated is more difficult than achieving superiority to the surveillance strategy (for example, investigating every observation in subsample 1 would amount to total costs of around $3,477 million, while not investigating would lead to total costs of $118 million for the 1:3 cost ratio).

[174] Taking the discussion about prominent cases of financial statement fraud from section 2.1.3 into consideration, outliers may very well exceed losses, amounting to a few billion rather than million USD.

[175] The results may furthermore be increased by searching for the optimal threshold that minimizes the costs. This study only focuses on nine different thresholds, but searching in smaller increments in areas of the AUC with FPR/TPR ratios resulting in low costs would probably improve the results.

[176] This resulted in the evaluation of a total of 648 computed cost matrices: 2 sampling approaches, 3 feature vectors, 9 thresholds and 12 subsamples over the entire timeframe.

In comparison, the matched sampling approach – or any other sample balancing – leads to a distortion of total costs for both strategies. Especially the pair-matched outcome shifts towards an artificial reality, in which it is more difficult for the detection model to outperform the naïve strategy than to refrain from investigating (for example, investigating every observation in subsample 1 would amount to total costs of around $75 million, while desisting from investigating would lead to total costs of $115 million for the 1:3 cost ratio). Again, this is the case for the cost ratios assumed in this study. Interpreting the matched sampling results is rather unintuitive. However, the feature vector performance in the artificial environment in which fraudulent cases are much more likely than in reality can still be assessed.

Realistic sampling

Cost ratio	Feature vector	Costs of model errors relative to naïve strategy	Costs of model errors relative to surveillance strategy
1:3	**Quant.**	0.928	0.027
	Qual.	0.671	0.019
	Quant. & qual.	0.657	0.019
1:10	**Quant.**	0.875	0.083
	Qual.	0.571	0.054
	Quant. & qual.	0.565	0.053

Matched sampling

Cost ratio	Feature vector	Costs of model errors relative to naïve strategy	Costs of model errors relative to surveillance strategy
1:3	**Quant.**	0.308	0.470
	Qual.	0.254	0.388
	Quant. & qual.	0.251	0.384
1:10	**Quant.**	0.095	0.474
	Qual.	0.089	0.446
	Quant. & qual.	0.089	0.443

The table reports outcomes compared to different detection strategies.
Cost ratio: costs per false positive/costs per false negative
Naïve Strategy: no investigation, all observations are regarded as non-fraudulent
Surveillance strategy: every observation will be investigated

Table 29 – Results for highest cost-saving potential

The results in Table 29 confirm the outcome of H4a and H4b. The quantitative feature vector is hardly better than a strategy in which no investigation would be carried out (which would result in values of >=1 in the column costs of model errors compared to naïve strategy). In contrast, the qualitative and combined feature vectors present considerably superior results for the more practically oriented realistic probability of fraud sampling approach.[177] In line with previous results, the combined feature vector is best suited for the fraud detection purpose. This holds for both sampling approaches.

A final conclusion can be made by estimating the absolute potential of savings that would be possible by utilizing the best performing feature vector and classifier that this study has identified for the 12 subsamples covering the 15-year timeframe. In this way, the absolute costs of misclassification are deducted from the absolute costs of both strategies. The savings compared to the naïve strategy accumulate to \$1.6 billion for the cost ratio of 1:10 (\$409 million for the cost ratio of 1:3). For the surveillance strategy in which every observation is investigated, \$33.4 billion for the cost ratio of 1:10 could be saved (\$34.7 billion for the cost ratio of 1:3).[178] Higher estimated costs of fraudulent cases, which is arguably reasonable, would only increase the absolute cost savings and therefore increase the relevance of this study's results.

5.2.5 Results Comparison

Comparing the results of this study with other fraud detection studies is hardly feasible, due to considerable differences regarding research design, especially sampling. Furthermore, the results are often reported in vastly different ways, which in conjunction with sampling differences causes even more problems when directly comparing the outcomes. Most likely, a research design that exploits common cross-validation and therefore does usually not rely on a holdout set comprise future, unseen observations, would report overly confident results compared to a real-world application of the respective model, as carried out in this study.[179] Table 30 captures the primary results of studies on fraud detection that have reported comparable detection performance metrics. The studies with the highest similarity regarding research design will be discussed hereafter.

[177] In the matched sampling approach, classifying everything as fraudulent often results in the lowest absolute costs because of the asymmetric cost structure. In this case, an accuracy of 50% and a recall of 100% are scored due to the equal size of both groups. The quantitative feature vector often reports this result when searching for the lowest total costs of misclassification. The combined feature vector, which reaches an accuracy of 72% and a recall of 92% in similar cases, results in rather small cost savings. Thus, Table 29 barely reveals the true cost saving potential of the feature vectors in the matched sampling approach.

[178] A detailed summary of the determination of the absolute costs can be found in Table 35 in Appendix E.

[179] See for example section 4.1.3 and 4.2.2 for further substantiations on the relevance of a separation of feature extraction, model training and testing.

The study that comes closest to this one is Brown et al. (2018) owing to similarities regarding the examination of qualitative data and the sampling design. This enables to report results that come close to a practical, real-world application, with the detection of multiple single-year holdout sets over a larger and in this case similar timeframe. Brown et al.'s highest average AUC of 0.778 is considerably worse than the 0.836 scored in this study, most likely attributable to their topic modelling design that limits the number of features through aggregation compared to the raw n-gram approach.

Cecchini et al. (2010a) have presented another of the few studies that utilize combined feature vectors in their detection model. Cecchini et al. (2010a) were constrained by the small sample size with only 61 fraudulent and 61 non-fraudulent observations between 1993 and 2002 and therefore had to rely on a leave-one-out analysis, which did not incorporate the assessment of the results on future unknown observations from subsequent years. With an accuracy of 81.97% and a recall of 80.77%, they have shown high discriminatory power for the model relying on quantitative and qualitative features. For a similar timeframe, under matched sampling and for the qualitative and quantitative feature vector, this study achieves an accuracy of 70.83% with a recall of 91.67%, but on a holdout set covering future unseen observations.

Cecchini et al. (2010b) and Abbasi et al. (2012) are similar to this study with regards to the detection of future unknown cases but differ as both studies utilize only quantitative features to create their detection models. Cecchini et al. (2010b), like Abbasi et al. (2012), validate their models on a holdout set covering the years 2001 to 2003 and a training set covering the years 1991 to 2000. Cecchini et al. (2010b) created their so called financial kernel that transforms input attributes into features for the classification via SVMs. With an AUC of 0.878, the detection performance of their financial kernel model based on quantitative features can be regarded as one of the highest available. In line with Abbasi et al. (2012), when limiting this study's results to the respective years, the performance accumulates to 0.891 for the subsamples with observations from 2001 to 2003.

Abbasi et al. (2012) have adopted a similar research design and have reported very high AUC values for their comprehensive MetaFraud framework. The main differences can be found in the use of quarterly and annual reports, the feature vector being derived solely from quantitative data, the sampling design encompassing a slightly different timeframe and the sample being subdivided into training and testing, cut off at the year 2000.[180] Furthermore, their sample has a 10% fraction of fraudulent observations (815/8,191), which does certainly not reflect the true occurrence of fraud. For their best model, an AUC of 0.931 indicates outstanding detection performance. On average across the 12 subsamples, a similar result could not be achieved by this study's models, but when limiting the 12 subsamples to reflect similar years to the testing set of Abbasi et al. (2012), an average AUC of 0.888 was achieved (including subsamples with only observations from 2000–2008), even peaking at 0.938 for the subsample starting in the year 2002. This was also the highest result achieved in this study under a realistic sampling approach and the combined feature vector with enhanced quantitative features utilizing an SVM. The high results may therefore be attributable to a sampling effect.

The problem with studies that report accuracy and recall or similar confusion matrix-based metrics rather than AUC values can be exemplified as follows. When determining the minimal costs of misclassification, different confusion matrices are generated based on various classification thresholds for every model, each representing an individual classification outcome. By using the empirically determined costs of false positives and false negatives, the lowest total costs are assessed. For the realistic sampling approach and a cost ratio of 1:10, this results in an average accuracy of above 99%, with a respective recall of 48%.[181] By choosing different thresholds, the recall may be increased, but accuracy would be decreased during the course, resulting in higher total costs, thus constituting an unsatisfactory result (e.g. subsample 4, optimal: accuracy of 99% and recall of 58%; suboptimal: accuracy of 96% and recall of 75%). In simple terms, sacrificing recall for additional accuracy is necessary owing to the large number of non-fraudulent observations relative to fraudulent cases in the sample.

Besides comparing raw detection outputs measured in abstract concepts like the AUC, it is also possible to compare the results to studies focussing on more economically relatable measures like the cost-saving potential resulting from the application of the detection models. In section 5.2.4, the costs of misclassification of the detection models were compared to costs of a surveillance and naïve a strategy, which investigate all or none of the observations respectively. This reflects one of the most common procedures when the

[180] Although their adaptive learning algorithm takes the evaluation of future observations at continuous points in time into consideration.

[181] Given that the AUCs vary over the 12 subsamples, so do accuracy and recall. For similar thresholds, the highest AUC of 0.93 resulted in an accuracy of 99% and a recall of 78%, while the lowest of 0.736 reported an accuracy of 99% and a recall of 29% for optimal thresholds in the 1:10 cost ratio setting.

performance of detection models is assessed based on the associated misclassification costs and allows to compare results across different sample sizes, which is not possible if absolute cost savings are reported. Persons (1995) reported costs of model errors relative to the naïve strategy of 0.971, indicating that the model is only slightly better than the naïve strategy in their matched sample for a cost ratio of 1:10. When recreating their findings with a similar cut-off probability, the quantitative feature vector in this study hardly reaches better results with around 0.931, however considerably better results are possible if choosing different threshold (around 0.1).[182] Beneish (1999) reported a cost of model error relative to the naïve strategy for a cost ratio of 1:10 of 0.680 for their model tested in a rather realistic probability of fraud sample.[183] Thereby, the model from Beneish beats the quantitative feature vector of this study, which reported an outcome of 0.875; however, the combined feature vector shows considerably better performance with a ratio of 0.565.

Summarizing the comparison, the detection performance of this study's models can be assumed to be at least on par with similar studies and usually scores even better than most of them when adjusting the overall results to comparable subsets. Furthermore, this study has proven to yield a reliable detection performance of the developed models over a timeframe of 15 years, through developing a comprehensive fraud detection approach that relies on thorough design and that is found to be robust on a matched and realistic sampling approach and close to real-world application. Especially the detection of fraud within unknown observations of future periods is a fundamental design decision that combines practical application and validation of the models and yields a decisive insight into the true detection performance. Overall, the developed models in this study yield high and reliable detection results, in a realistic setting on a generalizable level, which in conjunction has not been achieved so far. Furthermore, the study contributes to the financial statement fraud detection literature in numerous areas according to the 11 design and enhancing questions, while presenting one of the most comprehensive overviews of financial statement fraud detection to this date.

[182] Section 5.2.4 has already discussed the rather unintuitive nature and questionable usefulness of cost ratios in matched samples.

[183] The result has been chosen from the estimation sample of Beneish (1999) to account for similarities of this study's assessment of cost-saving potentials.

Quantitative features	**Fraud**	**Non-fraud**	**Setting**	**AUC**	**acc. (%)**	**rec. (%)**	**val.**	**f.**	**t.s.**
Loebbecke et al., 1989	77	305	USA			86	Training		
Persons, 1995	100	100	USA		72	64	Holdout		
Hansen et al., 1996	77	305	USA		89		Holdout		
Green & Choi, 1997	46	49	USA		72	68	Holdout		
Fanning & Cogger, 1998	102	102	USA		63	66	Holdout	Yes	
Summers & Sweeney, 1998	51	51	USA		60	68	Training		
Beneish, 1999	74	2,332	USA		90	54	Holdout		
Lee et al., 1999	56	60,453	USA			73	Holdout		
Bell & Carcello, 2010	77	305	USA		73		Holdout		
Feroz et al., 2000	42	90	USA		94		Holdout		
Spathis, 2002	38	38	Greece		84	84	Training		
Spathis et al., 2002	38	38	Greece		75	64	Holdout		
Lin et al., 2003	40	160	USA		86	35	Holdout		
Kaminski et al., 2004	79	79	USA		54	22	Holdout		
Kirkos et al., 2007	38	38	USA		90	92	Holdout		
Gaganis, 2009	199	199	USA		87	88	Holdout		
Lou & Wang, 2009	94	467	Taiwan		87		Holdout		
Cecchini et al., 2010(b)	132	3,187	USA	0.878	90	80	Holdout	Yes	
Dikmen & Küçükkocaoğlu, 2010	17	109	Turkey		67	81	Holdout	Yes	
Dechow et al., 2011	293	79,358	USA		64	69	Holdout	Yes	
Ravisankar et al., 2011	101	101	China	0.981	98	98	Holdout		
Alden et al., 2012	229	229	USA		76		Holdout		
Abbasi et al., 2012	815	8,191	USA	0.931		82	Holdout	Yes	(Yes)
Liu et al., 2015	138	160	China		88		Holdout		
Perols et al., 2017	51	15,934	USA	0.762			Holdout		
Qualitative features									
Goel et al., 2010	450	622	USA		89	89	Holdout		
Glancy & Yadav, 2011	11	20	USA		84	90	Holdout		
Humpherys et al., 2011	101	101	USA		65	71	Holdout		
Purda & Skillicorn, 2015	1,407	4,708	USA	0.890	83	81	Holdout		
Chen et al., 2017	45	135	China/Taiwan		85		Holdout		
Qualitative and quantitative features combined									
Cecchini et al., 2010(a)	61	61	USA		82	84	Holdout		
Dong et al., 2016	805	805	USA		82	92	Holdout		
Brown et al., 2018	459	37,806	USA	0.778		75	Holdout	Yes	Yes
This study's results									
Average*	805	84,960	USA	0.836			Holdout	Yes	Yes
Highest**				0.938	99**	59**	Holdout	Yes	Yes

acc. (accuracy): fraction of correctly classified observations rec. (recall): fraction of correctly classified fraudulent observations

val. (validation): the results are achieved by classifying the observations of the training or holdout set, for example through cross-validation

f. (future): where the design explicitly states that the observations of the holdout set are from future periods and feature extraction is limited to previous years

t.s. (time series): the results are assessed at different points in time in order to create a reliable detection model over the course of time

*Average performance across the final sample **Threshold resulting in the lowest absolute costs of misclassification for a single subset

Table 30 – Results comparison

5.3 Limitations

The interdisciplinary approach of this study involves a number of limitations in different areas. The majority of the limitations can be divided into either limitations concerned with validity and reliability of the measurements and the results or machine learning related limitations. Each limitation may be targeted by further research and will be discussed and complemented in detail in the subsequent section.

One of the most apparent limitations of this study concerns the reliability and validity of the results. As discussed in chapter 4.2.1 machine learning is often regarded as a black-box, where the relationship between the features and the relationship between the features and the outcome is difficult to observe. Although this study has tried to make the features relateable as depicted for example in chapter 5.2.1 and created the models based on elaborated fraud theories, it is still questionable how exactly fraud is expressed in the textual components of annual reports. Especially the qualitative features seem to be highly abstract and according to the results do vary considerably over the examined timeframe, which most likely indicates that little generalizable clues exist. Efforts to examine the textual components on a more relatable level are rather scarce (e.g. Goel et al., 2010; Brown et al., 2018) and often with narrow focus. Further examination with state of the art natural language processing methods and in conjunction with fraud factors operationalized from fraud theory could enhance the understanding of fraud considerably and most likely have a positive impact on detection results.

Regarding the temporal instability of the qualitative features, the study is furthermore limited through potential countermeasures by companies occurring over time. These countermeasures could involve adapting the reports to drop clues associated with known ones that are hinting at fraudulent manipulations, thus rendering it more difficult if not impossible to be detected. Although this study advocates updating qualitative features regularly to solve the problem to a certain extent, companies may find a way to mask themselves. Furthermore, the models require a solid data foundation, which this study achieved by limiting the final year of the sample to 2010, providing enough time for as many fraudulent cases as possible to be reported in the meantime. Having limited amounts of fraudulent cases each year because of ongoing investigations may hamper the detection results and make it more difficult to update the features as suggested.

Another limitation concerns the identification of fraudulent observations. Although AAERs are a common way of identifying fraudulent manipulations to financial statements in the literature, two problems arise from the fraud proxy. The first deals with the legal construct of fraud. Dechow et al. (2011) desist from using the term “fraud” and instead use “misstatement” for their AAER-based sample. They argue that although alterations to documents have been uncovered, managers rarely admit guilt with respect to allegations. Perols et al. (2017) suggest to only include firms for which fraud has been explicitly stated in SEC releases or other sources, reducing the sample considerably.[184] This procedure results in a precisely identified fraud group, in which it is likely that no wrongfully labelled fraudulent observations exist. In return, the non-fraud group might contain wrongfully labelled fraudulent observations due to the rather strict identification process. When looking into alternative fraud proxies like shareholder lawsuits, SOX internal control violations or restatements, it becomes apparent that neither is without limitations (Dechow et al., 2011, pp. 18–19). Shareholder lawsuits are biased towards firms with larger stock value declines, SOX violations do not capture older observations and often comprise younger or smaller firms with less developed accounting infrastructure and restatements are not strict in capturing fraud, which goes beyond honest mistakes and implies a level of intentionality.[185] The second problem results from the imperfect fraud proxy in terms of coverage. Samples relying on SEC enforcements are generally biased towards the SEC’s activity level (Dyck et al., 2017, p. 5). The amount of fraudulent cases hidden in the group of non-fraudulent observations is unknown and can hardly be estimated, potentially leading to a suboptimal identification of features used to distinguish between both groups. Dyck et al. (2017) estimate that for their sample of cases between 1996 and 2004, two-thirds of cases go undiscovered by common fraud proxies. However, according to Karpoff et al. (2017), samples constructed using AAERs possess, in comparison to other common sources used in financial misconduct research, the largest scope based on the fraction of cases they cover and one of the highest coverages of fraudulent cases in accordance with applicable fraud definitions. Summarizing, the imperfect identification of fraudulent observations and the associated unknown number of hidden fraud cases in the non-fraud group is rather a general problem in the financial statement fraud detection literature and not specific for this study.

The selection of the quantitative features represents another limitation that this and similar studies are concerned with. This study has focused on incorporating predictors that capture

[184] Perols et al. (2017) follow the sampling design of Beasley (1996).

[185] See section 2.1.2 for a discussion of intentionality of the misstatement in the identification of fraud and 4.1.1 for a comparison of data sources of financial misconduct research referring to Karpoff, Koester, Lee, and Martin (2017).

a large number of fraud factors based on elaborated fraud theories, predictors that have been commonly applied successfully in similar studies and that do not lead to a significant drop in sample size. This combination in conjunction with the data coverage in the Compustat universe has limited the scope to rather simple predictors from company financials, which is similar to other studies that have exploited quantitative and qualitative features like Cecchini et al. (2010a) or Brown et al. (2018). In both studies, the quantitative features are rather seen as a benchmark for the comparison of textual and combined models. However, Abbasi et al. (2012) for example have shown that complex quantitative features like features that compare the companies to industry means and others that create a direct relationship between the observations, or additional non-financial features can considerably improve detection performance, which was then partly addressed in this study's second enhancing question, as discussed in section 5.2.3, but certainly not to the greatest extent. Examining a combination of both, complex quantitative and qualitative features alike in one study would certainly contribute towards a better detection performance and allow a comparison of the two types of models on a more similar level.

Moreover, with regard to data availability, the fraud factors discussed in section 2.2 can hardly be implemented holistically into the detection models with the most commonly utilized data sources. Fraud theories offer a broad scope of factors and associated clues that may help to detect fraudulent actions (e.g. Feroz, Kwon, Pastena, & Park, 2000). However, this study and others have rarely attempted to operationalize them in greater detail, especially on the level of fraud perpetrators. Some management-specific, personal characteristics are potentially publicly available, although most likely not in annual reports and instead in other corporate communication sources, such as websites or outside of the company sphere in social networks.[186] Even the SAS No. 99 – Consideration of Fraud in a Financial Statement Audit (AU §316.85 and AU §316.86) hints at the attitude of top-level employees as a relevant source of clues. Incorporating data sources beyond the commonly utilized ones could not only increase detection performance but may also provide empirical justification for the applicability of fraud theories in the financial statement fraud context.

In addition to data availability and coverage, data authenticity may also influence the results, especially when relying on secondary data like Compustat instead of the actual data from the annual reports. Chychyla and Kogan (2015) examine the differences between

[186] To a certain extent, personal characteristics may also found in AAERs if an enforcement action directly concerns an officer.

Compustat data items and the applicable 10-K data. The results reveal significant differences between both sources. These differences may lead to the misrepresentation of peculiarities in the quantitative data, hinting at fraudulent manipulations and thus reducing the reliability and performance of the quantitative features. According to the authors' results, the divergence is furthermore industry- and size-specific. In the case of fraud, where subtle hints are supposedly relevant, it may be more appropriate to rely on the information as it is given in the annual report than use secondary data sources as commonly applied in the literature and in this study. Validating the results using the actual data source has so far not been carried out in the financial statement fraud detection context.

The generalizability of this study's results may also be limited as the sample is solely drawn from companies required to disclose their annual report on Form 10-K, resulting in mostly public US companies. In this regard, the detection models have also only been tested on texts from annual reports in the English language, although it might work equally well for related languages if the processing of texts in general and the preprocessing like stop word elimination and stemming, in particular, are adapted accordingly. The comparison of the qualitative features to different studies in section 5.2.1 has already hinted at a certain level of consensus in the identification of qualitative features even across languages.

A number of limitations also concern the machine learning methodology. The hyperparameter optimization was carried out once on a random dataset via 5-fold cross-validation for each feature vector and sampling approach. Detection performance would likely be improved if optimization were conducted via cross-validation on each training set individually across the different timeframes that were chosen. However, the computational effort must be limited and the optimization process needs to be reduced due to the large number of different tasks and controls this study aims for. The computational effort also limits the results for the artificial neural network. A more in-depth optimization procedure covering a larger number of parameter combinations would most likely lead to considerably better results for this classifier especially. Given that most studies report their classifier set-up only very superficially, it is hardly possible to ascertain whether similar constraints concern other studies, but it seems to be very likely.

5.4 Suggestions for Further Research

Derived from the limitations specifically, and the interdisciplinary nature of the fraud detection problem in general, the possibilities for further research are manifold. Thereby, the literature could focus on two areas. The first revolves around the theoretical concept of financial statement fraud, especially how it is conducted, who conducts it and the underlying reasons behind the fraud, while the second one is concerned with improving research methodologies and thereby the detection performance of models while ensuring reliability and validity.

Fraud theories lay the conceptual ground that can be used as the basis to develop detection models. However, most of the fraud theories, especially the less prevalent ones, are lacking empirical justification. The literature and practitioners are often only referring to the fraud triangle if adopting fraud theory at all into their work. Recapitulating section 2.2, the shortcomings of the fraud triangle coupled with the amount of elaborated offspring offers an array of fraud factors that may be operationalized in future research. Moreover, the development of new fraud theories precisely focussing on financial statement fraud may enhance the understanding of the subject and offer a better basis compared to the prevalent ones that are usually concerned with all kinds of fraud and often derived from one specific type, potentially limiting their direct applicability to financial statement fraud.

Besides the development or operationalization of new or unused fraud theories, the findings so far could also use additional confirmation from different data sources. The hints and clues of manipulation are supposed to be of subtle nature, so they may be best studied using the primary source, rather than potentially distorted databases. This study has partly achieved the goal of authenticity by relying on the very texts from annual reports, however constructing the quantitative features using data from the Compustat universe.[187] Utilizing for example XBRL data could shed additional light on the true nature of quantitative predictors and potentially reveal the required subtle insights.

Moreover, not all factors of fraud theories can be equally well operationalizable by the data sources commonly relied on. Examining different publically available sources of corporate communication for the purpose of identifying and extracting patterns that hint at accounting fraud – which could also be done in conjunction with for example annual reports – offers new research possibilities. Especially sources with less prescribed and standardized information (like information on websites) could supposedly contain vital hints on its own or in conjunction with financial information (e.g. Brazel et al., 2009). This could also involve examinations beyond textual and numerical data and capture for example graphical

[187] See Figure 11 in section 4.1.1 for a basic overview.

presentation styles or evaluate the availability or absence of information voluntarily provided in the respective sources.

Another area for further research is concerned with the specific creation of the detection models and may furthermore be separated into feature related improvements and classifier improvements. Regarding the latter, some studies have already adopted classifier enhancements like ensemble methods, where several single classifiers are combined to form a meta classifier, which for financial statement fraud detection but also in other contexts has already shown to be able to improve the results (e.g. Kotsiantis et al., 2006).[188] Evolutionary algorithms could also find increased usage in the context of financial statement fraud detection, especially when relying on predictors from textual data, where the algorithms can help identify the relevant patterns in the complex, and according to the results of this study, changing feature space when textual predictors are concerned (e.g. Alden et al., 2012).

The other potential field for improvements in this area is concerned with the way textual data is examined. So far, bag-of-words or bag-of-n-grams approaches (e.g. Purda & Skillicorn, 2015) besides topic modelling (e.g. Brown et al., 2018) and presentation styles (e.g. Goel & Gangolly, 2012) have been utilized for the examination of financial statement fraud. Fisher et al. (2016) or Lewis and Young (2019) discuss the implementation of textual analysis of corporate texts and propose roadmaps for future research, which implies avoiding pitfalls and relying on relevant, state of the art methods. Thereby, natural language processing offers great possibilities which in conjunction with the free availability of textual information and the access to tools and the required hardware can improve the understanding of financial statement fraud and its detection considerably.

[188] At first glance, this study's methodology would be predestined for ensemble learning, as multiple classifiers have been utilized from the beginning. After initial tests on a random dataset using 5-fold cross-validation, stacking and voting did not significantly beat the single best classifier, which is why further implementation has not been carried out. This may be because of the low number of classifiers (4 in the case of this study) utilized for the construction of the ensemble learners, which are usually more plentiful (e.g. 7 in the case of Kotsiantis, Koumanakos, Tzelepis, & Tampakas, 2006; 14 in the case of Abbasi, Albrecht, Vance, & Hansen, 2012).

6 Summary and Conclusion

Academics and practitioners alike have long been striving for approaches to identify financial statement fraud using publicly available data. Starting with financial ratios, toolkits have developed substantially over time. The relevance of truthful corporate reports for capital market efficiency is unquestionable. Capital market supervision and auditors are concerned with identifying fraudulent manipulations to ensure the trust of market participants in audited financial statements. Fraud theory suggests that complex models are necessary in order to capture fraudulent behaviour in its entirety. In recent years, the examination of textual information has become more popular and technological progress provides easier access to text analysis methods. The availability of machine-readable corporate texts and opportunities to process an increasing amount of textual data has opened the door for advancements in the area of accounting and finance research. In the domain of financial statement fraud detection, patterns extracted from textual data even seem to be superior to solely quantitative data like financial ratios.

The primary research goal of this study was the detection of future financial statement fraud using textual and financial data. A sound analysis of applicable fraud theories and a detailed assessment of financial statement fraud schemes have laid the basis for the development of a comprehensive detection approach. Therefore, 11 design questions refined by seven hypotheses have been formulated and answered by conducting an empirical analysis to create reliable fraud detection models that can efficiently distinguish fraudulent and non-fraudulent reports over a timeframe of 15 years, which can be seen in figure 35. The design questions, besides serving to build, improve, and assess the detection models, have also helped to explain financial statement fraud and add to the general understating of the clues in annual reports that can hint at fraudulent manipulations.

Goal: Detection of future financial statement fraud exploiting qualitative and quantitative information from annual reports.

Design questions

1. Size of qualitative feature vector
   Larger qualitative feature vectors score better results
2. Stability of qualitative features over time
   Qualitative feature are time-dependent
3. Varying time gaps between training and holdout sample
   Time gaps lead to deteriorating results
4. Varying training and holdout set sizes
   Sample sizes influence results
5. Detection rates over time
   Reliable detection performance over the entire timeframe
6. Feature vector performance
   Combined feature vectors score best results
7. Matched and realistic sampling
   Results hold for both sampling approaches
8. Classifier performance
   SVMs superior

Enhancing questions

1. First instance of fraud detection
   Results hold for first instance of fraud detection
2. Quantitative feature vector enhancements
   Enhancements do not lead to significantly better results
3. Cost-sensitive results
   Models are cost-efficient

Figure 35 – Summary of research goals

This study contributes in numerous areas to the financial statement fraud detection literature. To date, this is the first study to use longer multi-word phrases to identify accounting fraud from corporate texts. The methodology has facilitated the capturing of larger patterns in the narratives and has partially addressed context in order to reflect linguistic and narrative peculiarities more precisely. Furthermore, the texts were preprocessed by eliminating stop words and a word stemmer was introduced in accordance with empirical evidence regarding increased detection performance for preprocessed texts. In contrast to the rather static sampling designs of most studies, changes to detection results were assessed when the timeframe for pattern extraction and model building varied. In this regard, this study has also examined how text-based patterns alter over time and has shed light on the importance of updating qualitative features resulting from the textual analysis on a regular basis. In this way, it is possible to incorporate potential changes to narratives from corporate disclosure in general as well as to absorb fraud-relevant patterns, which are also assumed to be time-dependent. A failure to update regularly or to avoid time gaps between model building with pattern extraction and training and the actual fraud detection of unknown cases in future periods can seriously compromise detection performance. Thereby, the detection model is automated from the initial data gathering until the final results and compared to other techniques like topic modelling free from human involvement and the associated subjectivity.

This study's reference to practical applicability constitutes one of its most important traits, as previous literature has tended to empirically test fraud detection models on constructed, balanced samples alone, without creating a test environment that comes close to a real-world scenario. To thoroughly test the models, both a balanced sample through pair-matching fraudulent and non-fraudulent observations as well as a realistic probability of fraud sample, in which all available annual reports are taken into consideration, has been utilized. This research design decision has made it also possible to compare the results with previous studies, to examine potential differences in the evidence while maintaining the advantage of the control environment that the matched sampling generated and to present an outcome, based at the level of practical implementation. The study's sampling approaches have laid the foundation for its pivotal results, in which the detection performance of financial metrics, textual patterns and a combination of both have been tested, adopting a range of classifiers from basic techniques like naïve Bayes and k-nearest neighbour to more sophisticated ones like support vector machines and artificial neural networks, providing detection results over a 15-year timeframe using rolling subsamples. Moreover, by applying the matched sampling approach and an additional first instance of fraud detection test, the validity of the results has been assessed as effectively as possible, which has not been done to this extent in the literature so far.

In line with Cecchini et al. (2010a), Purda and Skillicorn (2015) or Brown et al. (2018) this study's results support the view that textual information is superior to financial ratios in identifying accounting fraud. Detection performance can also be slightly increased by combining both data types, yielding significantly better results for the realistic probability of fraud sampling approach and the best classifier. With an average AUC of 0.836 across 12 rolling subsamples between 1996 and 2010 and a single highest result of 0.939 for the subsample covering the years 2002 to 2005, high and reliable detection performance has been demonstrated. Due to design differences regarding sampling approaches, examined timeframes, the explicit design of trying to detect future fraudulent cases, and the fact that this study is solely focussing on cases in the SEC's regulatory sphere, the comparability to other studies is limited to a cursory view, which is a general constraint in the field of financial statement fraud detection. Nonetheless, the reported results of similar studies regarding the aforementioned design peculiarities reveal the potential improvements this study's results offer, being at least on par with the highest reported ones and even topping them on isolated, similar timeframes. The eminent detection performance becomes also apparent when examining the associated costs of misclassification. Besides reporting the rather abstract detection performance measures, this study facilitates on empirically derived cost ratios to assess the cost saving potential of the underlying detection models, scoring considerably better results than comparable studies and amounting to an absolute cost saving potential of \$34.7 billion over the entire timeframe under an conservative estimation of misclassification costs.

The results of this study can be especially relevant for everyone concerned with the identification of fraudulently altered financial statements. For example auditors, regulators and enforcement agencies can be supported by an automated detection system that helps to allocate resources to high risk targets and therefore increase efficiency. Auditors, for example, can identify needs for in-depth audits, while regulators and enforcement agencies can use the system for supervision purposes. This may also increase and restore the trust in audited financial statements and decrease the chances of severe accounting scandals like during the early 2000s. Supervision agencies may even benefit in the case of auditor involvement in accounting fraud or auditor inability to detect fraud and still secure trust in financial statements. Thereby, the detection models are not only serving as a source of fraud detection but may also deter fraud from happening when chances of getting caught are increased.

Despite the aforementioned obvious groups, all stakeholders may benefit from the detection models alike. For example, investors or cooperating companies could utilize the detection models to control the reported company financials, therefore reduce uncertainty and the risk to fall for manipulated numbers. As the detection system has been created to be as free as possible from human subjectivity and involvement and works on current hardware it hardly accrues additional costs.

Altogether, it can be concluded that the current stage of the development of financial statement fraud detection models that solely rely on publicly available data permits the identification of fraudulent cases with high accuracy, efficiency, and reliability. This study's unique research design and scope have not only demonstrated outstanding detection performance but furthermore increased the understanding of setting up detection models and thereby shed additional light on the way financial statement fraud expresses itself in the manipulated disclosure. In order to ensure the validity, practical application, and comparability of the results, future studies should try to rely on a comprehensive research design based on the findings of this and similar studies. The field of textual analysis for fraud detection (as well as in other areas of accounting and finance) offers great possibilities to study corporate disclosure and to ensure that its standards are high.

References

Abbasi, A., Albrecht, C. C., Vance, A., & Hansen, J. (2012). MetaFraud: A Meta-Learning Framework for Detecting Financial Fraud. *MIS Quarterly*, *36*(4), 1293–1327.

Abbott, L. J., Park, Y., & Parker, S. (2000). The Effects of Audit Committee Activity and Independence on Corporate Fraud. *Managerial Finance*, *26*(11), 55–68.

Abdullahi, R., & Mansor, N. (2015). Fraud Triangle Theory and Fraud Diamond Theory. Understanding the Convergent and Divergent For Future Research. *International Journal of Academic Research in Accounting, Finance and Management Sciences*, *5*(4), 38–45.

Abrahamson, E., & Park, C. (1994). Concealment of negative organizational outcomes: An agency theory perspective. *Academy of Management Journal*, *37*(5), 1302–1334.

Aggarwal, C. C., & Zhai, C. (2012). *Mining Text Data*. New York, NY: Springer.

Akerlof, G. A. (1970). The Market for "Lemons": Quality Uncertainty and the Market Mechanism. *The Quarterly Journal of Economics*, *84*(3), 488–500.

Albrecht, C. C., Albrecht, W. S., & Dunn, J. G. (2001). Can Auditors Detect Fraud: A Review of the Research Evidence. *Journal of Forensic Accounting*, *2*, 1–12.

Albrecht, W. S., Albrecht, C. O., Albrecht, C. C., & Zimbelman, M. F. (2016). *Fraud Examination* (5th edition). Boston, MA: Cengage Learning.

Albrecht, W. S., Howe, K. R., & Romney, M. B. (1984). *Deterring Fraud: The Internal Auditor's Perspective*. Altamonte Springs, FL: Institute of Internal Auditors Research Foundation.

Albrecht, W. S., & Hoopes, J. L. (2014). Why Audits Cannot Detect All Fraud: Real Examples and Insights from an Expert Witness. *The CPA Journal*, *84*(10), 12–21.

Alden, M. E., Bryan, D. M., Lessley, B. J., & Tripathy, A. (2012). Detection of Financial Statement Fraud Using Evolutionary Algorithms. *Journal of Emerging Technologies in Accounting*, *9*(1), 71–94.

Anderson, E. (1935). The Irises of the Gaspe Peninsula. *Bulletin of the American Iris Society*, *59*, 2–5.

Arens, A. A., Loebbecke, J. K., Elder, R. J., & Beasley, M. S. (2000). *Auditing: An Integrated Approach* (8th edition). Upper Saddle River, NJ: Prentice Hall.

Arrow, K. J. (1963). Uncertainty and the Welfare Economics of Medical Care. *The American Economic Review*, *53*(5), 941–973.

Baker, T. (1996). On the Genealogy of Moral Hazard. *Texas Law Review*, *75*(2), 237–292.

Bamber, L. S., Jiang, J., & Wang, I. Y. (2010). What's My Style? The Influence of Top Managers on Voluntary Corporate Financial Disclosure. *The Accounting Review*, *85*(4), 1131–1162.

Barber, D. (2012). *Bayesian Reasoning and Machine Learning* (Online Version). Cambridge, GB: Cambridge University Press.

Baudot, L. (2014). GAAP convergence or convergence Gap: unfolding ten years of accounting change. *Accounting, Auditing & Accountability Journal*, *27*(6), 956–994.

Bayes, T., & Price, R. (1763). An Essay towards Solving a Problem in the Doctrine of Chances. *Philosophical Transactions*, *53*, 370–418.

Bazeley, P. (2004). Issues in Mixing Qualitative and Quantitative Approaches to Research. In R. Buber, J. Gadner, & L. Richards (Eds.), *Applying Qualitative Methods to Marketing Management Research* (pp. 141–156). Basingstoke, GB: Palgrave Macmillan.

Beasley, M. S. (1996). An Empirical Analysis of the Relation Between the Board of Director Composition and Financial Statement Fraud. *The Accounting Review*, *71*(4), 443–465.

Beasley, M. S. (2003). SAS No. 99: A New Look at Auditor Detection of Fraud. *Journal of Forensic Accounting*, *4*(1), 1–20.

Beasley, M. S., Carcello, J. V., Hermanson, D. R., & Lapides, P. D. (2000). Fraudulent Financial Reporting: Consideration of Industry Traits and Corporate Governance Mechanisms. *Accounting Horizons*, *14*(4), 441–454.

Bell, T. B., & Carcello, J. V. (2000). A Decision Aid for Assessing the Likelihood of Fraudulent Financial Reporting. *Auditing: A Journal of Practice & Theory*, *19*(1), 169–184.

Beneish, M. D. (1997). Detecting GAAP violation: implications for assessing earnings management among firms with extreme financial performance. *Journal of Accounting and Public Policy*, *16*(3), 271–309.

Beneish, M. D. (1999). The Detection of Earnings Manipulation. *Financial Analysts Journal*, *55*(5), 24–36.

Benford, F. (1938). The Law of Anomalous Numbers. *Proceedings of the American Philosophical Society*, *78*(4), 551–572.

Ben-Hur, A., & Weston, J. (2010). A User's Guide to Support Vector Machines. In O. Carugo & F. Eisenhaber (Eds.), *Methods in Molecular Biology: Vol. 609. Data Mining Techniques for the Life Sciences* (pp. 223–239). New York, NY: Humana Press.

Benston, G. J. (2006). Fair-value accounting: A cautionary tale from Enron. *Journal of Accounting and Public Policy*, *25*(4), 465–484.

Bergstra, J., & Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, *13*(Feb.), 281–305.

Bierstaker, J. L., Brody, R. G., & Pacini, C. (2006). Accountants' perceptions regarding fraud detection and prevention methods. *Managerial Auditing Journal*, *21*(5), 520–535.

Bishop, T. J.F. (2004). Preventing, Deterring, and Detecting Fraud: What Works and What Doesn't. *Journal of Investment Compliance*, *5*(2), 120–127.

Block, M. K., & Gerety, V. E. (1995). Some Experimental Evidence on Differences between Student and Prisoner Reactions to Monetary Penalties and Risk. *The Journal of Legal Studies*, *24*(1), 123–138.

Bloomfield, R. J. (2002). The "Incomplete Revelation Hypothesis" and Financial Reporting. *Accounting Horizons*, *16*(3), 233–243.

Boo, E., & Simnett, R. (2002). The Information Content of Management's Prospective Comments in Financially Distressed Companies: A Note. *Abacus*, *38*(2), 280–295.

Brazel, J. F., Jones, K. L., & Zimbelman, M. F. (2009). Using Nonfinancial Measures to Assess Fraud Risk. *Journal of Accounting Research*, *47*(5), 1135–1166.

Brown, N. C., Crowely, R. M., & Elliott, W. B. (2018). What are You Saying? Using Topic to Detect Financial Misreporting. *27th Annual Conference on Financial Economics and Accounting Paper, Toronto, CA.* Advance online publication.

Buller, D. B., & Burgoon, J. K. (1996). Interpersonal Deception Theory. *Communication Theory*, *6*(3), 203–242.

Byrnes, P., Al-Awadhi, A., Gullvist, B., Brown-Liburd, H., Teeter, R., Warren, J., & Vasarhelyi, M. (2018). Evolution of Auditing: From the Traditional Approach to the Future Audit. In Chan, D. Chiu, V. & M. Vasarhelyi (Eds.), *Continuous Auditing. Rutgers Studies in Accounting Analytics* (pp. 285–297). Bingley, GB: Emerald Group Publishing.

Carney, W. J. (1989). The Limits of the Fraud on the Market Doctrine. *The Business Lawyer*, *44*(4), 1259–1292.

Cavnar, W. B., & Trenkle, J. M. (1994). N-Gram-Based Text Categorization. *Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval (SDAIR-94), Las Vegas, USA*, 161–175.

Cecchini, M., Aytug, H., Koehler, G. J., & Pathak, P. (2010a). Making words work: Using financial text as a predictor of financial events. *Decision Support Systems*, *50*(1), 164–175.

Cecchini, M., Aytug, H., Koehler, G. J., & Pathak, P. (2010b). Detecting Management Fraud in Public Companies. *Management Science*, *56*(7), 1146–1160.

Chapelle, O., Schölkopf, B., & Zien, A. (2010). *Semi-Supervised Learning* (1st edition). Cambridge, MA: The MIT Press.

Chen, C. J. P., Ding, Y., & Xu, B. (2014). Convergence of Accounting Standards and Foreign Direct Investment. *The International Journal of Accounting*, *49*(1), 53–86.

Chen, S. (2016). Detection of fraudulent financial statements using the hybrid data mining approach. *SpringerPlus*, *5*(89), 1–16.

Chen, Y., Hu, G., Lin, L., & Xiao, M. (2015). GAAP Difference or Accounting Fraud? Evidence from Chinese Reverse Mergers Delisted from U.S. Markets. *Journal of Forensic and Investigative Accounting*, *7*(1), 122–145.

Chen, Y.-J., Wu, C.-H., Chen, Y.-M., Li, H.-Y., & Chen, H.-K. (2017). Enhancement of fraud detection for narratives in annual reports. *International Journal of Accounting Information Systems*, *26*(C), 32–45.

Christie, R., & Geis, F. (1970). *Studies in Machiavellianism*. New York, NY: Academic Press.

Christopher, J., Sarens, G., & Leung, P. (2009). A critical analysis of the independence of the internal audit function: Evidence from Australia. *Accounting, Auditing & Accountability Journal*, *22*(2), 200–220.

Chychyla, R., & Kogan, A. (2015). Using XBRL to Conduct a Large-Scale Study of Discrepancies between the Accounting Numbers in Compustat and SEC 10-K Filings. *Journal of Information Systems*, *29*(1), 37–72.

Chye Koh, H., & Woo, E.-S. (1998). The expectation gap in auditing. *Managerial Auditing Journal*, *13*(3), 147–154.

Coates, J. C., IV (2007). The Goals and Promise of the Sarbanes-Oxley Act. *Journal of Economic Perspectives*, *21*(1), 91–116.

Coleman, J. W. (1987). Toward an Integrated Theory of White-Collar Crime. *American Journal of Sociology*, *93*(2), 406–439.

Coomans, D., & Massart, D. L. (1982). Alternative k-Nearest Neighbour Rules in Supervised Pattern Recognition: Part 1. k-Nearest Neighbour Classification by Using Alternative Voting Rules. *Analytica Chimica Acta*, *136*, 15–27.

Cooper, S., & Peterson, C. (1980). Machiavellianism and Spontaneous Cheating in Competition. *Journal of Research in Personality*, *14*(1), 70–75.

Coram, P., Ferguson, C., & Moroney, R. (2008). Internal audit, alternative internal audit structures and the level of misappropriation of assets fraud. *Accounting & Finance*, *48*(4), 543–559.

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, *20*(3), 273–297.

Courtis, J. K. (1998). Annual report readability variability: tests of the obfuscation hypothesis. *Accounting, Auditing & Accountability Journal*, *11*(4), 459–472.

Cowans, P. J. (2006). *Probabilistic Document Modelling*: University of Cambridge, Dissertation.

Cox, R. A.K., & Weirich, T. R. (2002). The stock market reaction to fraudulent financial reporting. *Managerial Auditing Journal*, *17*(7), 374–382.

Cressey, D. R. (1950). The Criminal Violation of Financial Trust. *American Sociological Review*, *15*(6), 738–743.

Cressey, D. R. (1953). *Other People's Money: A Study in the Social Psychology of Embezzlement*. New York, NY: Free Press.

Cullinan, C. P., Earley, C. E., & Roush, P. B. (2013). Multiple Auditing Standards and Standard Setting: Implications for Practice and Education. *Current Issues in Auditing*, *7*(1), C1–C10.

Cybenko, G. (1989). Approximation by Superpositions of a Sigmoidal Function. *Mathematics of Control, Signals and Systems*, *2*(4), 303–314.

Dalton, D. R., Hitt, M. A., Certo, S. T., & Dalton, C. M. (2007). The Fundamental Agency Problem and Its Mitigation. *The Academy of Management Annals*, *1*(1), 1–64.

Daske, H. (2006). Economic Benefits of Adopting IFRS or US-GAAP - Have the Expected Cost of Equity Capital Really Decreased? *Journal of Business Finance*, *33*(3-4), 329–373.

Davidson, R. H., Dey, A., & Smith, A. J. (2019). CEO Materialism and Corporate Social Responsibility. *The Accounting Review*, *94*(1), 101–126.

Davis, A. K., Piger, J. M., & Sedor, L. M. (2012). Beyond the Numbers: Measuring the Information Content of Earnings Press Release Language. *Contemporary Accounting Research*, *29*(3), 845–868.

Dechow, P. M., Ge, W., Larson, C. R., & Sloan, R. G. (2011). Predicting Material Accounting Misstatements. *Contemporary Accounting Research*, *28*(1), 17–82.

Dechow, P. M., Sloan, R. G., & Sweeney, A. P. (1995). Detecting Earnings Management. *The Accounting Review*, *70*(2), 193–225.

Dechow, P. M., Sloan, R. G., & Sweeney, A. P. (1996). Causes and Consequences of Earnings Manipulation: An Analysis of Firms Subject to Enforcement Actions by the SEC. *Contemporary Accounting Research*, *13*(1), 1–36.

Dellaportas, S. (2013). Conversations with inmate accountants: Motivation, opportunity and the fraud triangle. *Accounting Forum*, *37*(1), 29–39.

Dikmen, B., & Küçükkocaoğlu, G. (2010). The Detection of Earnings Manipulation: The Three-Phase Cutting Plane Algorithm using Mathematical Programming. *Journal of Forecasting*, *29*(5), 442-466.

Dixon, M. I. (1995). The Re-Defining of White Collar Crime. *Dickinson Journal of International Law*, *13*(3), 561–566.

Domingos, P. (1999). The Role of Occam's Razor in Knowledge Discovery. *Data Mining and Knowledge Discovery*, *3*(4), 409–425.

Dong, W., Liao, S., & Liang, L. (2016). FIinancial Statement Fraud Detection using Text Mining: A Systemic Functional Linguistic Theory Perspective. *Proceedings of the 20th Pacific Asia Conference on Information Systems (PACIS 2016), Chiayi, TW.*

Dorminey, J., Fleming, A. S., Kranacher, M.-J., & Riley, R. A. (2012). The Evolution of Fraud Theory. *Issues in Accounting Education*, *27*(2), 555–579.

Dorminey, J. W., Fleming, A. S., Kranacher, M.-J., & Riley Jr., R. A. (2010). Beyond the Fraud Triangle. *The CPA Journal*, *80*(7), 17–23.

Drank, P. D., & Nigrini, M. J. (2000). Computer assisted analytical procedures using Benford's Law. *Journal of Accounting Education*, *18*(2), 127–146.

Durtschi, C., Hillison, W., & Pacini, C. (2004). The Effective Use of Benford's Law to Assist in Detecting Fraud in Accounting Data. *Journal of Forensic Accounting*, *5*, 17–34.

Dyck, I. J. A., Morse, A., & Zingales, L. (2013). How Pervasive is Corporate Fraud? *Rotman School of Management Working Paper No. 2222608.* Advance online publication.

Dyck, I. J. A., Morse, A., & Zingales, L. (2017). How Pervasive is Corporate Fraud? *Working Paper, NYU Law*.

Dyer, T., Lang, M., & Stice-Lawrence, L. (2017). The evolution of 10-K textual disclosure: Evidence from Latent Dirichlet Allocation. *Journal of Accounting and Economics*, *64*(2-3), 221–245.

Easley, D., & O'Hara, M. (2004). Information and the Cost of Capital. *The Journal of Finance*, *59*(4), 1553–1583.

Eining, M. M., Jones, D. R., & Loebbecke, J. K. (1997). Reliance on Decision Aids: An Examination of Auditors' Assessment of Management Fraud. *Auditing: A Journal of Practice & Theory*, *16*(2), 1–19.

Eisenhardt, K. M. (1988). Agency- and Institutional-Theory Explanations: The Case of Retail Sales Compensation. *Academy of Management Journal*, *31*(3), 488–511.

Elliott, R. K., & Willingham, J. J. (Eds.). (1980). *Management Fraud: Detection and Deterrence*. Princeton, NJ: Petrocelli Books.

Elsayed, A. (2017). Fraud Theories: Explanation of Financial Statement Fraud. *Working Paper.* Advance online publication.

Epstein, M. J., & Geiger, M. (1994). Investor Views of Audit Assurance: Recent Evidence of the Expectation Gap. *Journal of Accountancy*, *178*(5), 60–66.

Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. *The Journal of Finance*, *25*(2), 383–417.

Fama, E. F., & Jensen, M. C. (1983). Separation of Ownership and Control. *Journal of Law and Economics*, *26*(2), 301–325.

Fanning, K. M., & Cogger, K. O. (1998). Neural Network Detection of Management Fraud using Published Financial Data. *International Journal of Intelligent Systems in Accounting, Finance & Management*, *7*(1), 21–41.

Farber, D. B. (2005). Restoring Trust after Fraud: Does Corporate Governance Matter? *The Accounting Review*, *80*(2), 539-561.

Farrell, B. R., & Healy, P. (2000). White Collar Crime: A Profile of the Perpetrator and an Evaluation of the responsibilities for its Prevention and Detection. *Journal of Forensic Accounting*, *1*, 17–34.

Fawcett, T. (1997). Adaptive Fraud Detection. *Data Mining and Knowledge Discovery*, *1*(3), 291–316.

Feldman, R., & Sanger, J. (2007). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge, GB: Cambridge University Press.

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

Fischel, D. (1989). Efficient Capital Markets the Crash and the Fraud on the Market Theory. *Cornell Law Review*, *74*(5), 907–922.

Fisher, I. E., Garnsey, M. R., & Hughes, M. E. (2016). Natural Language Processing in Accounting, Auditing and Finance: A Synthesis of the Literature with a Roadmap for Future Research. *Intelligent Systems in Accounting, Finance and Management*, *23*(3), 157–214.

Fisher, R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, *7*(2), 179–188.

Fleming, A. S., Hermanson, D. R., Kranacher, M.-J., & Riley, R. A. (2016). Financial Reporting Fraud: Public and Private Companies. *Journal of Forensic Accounting Research*, *1*(1), A27–A41.

Flesher, D. L. (1996). *Internal Auditing – Standards and Practices*. Altamonte Springs, FL: The Institute of Internal Auditors.

Fletcher, R. (2010). The Sequential Quadratic Programming Method. In I. M. Bomze, V. F. Demjanov, R. Fletcher, T. Terlaky, G. Di Pillo, & F. Schoen (Eds.), *Lecture Notes in Mathematics: Vol. 1989. Nonlinear Optimization* (pp. 165–214). Berlin, DE: Springer.

Friesen, L. (2012). Certainty of Punishment versus Severity of Punishment: An Experimental Investigation. *Southern Economic Journal*, *79*(2), 399–421.

Gaganis, C. (2009). Classification Techniques for the Identification of Falsified Financial Statements: A Comparative Analysis. *Intelligent Systems in Accounting, Finance and Management*, *16*(3), 207–229.

Gao, L., & Srivastava, R. P. (2007). The Anatomy of Management Fraud Schemes: Analyses and Implications. *Indian Accounting Review*, *15*(1), 1–27.

Gee, S. (2015). *Fraud and Fraud Detection: A Data Analytics Approach*. Hoboken, NJ: John Wiley & Sons.

Geerts, G. L. (2011). A Design Science Research Methodology and its Application to Accounting Information Systems Research. *International Journal of Accounting Information Systems*, *12*(2), 142–151.

Gerard, J. A., & Weber, C. M. (2014). How Agency Theory Informs a $30 Million Fraud. *Journal of Finance, Accounting & Management*, *5*(1), 16–47.

Gilson, R. J. (2001). Globalizing Corporate Governance: Convergence of Form or Function. *The American Journal of Comparative Law*, *49*(2), 329–357.

Goel, S., & Gangolly, J. S. (2012). Beyond The Numbers: Mining The Annual Reports For Hidden Cues Indicative Of Financial Statement Fraud. *Intelligent Systems in Accounting, Finance and Management*, *19*(2), 75–89.

Goel, S., Gangolly, J. S., Faerman, S. R., & Uzuner, O. (2010). Can Linguistic Predictors Detect Fraudulent Financial Filings? *Journal of Emerging Technologies in Accounting*, *7*(1), 25–46.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. Cambridge, MA: The MIT Press.

Gordon, R., & Bovenberg, A. L. (1996). Why is Capital so Immobile Internationally? Possible Explanations and Implications for Capital Income Taxation. *The American Economic Review*, *86*(5), 1057–1075.

Gottschalk, P. (2018). *Investigating White-Collar Crime: Evaluation of Fraud Examinations*. Cham, CH: Springer International Publishing.

Green, B. P., & Choi, J. H. (1997). Assessing the Risk of Management Fraud Through Neural Network Technology. *Auditing: A Journal of Practice & Theory*, *16*(1), 14–28.

Greller, M. M. (1980). Management Fraud: Its Social Psychology and Relation to Management Fraud. In R. K. Elliott & J. J. Willingham (Eds.), *Management Fraud: Detection and Deterrence* (pp. 171–184). Princeton, NJ: Petrocelli Books.

Griffin, P. A. (2003). Got Information? Investor Response to Form 10-K and Form 10-Q EDGAR Filings. *Review of Accounting Studies*, *8*(4), 433–460.

Grove, H., & Cook, T. (2007). Fraudulent Financial Reporting Detection: Corporate Governance Red Flags. *Corporate Ownership & Control*, *4*(4), 254–261.

Grove, H., Cook, T., Streeper, E., & Throckmorton, G. (2010). Bankruptcy and Fraud Analysis: Shorting and Selling Stocks. *Journal of Forensic and Investigative Accounting*, *2*(2), 276–293.

Grüning, M. (2011). Artificial Intelligence Measurement of Disclosure (AIMID). *European Accounting Review*, *20*(3), 485–519.

Gunnthorsdottir, A., McCabe, K., & Smith, V. (2002). Using the Machiavellianism instrument to predict trustworthiness in a bargaining game. *Journal of Economic Psychology*, *23*(1), 49–66.

Hake, E. R. (2016). Financial Illusion: Accounting for Profits in an Enron World. *Journal of Economic Issues*, *39*(3), 595–611.

Hanczar, B., Hua, J., Sima, C., Weinstein, J., Bittner, M., & Dougherty, E. R. (2010). Small-sample precision of ROC-related estimates. *Bioinformatics*, *26*(6), 822–830.

Hansen, J. V., McDonald, J. B., Messier, W. F., JR., & Bell, T. B. (1996). A Generalized Qualitative-response Model and the Analysis of Management Fraud. *Management Science*, *42*(7), 1022–1032.

Hansen, L. L. (2009). Corporate financial crime: Social diagnosis and treatment. *Journal of Financial Crime*, *16*(1), 28–40.

Haque, F., Arun, T., & Kirkpatrick, C. (2008). Corporate Governance and Financial Markets: A Conceptual Framework. *Corporate Ownership and Control*, *5*(2), 264–276.

Harding, N., & Mckinnon, J. (1997). User Involvement in the Standard-Setting Process: A Research Note on the Congruence of Accountant and User Perceptions of Decision Usefulness. *Accounting, Organizations and Society*, *22*(1), 55–67.

Hassanat, A. B., Abbadi, M. A., Altarawneh, G. A., & Alhasanat, A. A. (2014). Solving the Problem of the K Parameter in the KNN Classifier Using an Ensemble Learning

Approach. *International Journal of Computer Science and Information Security*, *12*(8), 33–39.

Hastie, T. J., Tibshirani, R. J., & Friedman, J. (2016). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd edition). New York, NY: Springer.

Heath, J. (2008). Business Ethics and Moral Motivation: A Criminological Perspective. *Journal of Business Ethics*, *83*(4), 595–614.

Henselmann, K., Scherr, E., & Ditter, D. (2013). Applying Benford's Law to individual financial reports: An empirical investigation on the basis of SEC XBRL filings. *Working Papers in Accounting Valuation Auditing, No. 2012-1 [rev.]*.

Herz, R. H., & Petrone, K. R. (2004). International Convergence of Accounting Standards - Perspectives from the FASB on Challenges and Opportunities. *Northwestern Journal of International Law & Business*, *25*(3), 631–660.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design Science in Information Systems Research. *MIS Quarterly*, *28*(1), 75–105.

Hill, C. W. L., & Jones, T. M. (1992). Stakeholder-Agency Theory. *Journal of Management Studies*, *29*(2), 131–154.

Hoberg, G., & Lewis, C. (2017). Do fraudulent firms produce abnormal disclosure? *Journal of Corporate Finance*, *43*(C), 58–85.

Hofstedt, T. R. (1976). Behavioral accounting research: Pathologies, paradigms and prescriptions. *Accounting, Organizations and Society*, *1*(1), 43–58.

Hornik, K. (1991). Approximation Capabilities of Multilayer Feedforward Networks. *Neural Networks*, *4*(2), 251–257.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd edition). *Wiley Series in Probability and Statistics*. New York, NY: John Wiley & Sons.

Hsu, C.-W., Chang, C.-C., & Lin, C.-J. (2003). A Practical Guide to Support Vector Classification. *Department of Computer Science, National Taiwan University*.

Huang, S. Y., Lin, C.-C., Chiu, A.-A., & Yen, D. C. (2017). Fraud detection using fraud triangle risk factors. *Information Systems Frontiers*, *19*(6), 1343–1356.

Huber, W. D. (2017). Forensic Accounting, Fraud Theory, and the End of the Fraud Triangle. *Journal of Theoretical Accounting Research*, *12*(2), 28–49.

Humpherys, S. L., Moffitt, K. C., Burns, M. B., Burgoon, J. K., & Felix, W. F. (2011). Identification of fraudulent financial statements using linguistic credibility analysis. *Decision Support Systems*, *50*(3), 585–594.

Hutter, F., Kotthoff, L., & Vanschoren, J. (2019). *Automated Machine Learning: Methods, Systems, Challenges*. *The Springer Series on Challenges in Machine Learning*. Cham, CH: Springer International Publishing.

Hutter, F., Lücke, J., & Schmidt-Thieme, L. (2015). Beyond Manual Tuning of Hyperparameters. *KI - Künstliche Intelligenz*, *29*(4), 329–337.

Jackson, C. W. (2015). *Detecting Accounting Fraud: Analysis and Ethics* (1st global edition). Harlow, GB: Pearson Education.

Jain, A. K., Mao, J., & Mohiuddin, K. M. (1996). Artificial neural networks: A tutorial. *Computer*, *29*(3), 31–44.

Jain, T., & Jamali, D. (2016). Looking Inside the Black Box: The Effect of Corporate Governance on Corporate Social Responsibility. *Corporate Governance: An International Review*, *24*(3), 253–273.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An Introduction to Statistical Learning: With Applications in R* (Corrected 8th printing). *Springer Texts in Statistics*. New York, NY: Springer.

Jans, M., Lybaert, N., & Vanhoof, K. (2009). A Framework for Internal Fraud Risk Reduction at IT Integrating Business Processes: The IFR² Framework. *The International Journal of Digital Accounting Research*, *9*, 1–29.

Janvrin, D. J., Payne, E. A., Byrnes, P., Schneider, G. P., & Curtis, M. B. (2012). The Updated COSO Internal Control—Integrated Framework: Recommendations and Opportunities for Future Research. *Journal of Information Systems*, *26*(2), 189–213.

Jensen, M. C., & Meckling, W. H. (1976). Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure. *Journal of Financial Economics*, *3*(4), 305–360.

Jizi, M. I., Salama, A., Dixon, R., & Stratling, R. (2014). Corporate Governance and Corporate Social Responsibility Disclosure: Evidence from the US Banking Sector. *Journal of Business Ethics*, *125*(4), 601–615.

John, G. H., & Langley, P. (1995). Estimating Continuous Distributions in Bayesian Classifiers. *Proceeding UAI'95 Proceedings of the Eleventh conference on Uncertainty in artificial intelligence, Montreal, CA*, 338–345.

Jovanovic, F., Andreadakis, S., & Schinckus, C. (2016). Efficient market hypothesis and fraud on the market theory a new perspective for class actions. *Research in International Business and Finance*, *38*, 177–190.

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

Karlik, B., & Olgac, A. V. (2011). Performance Analysis of Various Activation Functions in Generalized MLP Architectures of Neural Networks. *International Journal of Artificial Intelligence and Expert Systems*, *1*(4), 111–122.

Karpoff, J. M., Koester, A., Lee, D. S., & Martin, G. S. (2017). Proxies and Databases in Financial Misconduct Research. *The Accounting Review*, *92*(6), 129–163.

Kassem, R., & Higson, A. (2012). The new fraud triangle model. *Journal of Emerging Trends in Economics and Management Sciences*, *3*(3), 191–195.

Kavzoglu, T. (1999). Determining Optimum Structure for Artificial Neural Networks. *In Proceedings of the 25th Annual Technical Conference and Exhibition of the Remote Sensing Society, Cardiff, GB*, 675–682.

Keasey, K., Thompson, S., & Wright, M. (2005). Introduction. In K. Keasey, S. Thompson, & M. Wright (Eds.), *Corporate Governance: Accountability, Enterprise and International Comparisons* (pp. 1–19). Hoboken, NJ: John Wiley & Sons.

Kinney, W. R., & McDaniel, L. S. (1989). Characteristics of Firms Correcting Previously Reported Quarterly Earnings. *Journal of Accounting and Economics*, *11*(1), 71–93.

Kinney, W. R., & Nelson, M. W. (1996). Outcome Information and the "Expectation Gap": The Case of Loss Contingencies. *Journal of Accounting Research*, *34*(2), 281–299.

Kirkos, E., Spathis, C. T., & Manolopoulos, Y. (2007). Data Mining techniques for the detection of fraudulent financial statements. *Expert Systems with Applications*, *32*(4), 995–1003.

Korsmo, C. (2014). Market Efficiency and Fraud on the Market: The Promise and Peril of Halliburton. *Lewis & Clark Law Review*, *18*(4), 827–892.

Kotsiantis, S., Koumanakos, E., Tzelepis, D., & Tampakas, V. (2006). Forecasting Fraudulent Financial Statements using Data Mining. *International Journal of Computational Intelligence*, *3*(2), 104–110.

Kotu, V., & Deshpande, B. (2014). *Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner*. Amsterdam, NL: Elsevier.

Kowalski, G. J., & Maybury, M. T. (2000). *Information Storage and Retrieval Systems: Theory and Implementation* (2nd edition). Boston, MA: Kluwer.

Kranacher, M.-J., Riley, R., & Wells, J. T. (op. 2011). *Forensic Accounting and Fraud Examination* (2nd edition). Hoboken, NJ: John Wiley & Sons.

Krogh, A., & Hertz, J. (1992). A Simple Weight Decay Can Improve Generalization. In J. E. Moody, S. J. Hanson, & R. P. Lippmann (Eds.), *Advances in Neural Information Processing Systems 4* (pp. 950–957). San Mateo, CA: Morgan Kaufmann Publishers.

Lambert, R., Leuz, C., & Verrecchia, R. E. (2007). Accounting Information, Disclosure, and the Cost of Capital. *Journal of Accounting Research*, *45*(2), 385–420.

Lawrence, A. (2013). Individual investors and financial disclosure. *Journal of Accounting and Economics*, *56*(1), 130–147.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444.

Lee, C., & Lee, G. G. (2006). Information gain and divergence-based feature selection for machine learning-based text categorization. *Information Processing & Management*, *42*(1), 155–165.

Lee, T. A., Ingram, R. W., & Howard, T. P. (1999). The Difference between Earnings and Operating Cash Flow as an Indicator of Financial Reporting Fraud. *Contemporary Accounting Research*, *16*(4), 749–786.

Lee, Y.-J. (2012). The Effect of Quarterly Report Readability on Information Efficiency of Stock Prices. *Contemporary Accounting Research*, *29*(4), 1137–1170.

Lehavy, R., Li, F., & Merkley, K. (2011). The Effect of Annual Report Readability on Analyst Following and the Properties of Their Earnings Forecasts. *The Accounting Review*, *86*(3), 1087–1115.

Lendez, A. M., & Lorevec, J. J. (1999). How to Prevent and Detect Financial Statement Fraud. *Journal of Corporate Accounting & Finance*, *11*(1), 47–54.

Leuz, C. (2003). IAS Versus U.S. GAAP: Information Asymmetry-Based Evidence from Germany's New Market. *Journal of Accounting Research*, *41*(3), 445–472.

Lev, B., & Ohlson, J. A. (1982). Market-Based Empirical Research in Accounting: A Review, Interpretation, and Extension. *Journal of Accounting Research*, *20*(Supplement), 249–322.

Lewis, C., & Young, S. (2019). Fad or Future? Automated Analysis of Financial Text and its Implications for Corporate Reporting. *Accounting and Business Research*, *49*(5), 587–615.

Li, E. X., & Ramesh, K. (2009). Market Reaction Surrounding the Filing of Periodic SEC Reports. *The Accounting Review*, *84*(4), 1171–1208.

Li, F. (2008). Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and Economics*, *45*(2-3), 221–247.

Li, F. (2010a). Managers' Self-Serving Attribution Bias and Corporate Financial Policies. *SSRN Electronic Journal.* Advance online publication.

Li, F. (2010b). The Information Content of Forward-Looking Statements in Corporate Filings-A Naïve Bayesian Machine Learning Approach. *Journal of Accounting Research*, *48*(5), 1049–1102.

Liggio, C. D. (1974). The Expectation Gap: The Accountant's Legal Waterloo? *Journal of Contemporary Business*, *3*(3), 27–44.

Lin, J. W., Hwang, M. I., & Becker, J. D. (2003). A fuzzy neural network for assessing the risk of fraudulent financial reporting. *Managerial Auditing Journal*, *18*(8), 657–665.

Lin, S., Pizzini, M., Vargus, M., & Bardhan, I. R. (2011). The Role of the Internal Audit Function in the Disclosure of Material Weaknesses. *The Accounting Review*, *86*(1), 287–323.

Lindgren, H. (1982). The Kreuger Crash of 1932 in memory of a financial genius, or was he a simple swindler? *Scandinavian Economic History Review*, *30*(3), 189–206.

Liu, C., Chan, Y., Alam Kazmi, S. H., & Fu, H. (2015). Financial Fraud Detection Model: Based on Random Forest. *International Journal of Economics and Finance*, *7*(7), 178–188.

Lobo, J. M., Jiménez-Valverde, A., & Real, R. (2008). AUC: A misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, *17*(2), 145–151.

Loebbecke, J. K., Eining, M. M., & Willingham, J. J. (1989). Auditor's Experience with Material Irregularities: Frequency, Nature, and Detectability. *Auditing: A Journal of Practice & Theory*, *Fall*, 1–28.

Lokanan, M. E. (2015). Challenges to the fraud triangle: Questions on its usefulness. *Accounting Forum*, *39*(3), 201–224.

Lou, Y.-I., & Wang, M.-L. (2009). Fraud Risk Factor Of The Fraud Triangle Assessing The Likelihood Of Fraudulent Financial Reporting. *Journal of Business & Economics Research*, *7*(2), 61–78.

Loughran, T., & McDonald, B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*, *66*(1), 35–65.

Loughran, T., & McDonald, B. (2016). Textual Analysis in Accounting and Finance: A Survey. *Journal of Accounting Research*, *54*(4), 1187–1230.

Lu, Z., Pu, H., Wang, F., Hu, Z., & Wang, L. (2017). The Expressive Power of Neural Networks: A View from the Width. *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*, 6232–6240.

Macey, J. R. (2013). *The Death of Corporate Reputation: How Integrity Has Been Destroyed on Wall Street*. Upper Saddle River, NJ: Prentice Hall.

Macey, J. R., & Miller, G. P. (1990). Good Finance, Bad Economics: An Analysis of the Fraud-on-the-Market Theory. *Stanford Law Review*, *42*(4), 1059–1092.

Mackevičius, J., & Kkazlauskienė, L. (2009). The Fraud Tree and its Investigation in Audit. *Ekonomika*, *85*(1), 90–101.

Mandrekar, J. N. (2010). Receiver Operating Characteristic Curve in Diagnostic Test Assessment. *Journal of Thoracic Oncology*, *5*(9), 1315–1316.

Maragno, L. M. D., & Borba, J. A. (2017). Conceptual map of fraud: theoretical and empirical configuration of international studies and future research opportunities. *Journal of Education and Research in Accounting, Revista de Educação e Pesquisa em Contabilidade*, *11*(3 Special Edition), 41–66.

Mason, S. J., & Graham, N. E. (2002). Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. *Quarterly Journal of the Royal Meteorological Society*, *128*(584), 2145–2166.

Mayew, W. J., Sethuraman, M., & Venkatachalam, M. (2015). MD&A Disclosure and the Firm's Ability to Continue as a Going Concern. *The Accounting Review*, *90*(4), 1621–1651.

McAfee, K., & Guth, L. (2014). Markopolizing Conversion Fraud: Understanding and Identifying Opportunities for US Financial Reporting Conversion Fraud. *Journal of Legal, Ethical and Regulatory Issues*, *17*(1), 57–65.

McCallum, A., & Nigam, K. (1998). A Comparison of Event Models for Naive Bayes Text Classification. *Learning for Text Categorization: Papers from the 1998 AAAI Workshop*, 41–48.

McGuire, J. B. (1988). Agency Theory and Organizational Analysis. *Managerial Finance*, *14*(4), 6–9.

Meeks, G., & Swann, G. P. (2009). Accounting Standards and the Economics of Standards. *Accounting and Business Research*, *39*(3), 191–210.

Merchant, K. A. (1990). The Effects of Financial Controls on Data Manipulation and Management Myopia. *Accounting, Organizations and Society*, *15*(4), 297–313.

Miner, G., Elder, J., Fast, A., Hill, T., Nisbet, R., & Delen, D. (2012). *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*. Amsterdam, NL: Academic Press.

Moyes, G. D., & Baker, C. R. (2003). Auditor's beliefs about the fraud detection effectiveness of standard audit procedures. *Journal of Forensic Accounting*, *4*, 199–216.

Murphy, P. R. (2012). Attitude, Machiavellianism and the Rationalization of Misreporting. *Accounting, Organizations and Society*, *37*(4), 242–259.

Murphy, P. R., & Dacin, M. T. (2011). Psychological Pathways to Fraud: Understanding and Preventing Fraud in Organizations. *Journal of Business Ethics*, *101*(4), 601–618.

Nakashima, M. (2017). Can The Fraud Triangle Predict Accounting Fraud? Evidence from Japan. *Working Paper, presented at the 8th Conference of the Japanese Accounting Review, Kobe, JP*.

Ndofor, H. A., Wesley, C., & Priem, R. L. (2012). Providing CEOs With Opportunities to Cheat. *Journal of Management*, *41*(6), 1774–1797.

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying Words: Predicting Deception From Linguistic Styles. *Personality & Social Psychology Bulletin*, *29*(5), 665–675.

Niederhoffer, V., & Osborne, M. F. M. (1966). Market Making and Reversal on the Stock Exchange. *Journal of the American Statistical Association*, *61*(316), 897–916.

Nigrini, M. J. (1996). A Taxpayer Compliance Application of Benford's Law. *The Journal of the American Taxation Association*, *18*(1), 72–91.

North, M. (2012). *Data Mining for the Masses*. *A Global Text Project Book*: Global Text.

Oduntan, O. E. (2018). Performance Analysis of K-Nearest Neighbour Classifier and Cosine Similarity Measure in Generating Weighted Scores for an Automated Essay Grading System. *International Journal of Scientific Engineering Research*, *6*(11), 57–62.

Ogoun, S., & Obara, L. C. (2013). Curbing Occupational and Financial Reporting Fraud: An Alternative Paradigm. *International Journal of Business and Social Science*, *4*(9), 123–132.

Othman, I. W., Hasan, H., Tapsir, R., Rahman, N. A., Tarmuji, I., Majdi, S., . . . Omar, N. (2012). Text Readability and Fraud Detection. *Proceedings of IEEE: Symposium on Business, Engineering and Industrial Applications (ISBEIA), Bandung, ID*, 296–301.

Paulhus, D. L., & Williams, K. M. (2002). The Dark Triad of personality: Narcissism, Machiavellianism, and psychopathy. *Journal of Research in Personality*, *36*(6), 556–563.

Pedneault, S., Rudewicz, F., Silverstone, H., & Sheetz, M. (2012). *Forensic Accounting and Fraud Investigation for Non-Experts* (3rd edition). Hoboken, NJ: Wiley.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2014). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, *24*(3), 45–77.

Pepe, M. S. (2000). Receiver Operating Characteristic Methodology. *Journal of the American Statistical Association*, *95*(449), 308–311.

Perols, J. (2011). Financial Statement Fraud Detection: An Analysis of Statistical and Machine Learning Algorithms. *Auditing: A Journal of Practice & Theory*, *30*(2), 19–50.

Perols, J. L., Bowen, R. M., Zimmermann, C., & Samba, B. (2017). Finding Needles in a Haystack: Using Data Analytics to Improve Fraud Prediction. *The Accounting Review*, *92*(2), 221–245.

Perri, F. S., Lichtenwald, T. G., & Mieczkowska, E. M. (2014). Sutherland, Cleckley and Beyond: White-Collar Crime and Psychopathy. *International Journal of Psychological Studies*, *6*(4), 71–88.

Persons, O. S. (1995). Using Financial Statement Data to Identify Using Financial Statement Data to Identify Factors Associated with Fraudulent Financial Reporting. *Journal of Applied Business Research*, *11*(3), 38–46.

Peterson, L. (2009). K-nearest neighbor. *Scholarpedia*, *4*(2), 1883.

Phillips, V. L., Saks, M. J., & Peterson, J. L. (2001). The Application of Signal Detection Theory to Decision-Making in Forensic Science. *Journal of Forensic Sciences*, *46*(2), 294-308.

Porter, B. (1993). An Empirical Study of the Audit Expectation-Performance Gap. *Accounting and Business Research*, *24*(93), 49–68.

Prawitt, D. F., Smith, J. L., & Wood, D. A. (2009). Internal Audit Quality and Earnings Management. *The Accounting Review*, *84*(4), 1255–1280.

Purda, L., & Skillicorn, D. (2015). Accounting Variables, Deception, and a Bag of Words: Assessing the Tools of Fraud Detection. *Contemporary Accounting Research*, *32*(3), 1193–1223.

Quick, R., & Wolz, M. (2003). Benford's Law in deutschen Rechnungslegungsdaten. *Betriebswirtschaftliche Forschung und Praxis*, *55*(2), 208–224.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81–106.

Ramamoorti, S. (2008). The Psychology and Sociology of Fraud: Integrating the Behavioral Sciences Component Into Fraud and Forensic Accounting Curricula. *Issues in Accounting Education*, *23*(4), 521–533.

Ramamoorti, S., Pope, K. R., Morrison, D. E., & Koletar, J. W. (2013). *A.B.C.'s of Behavioral Forensics: Using Psychology to Prevent, Deter and Detect Fraud*. Hoboken, NJ: John Wiley & Sons.

Ravisankar, P., Ravi, V., Raghava Rao, G., & Bose, I. (2011). Detection of financial statement fraud and feature selection using data mining techniques. *Decision Support Systems*, *50*(2), 491–500.

Reardon, P. (2017). A Corporate & Securities Attorney's Comparison of Public vs. Private Companies. Retrieved from https://www.sec.gov/info/smallbus/acsec/reardon-comparison-public-vs-private-companies.pdf

Rezaee, Z. (2003). High-Quality Financial reporting: The Six-Legged Stool. *Strategic Finance*, *84*(8), 26–30.

Rezaee, Z., & Riley, R. (2009). *Financial Statement Fraud: Prevention and Detection* (2nd edition). Hoboken, NJ: John Wiley & Sons.

Robins, N. (2007). This Imperious Company: The English East India Company and its Legacy for Corporate Accountability. *Journal of Corporate Citizenship*, *25*(Spring), 31–42.

Robins, N. (2012). *The Corporation That Changed the World: How the East India Company Shaped the Modern Multinational* (2nd edition). London, GB: Pluto Press.

Rogovin, C. H., & Martens, F. T. (1992). The Evil That Men Do. *Journal of Contemporary Criminal Justice*, *8*(1), 62–79.

Rose, A. M. (2010). The Multi-Enforcer Approach To Securities Fraud Deterrence: A Critical Analysis. *University of Pennsylvania Law Review*, *158*(7), 2173–2231.

Rouf, A. (2011). Corporate characteristics, governance attributes and the extent of voluntary disclosure in Bangladesh. *African Journal of Business Management*, *5*(19), 7836–7845.

Ryan, B., Scapens, R. W., & Theobald, M. (2002). *Research Method and Methodology in Finance and Accounting* (2nd edition). London, GB: Cengage Learning.

Ryerson, F. E., III (2009). Improper Capitalization and the Management of Earnings. *Proceedings of the American Society of Business and Behavioral Sciences Annual Conference (ASBBS 2009), Las Vegas, NV*, *16*(1).

Saari, C. P. (1977). The Efficient Capital Market Hypothesis, Economic Theory and the Regulation of the Securities Industry. *Stanford Law Review*, *29*(5), 1031–1076.

Sabau, A. S. (2012). Survey of Clustering based Financial Fraud Detection Research. *Informatica Economica*, *16*(1), 110–122.

Saville, A. (2006). Using Benford’s Law to detect data error and fraud: An examination of companies listed on the Johannesburg Stock Exchange. *South African Journal of Economic and Management Sciences*, *9*(3), 341–354.

Scharff, M. M. (2005). Understanding WorldCom's Accounting Fraud: Did Groupthink Play a Role? *Journal of Leadership & Organizational Studies*, *11*(3), 109–118.

Scheff, J. D., Almon, R. R., Dubois, D. C., Jusko, W. J., & Androulakis, I. P. (2011). Assessment of Pharmacologic Area Under the Curve When Baselines are Variable. *Pharmaceutical Research*, *28*(5), 1081–1089.

Scholes, M. S. (1970). *A Test of the competitive Market Hypothesis: The Market for New Issues and Secondary Offerings*: University of Chicago, Dissertation.

Schroeder, R. G., Clark, M. W., & Cathey, J. M. (2019). *Financial Accounting Theory and Analysis: Text and Cases* (13th edition). Hoboken, NJ: John Wiley & Sons.

Schuchter, A., & Levi, M. (2016). The Fraud Triangle revisited. *Security Journal*, *29*(2), 107–121.

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge, GB: Cambridge University Press.

Shapiro, S. P. (1990). Collaring the Crime, not the Criminal: Reconsidering the Concept of White-Collar Crime. *American Sociological Review*, *55*(3), 346–365.

Sharma, V. D. (2004). Board of Director Characteristics, Institutional Ownership, and Fraud: Evidence from Australia. *Auditing: A Journal of Practice & Theory*, *23*(2), 105–117.

Shi, W., Connelly, B. L., & Hoskisson, R. E. (2017). External Corporate Governance and Financial Fraud: Cognitive Evaluation Theory Insights on Agency Theory Prescriptions. *Strategic Management Journal*, *38*(6), 1268–1286.

Shil, N. C., Das, B., & Pramanik, A. K. (2009). Harmonization of Accounting Standards through Internationalization. *International Business Research*, *2*(2), 194–201.

Sidorsky, R. (2006). Assessing the Risks of Accounting Fraud. *Commercial Lending Review*, *21*(6), 9–17.

Sigletos, G., Paliouras, G., Spyropoulos, C. D., & Hatzopoulos, M. (2005). Combining Information Extraction Systems Using Voting and Stacked Generalization. *Journal of Machine Learning Research*, *6*(Nov), 1751–1782.

Singleton, T. W., & Singleton, A. J. (2010). *Fraud Auditing and Forensic Accounting* (4th Edition). Hoboken, NJ: John Wiley & Sons.

Skousen, C. J., Smith, K. R., & Wright, C. J. (2009). Detecting and Predicting Financial Statement Fraud: The Effectiveness of the Fraud Triangle and SAS No. 99. *Advances in Financial Economics*, *13*, 53–81.

Sloane, E. H. (1944). Rationalization. *The Journal of Philosophy*, *41*(1), 12–21.

Smith, A. (1776). *An Inquiry into the Nature and Causes of the Wealth of Nations, edited with an Introduction, Notes, Marginal Summary and an Enlarged Index by Edwin Cannan (1904)* (Vol. 2). London, GB: Methuen.

Smith, L. N. (2018). A Disciplined Approach to Neural Network Hyper-Parameters: Part 1 - Learning Rate, Batch Size, Momentum, and Weight Decay. *US Naval Research Laboratory*, *Technical Report 5510-026*.

Smith, M. (1996). Qualitative characteristics in accounting disclosures: a desirability trade-off. *Managerial Auditing Journal*, *11*(3), 11–16.

Smith, M., & Taffler, R. J. (2000). The chairman's statement - A content analysis of discretionary narrative disclosures. *Accounting, Auditing & Accountability Journal*, *13*(5), 624–647.

Snider, L. (1982). Traditional and Corporate Theft - A Comparison of Sanctions. In P. Wickman & T. Dailey (Eds.), *From White-Collar and Economic Crime* (pp. 235–258). New York, NY: Lexington Books.

Soh, D. S. B., & Martinov-Bennie, N. (2011). The internal audit function: Perceptions of internal audit roles, effectiveness and evaluation. *Managerial Auditing Journal*, *26*(7), 605–622.

Spathis, C. T. (2002). Detecting false financial statements using published data: some evidence from Greece. *Managerial Auditing Journal*, *17*(4), 179–191.

Spathis, C. T., Doumpos, M., & Zopounidis, C. (2002). Detecting falsified financial statements: a comparative study using multicriteria analysis and multivariate statistical techniques. *European Accounting Review*, *11*(3), 509–535.

Stotland, E. (1977). White Collar Criminals. *Journal of Social Issues*, *33*(4), 179–196.

Street, D. L., & Shaughnessy, K. A. (1998a). The Evolution of the G4 + 1 and Its Impact on International Harmonization of Accounting Standards. *Journal of International Accounting, Auditing and Taxation*, *7*(2), 131–161.

Street, D. L., & Shaughnessy, K. A. (1998b). The Quest for International Accounting Harmonization: A Review of the Standard Setting Agendas of the IASC, US, UK, Canada, and Australia, 1973–1997. *The International Journal of Accounting*, *33*(2), 179–209.

Sugiarto, T., Noorzaman, S., Madu, L., Subagyo, A., & Amiri, A.M. (2017). First Digit Lucas, Fibonacci and Benford Number in Financial Statement. In H. Xue (Ed.), *Proceedings of the 2017 International Conference on Economic Development and Education Management (ICEDEM 2017)* (pp. 22–24). Paris, FR: Atlantis Press.

Summers, S. L., & Sweeney, J. T. (1998). Fraudulently Misstated Financial Statements and Insider Trading: An Empirical Analysis. *The Accounting Review*, *73*(1), 131–146.

Sunder, S. (1988). Political Economy of Accounting Standards. *Journal of Accounting Literature*, *7*, 31–41.

Sutherland, E. H. (1939). *Principles of Criminology* (3rd edition). Chicago, IL: J.B. Lippincott Company.

Sutherland, E. H. (1947). *Principles of Criminology* (4th edition). Chicago, IL: J.B. Lippincott Company.

Sutherland, E. H. (1940). White-Collar Criminality. *American Sociological Review*, *5*(1), 1–12.

Sutherland, E. H. (1944). Is "White Collar Crime" Crime? *American Sociological Review*, *10*(2), 132–139.

Sutton, T. (2006). *Corporate Financial Accounting and Reporting* (2nd edition). Harlow, GB: Pearson Education.

Sykes, G. M., & Matza, D. (1957). Techniques of Neutralization: A Theory of Delinquency. *American Sociological Review*, *22*(6), 664–670.

Szalma, J. L., & Hancock, P. A. (2013). A Signal Improvement to Signal Detection Analysis: Fuzzy SDT on the ROCs. *Journal of Experimental Psychology. Human Perception and Performance*, *39*(6), 1741–1762.

Tarca, A. (2004). International Convergence of Accounting Practices: Choosing between IAS and US GAAP. *Journal of International Financial Management and Accounting*, *15*(1), 60–91.

Tennyson, B. M., Ingram, R. W., & Dugan, M. T. (1990). Assessing the Information Content of narrative Disclosures in explaining Bankruptcy. *Journal of Business Finance & Accounting*, *17*(3), 391–410.

Theodoridis, S. (2015). *Machine Learning: A Bayesian and Optimization Perspective*. London, GB: Academic Press.

Thomsen, S. (2003). The Convergence of Corporate Governance Systems to European and Anglo-American Standards. *European Business Organization Law Review*, *4*(1), 31–50.

Thomsen, S., & Conyon, M. J. (2012). *Corporate Governance: Mechanisms and Systems*. London, GB: McGraw-Hill.

Trotman, K. T., & Bradley, G. W. (1981). Associations between social responsibility disclosure and characteristics of companies. *Accounting, Organizations and Society*, *6*(4), 355–362.

Tunley, M. (2011). Need, greed or opportunity? An examination of who commits benefit fraud and why they do it. *Security Journal*, *24*(4), 302–319.

Tweedie, D., & Seidenstein, T. R. (2005). Setting a Global Standard: The Case for Accounting Convergence. *Northwestern Journal of International Law and Business*, *25*(3), 589–608.

Uzun, H., Szewczyk, S. H., & Varma, R. (2019). Board Composition and Corporate Fraud. *Financial Analysts Journal*, *60*(3), 33–43.

Van Aken, J. E. (2005). Management Research as a Design Science: Articulating the Research Products of Mode 2 Knowledge Production in Management. *British Journal of Management*, *16*(1), 19–36.

Vanasco, R. R. (1998). Fraud auditing. *Managerial Auditing Journal*, *13*(1), 4–71.

Vijayarani, S., Ilamathi, M., & Nithya, M. (2015). Preprocessing Techniques for Text Mining-An Overview. *International Journal of Computer Science & Communication Networks*, *5*(1), 7–16.

Volmer, P. B., Werner, J. R., & Zimmermann, J. (2007). New governance modes for Germany's financial reporting system: another retreat of the nation state? *Socio-Economic Review*, *5*(3), 437–465.

Vousinas, G. L. (2019). Advancing theory of fraud: The S.C.O.R.E. model. *Journal of Financial Crime*, *26*(1), 372–381.

Wallace, W. A. (1995). *Auditing*. Cincinnati, OH.: South-Western College Publishing.

Walsh, J. P., & Seward, J. K. (1990). On the Efficiency of Internal and External Corporate Control Mechanisms. *Academy of Management Review*, *15*(3), 421–458.

Wang, C., Venkatesh, S. S., & Judd, J. S. (1994). Optimal Stopping and Effective Machine Complexity in Learning. In J. D. Cowan, G. Tesauro, & J. Alspector (Eds.), *Advances in Neural Information Processing Systems 6* (pp. 303–310). Burlington, MA: Morgan Kaufmann Publishers.

Watrin, C., Struffert, R., & Ullmann, R. (2008). Benford's Law: an instrument for selecting tax audit targets? *Review of Managerial Science*, *2*(3), 219–237.

Watts, R. L., & Zimmerman, J. L. (1990). Positive Accounting Theory: A Ten Year Perspective. *The Accounting Review*, *65*(1), 131–156.

Weir, C., Laing, D., & McKnight, P. J. (2002). Internal and External Governance Mechanisms: Their Impact on the Performance of Large UK Public Companies. *Journal of Business Finance*, *29*(5-6), 579–611.

Wells, J. T. (2004). New Approaches to Fraud Deterrence: It's Time to Take a New Look at the Auditing Process. *Journal of Accountancy*, *197*(2), 72–76.

Wells, J. T. (2017). *Corporate Fraud Handbook: Prevention and Detection* (5th edition). Hoboken, NJ: John Wiley & Sons.

West, J., & Bhattacharya, M. (2016). Intelligent financial fraud detection: A comprehensive review. *Computers & Security*, *57*, 47–66.

Wilamowski, B. (2009). Neural Network Architectures and Learning Algorithms: How not to be Frustrated with Neural Networks. *IEEE Industrial Electronics Magazine*, *3*(4), 56–63.

Wilbanks, R. M., Hermanson, D. R., & Sharma, V. D. (2017). Audit Committee Oversight of Fraud Risk: The Role of Social Ties, Professional Ties, and Governance Characteristics. *Accounting Horizons*, *31*(3), 21–38.

Wildman, J. L., Salas, E., & Scott, C. P. R. (2014). Measuring Cognition in Teams: A Cross-Domain Review. *Human Factors*, *56*(5), 911–941.

Wilhelm, W. K. (2004). The Fraud Management Lifecycle Theory: A Holistic Approach to Fraud Management. *Journal of Economic Crime Management*, *2*(2), 1–38.

Wilks, T. J., & Zimbelman, M. F. (2004). Using Game Theory and Strategic Reasoning Concepts to Prevent and Detect Fraud. *Accounting Horizons*, *18*(3), 173–184.

Wiseman, R. M., Rodríguez, G. C., & Gomez-Mejia, L. (2012). Towards a Social Theory of Agency. *Journal of Management Studies*, *49*(1), 202–222.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2017). *Data Mining: Practical Machine Learning Tools and Techniques* (4th edition). Cambridge, MA: Morgan Kaufmann Publishers.

Wolfe, D., & Hermanson, D. (2004). The Fraud Diamond: Considering the Four Elements of Fraud. *The CPA Journal*, *74*(12), 38–42.

Woodward, J. D., Orlans, N. M., & Higgins, P. T. (2003). *Biometrics: Identity Assurance in the Information Age*. New York, NJ: McGraw-Hill.

Yoshikawa, T., & Rasheed, A. A. (2009). Convergence of Corporate Governance: Critical Review and Future Directions. *Corporate Governance: An International Review*, *17*(3), 388–404.

Zack, G. M. (2009). *Fair Value Accounting Fraud: New Global Risks and Detection Techniques*. Hoboken, NJ: John Wiley & Sons.

Zack, G. M. (2013). *Financial Statement Fraud: Strategies for Detection and Investigation*. Hoboken, NJ: John Wiley & Sons.

Zhou, L., Burgoon, J. K., Nunamaker, J. F., & Twitchell, D. (2004). Automating Linguistics-Based Cues for Detecting Deception in Text-Based Asynchronous Computer-Mediated Communications. *Group Decision and Negotiation*, *13*(1), 81–106.

Zhou, W., & Kapoor, G. (2011). Detecting evolutionary financial statement fraud. *Decision Support Systems*, *50*(3), 570–575.

Zurada, J. (1992). *Introduction to Artificial Neural Systems*. St. Paul, MS: West Publishing.

Appendix

The appendix contains tables with detailed results, covering design questions 5 to 8 and enhancing questions 2 and 3. The Appendix serves to avoid the convoluted description of results for the design and enhancing questions. The computations of design questions 5 to 8 and the respective hypothesis tests rely on the detailed results presented in Tables 31-33.

To create a comparable basis for the quantitative feature vector enhancements, a hyperparameter tuning akin to the initial three feature vectors has been carried out for an SVM and is reported in Table 34. The SVM was chosen due to its superior performance, as the results of design question 8 have shown. Table 35 supports the findings of enhancing question 3 and provides detailed results for the absolute cost saving potential.

- Tables 31-33: Detailed results for the final sample.
- Table 34: SVM hyperparameter tuning for quantitative feature enhancements.
- Table 35: Determination of absolute cost savings.

			SVM			
		Matched			Realistic	
Subsample	**Qual.**	**Quant.**	**Qual. & quant.**	**Qual.**	**Quant.**	**Qual. & quant.**
f1	0.755	0.657	0.771	0.809	0.687	0.817
f2	0.767	0.673	0.765	0.766	0.603	0.783
f3	0.748	0.698	0.719	0.799	0.680	0.808
f4	0.803	0.731	0.803	0.853	0.720	0.862
f5	0.786	0.706	0.767	0.856	0.760	0.873
f6	0.842	0.706	0.849	0.901	0.740	0.907
f7	0.837	0.619	0.889	0.931	0.746	0.930
f8	0.882	0.615	0.892	0.821	0.738	0.821
f9	0.819	0.585	0.819	0.879	0.755	0.907
f10	0.802	0.595	0.808	0.727	0.583	0.736
f11	0.762	0.634	0.777	0.787	0.537	0.776
f12	0.818	0.618	0.802	0.783	0.616	0.783

			ANN			
Subsample	**Qual.**	**Quant.**	**Qual. & quant.**	**Qual.**	**Quant.**	**Qual. & quant.**
f1	0.759	0.556	0.787	0.689	0.731	0.740
f2	0.745	0.560	0.758	0.741	0.734	0.788
f3	0.727	0.561	0.717	0.752	0.750	0.797
f4	0.794	0.533	0.822	0.744	0.738	0.816
f5	0.758	0.597	0.764	0.775	0.688	0.791
f6	0.862	0.554	0.864	0.788	0.688	0.822
f7	0.845	0.592	0.841	0.822	0.696	0.825
f8	0.841	0.588	0.861	0.751	0.623	0.721
f9	0.771	0.523	0.764	0.789	0.539	0.763
f10	0.824	0.589	0.831	0.685	0.513	0.692
f11	0.723	0.616	0.729	0.840	0.503	0.840
f12	0.769	0.510	0.773	0.738	0.614	0.781

AUC values for individual subsamples f1-f12

Subsample	Training	Holdout	Subsample	Training	Holdout
f1	1996–1998	1999	f7	2002–2004	2005
f2	1997–1999	2000	f8	2003–2005	2006
f3	1998–2000	2001	f9	2004–2006	2007
f4	1999–2001	2002	f10	2005–2007	2008
f5	2000–2002	2003	f11	2006–2008	2009
f6	2001–2003	2004	f12	2007–2009	2010

The training includes the model generation process with individual feature extraction and model training. The holdout set comprises unknown cases from the subsequent period.

Table 31 – Appendix A: Detailed ANN and SVM final results

KNN

	Matched			Realistic		
Subsample	**Qual.**	**Quant.**	**Qual. & quant.**	**Qual.**	**Quant.**	**Qual. & quant.**
f1	0.727	0.604	0.616	0.743	0.700	0.708
f2	0.711	0.608	0.619	0.723	0.688	0.684
f3	0.756	0.650	0.651	0.787	0.711	0.725
f4	0.825	0.723	0.749	0.874	0.790	0.786
f5	0.793	0.710	0.718	0.833	0.743	0.747
f6	0.857	0.676	0.651	0.905	0.747	0.747
f7	0.899	0.746	0.760	0.913	0.725	0.824
f8	0.784	0.649	0.651	0.803	0.733	0.736
f9	0.632	0.642	0.644	0.831	0.772	0.771
f10	0.822	0.624	0.638	0.627	0.605	0.604
f11	0.527	0.537	0.649	0.725	0.602	0.603
f12	0.657	0.512	0.573	0.735	0.643	0.668

NB

Subsample	**Qual.**	**Quant.**	**Qual. & quant.**	**Qual.**	**Quant.**	**Qual. & quant.**
f1	0.684	0.530	0.625	0.690	0.676	0.717
f2	0.598	0.565	0.571	0.665	0.690	0.700
f3	0.576	0.577	0.588	0.661	0.715	0.729
f4	0.591	0.519	0.538	0.691	0.650	0.678
f5	0.608	0.566	0.580	0.668	0.655	0.679
f6	0.636	0.545	0.571	0.701	0.618	0.675
f7	0.637	0.543	0.566	0.754	0.622	0.691
f8	0.690	0.616	0.664	0.647	0.675	0.686
f9	0.618	0.502	0.523	0.595	0.563	0.575
f10	0.698	0.548	0.607	0.636	0.565	0.588
f11	0.552	0.715	0.717	0.670	0.581	0.634
f12	0.748	0.506	0.537	0.737	0.572	0.601

AUC values for individual subsamples f1-f12

Subsample	Training	Holdout	Subsample	Training	Holdout
f1	1996–1998	1999	f7	2002–2004	2005
f2	1997–1999	2000	f8	2003–2005	2006
f3	1998–2000	2001	f9	2004–2006	2007
f4	1999–2001	2002	f10	2005–2007	2008
f5	2000–2002	2003	f11	2006–2008	2009
f6	2001–2003	2004	f12	2007–2009	2010

The training includes the model generation process with individual feature extraction and model training. The holdout set comprises of unknown cases from the subsequent period.

Table 32 – Appendix B: Detailed NB and KNN final results

Aggregated best results over time

	Matched			Realistic		
Subsample	**Qual.**	**Quant.**	**Qual. & quant.**	**Qual.**	**Quant.**	**Qual. & quant.**
f1	0.759	0.657	0.787	0.809	0.731	0.817
f2	0.767	0.673	0.765	0.766	0.734	0.788
f3	0.756	0.698	0.719	0.799	0.750	0.808
f4	0.825	0.731	0.822	0.874	0.790	0.862
f5	0.793	0.710	0.767	0.856	0.760	0.873
f6	0.862	0.706	0.864	0.905	0.747	0.907
f7	0.899	0.746	0.889	0.931	0.746	0.930
f8	0.882	0.649	0.892	0.821	0.738	0.821
f9	0.819	0.642	0.819	0.879	0.772	0.907
f10	0.824	0.624	0.831	0.727	0.605	0.736
f11	0.762	0.715	0.777	0.840	0.602	0.840
f12	0.818	0.618	0.802	0.783	0.643	0.783

Quantitative feature enhancement

	Matched			Realistic		
Subsample	**Qual.***	**Quant.**	**Qual. & quant.**	**Qual.***	**Quant.**	**Qual. & quant.**
f1	-	0.637	0.770	-	0.612	0.819
f2	-	0.654	0.762	-	0.605	0.776
f3	-	0.643	0.718	-	0.614	0.797
f4	-	0.686	0.786	-	0.645	0.870
f5	-	0.645	0.763	-	0.671	0.872
f6	-	0.731	0.843	-	0.700	0.908
f7	-	0.657	0.888	-	0.793	0.939
f8	-	0.690	0.884	-	0.707	0.827
f9	-	0.581	0.806	-	0.710	0.893
f10	-	0.544	0.839	-	0.602	0.744
f11	-	0.627	0.701	-	0.617	0.780
f12	-	0.586	0.779	-	0.577	0.812

AUC values for individual subsamples f1-f12

*The quantitative feature enhancements did not affect the qualitative feature vector.

Subsample	Training	Holdout	Subsample	Training	Holdout
f1	1996–1998	1999	f7	2002–2004	2005
f2	1997–1999	2000	f8	2003–2005	2006
f3	1998–2000	2001	f9	2004–2006	2007
f4	1999–2001	2002	f10	2005–2007	2008
f5	2000–2002	2003	f11	2006–2008	2009
f6	2001–2003	2004	f12	2007–2009	2010

The training includes the model generation process with individual feature extraction and model training. The holdout set comprises unknown cases from the subsequent period.

Table 33 – Appendix C: Aggregated best results over time and feature enhancements

Matched sampling approach

Boosted quantitative feature vector | **Boosted combined feature vector**

Polynomial kernel function

		n						n			
		1	2	3	4			1	2	3	4
C	-1	0.463	0.571	0.501	0.500	C	-1	0.938	0.850	0.818	0.500
	0	0.463	0.571	0.501	0.500		0	0.938	0.850	0.818	0.500
	5	0.440	0.550	0.558	0.549		5	0.918	0.896	0.874	0.500
	10	0.440	0.539	0.558	0.549		10	0.918	0.896	0.874	0.500

Dot kernel function

C	-1	0.468	C	-1	0.955
	0	0.468		0	0.955
	5	0.443		5	0.931
	10	0.443		10	0.931

Radial kernel function

		γ						γ			
		0	1	2	4			0	1	2	4
C	-1	0.500	0.560	0.510	0.513	C	-1	0.500	0.560	0.510	0.505
	0	0.500	0.560	0.510	0.573		0	0.500	0.560	0.510	0.513
	5	0.500	0.517	0.505	0.573		5	0.500	0.517	0.510	0.513
	10	0.500	0.522	0.505	0.573		10	0.500	0.522	0.505	0.516

Realistic sampling approach

Boosted quantitative feature vector | **Boosted combined feature vector**

Polynomial kernel function

		n						n			
		1	2	3	4			1	2	3	4
C	-1	0.630	0.555	0.500	0.500	C	-1	0.849	0.871	0.500	0.500
	0	0.630	0.555	0.500	0.500		0	0.849	0.871	0.500	0.500
	5	0.627	0.568	0.585	0.530		5	0.814	0.787	0.728	0.500
	10	0.627	0.568	0.585	0.530		10	0.814	0.787	0.728	0.500

Dot kernel function

C	-1	0.631	C	-1	0.865
	0	0.631		0	0.865
	5	0.631		5	0.781
	10	0.631		10	0.781

Radial kernel function

		γ						γ			
		0	1	2	4			0	1	2	4
C	-1	0.500	0.520	0.508	0.503	C	-1	0.500	0.505	0.502	0.501
	0	0.500	0.520	0.508	0.503		0	0.500	0.505	0.502	0.501
	5	0.500	0.524	0.504	0.501		5	0.500	0.506	0.502	0.501
	10	0.500	0.521	0.502	0.507		10	0.500	0.506	0.502	0.501

The table reports AUC values of parameter combinations for different feature vectors and sampling approaches. The parameter setup has been carried out on a random, unused dataset via stratified 5-fold cross-validation (mean AUC values are reported). Darker grey areas indicate better detection performance. n: polynomial degree; C: penalty value; γ: flexibility parameter for radial kernel function.

Table 34 – Appendix D: SVM parameter optimization for feature vector enhancements

Cost Ratio	Sample	Error costs of model (in million USD)	Costs of surveillance strategy (in million USD)	Costs of naïve strategy (in million USD)	Costs of model errors relative to surveillance strategy	Costs of model errors relative to naïve strategy
1:3	f1	82.40	3,476.90	118.40	0.024	0.696
	f2	100.06	3,399.50	148.80	0.029	0.672
	f3	114.92	3,201.81	177.60	0.036	0.647
	f4	76.98	3,033.92	142.40	0.025	0.541
	f5	80.24	2,792.30	136.00	0.029	0.590
	f6	44.95	2,780.79	96.00	0.016	0.468
	f7	27.01	2,743.66	72.00	0.010	0.375
	f8	32.95	2,704.96	49.60	0.012	0.664
	f9	31.91	2,695.02	38.40	0.012	0.831
	f10	29.82	2,837.28	35.20	0.011	0.847
	f11	29.82	2,857.67	35.20	0.010	0.847
	f12	26.03	2,836.23	36.80	0.009	0.707
	Total	677.09	35,360.03	1,086.40	-	-
	Mean	-	-	-	0.019	0.657
	Saved	-	34,682.94	409.31	-	-
1:10	f1	238.49	3,476.90	387.02	0.069	0.616
	f2	303.34	3,399.50	486.39	0.089	0.624
	f3	343.61	3,201.81	580.53	0.107	0.592
	f4	211.29	3,033.92	465.47	0.070	0.454
	f5	229.07	2,792.30	444.55	0.082	0.515
	f6	103.03	2,780.79	313.80	0.037	0.328
	f7	66.94	2,743.66	235.35	0.024	0.284
	f8	94.66	2,704.96	162.13	0.035	0.584
	f9	80.02	2,695.02	125.52	0.030	0.638
	f10	91.53	2,837.28	115.06	0.032	0.795
	f11	84.73	2,857.67	115.06	0.030	0.736
	f12	73.22	2,836.23	120.29	0.026	0.609
	Total	1,919.93	35,360.03	3,551.17	-	-
	Mean	-	-	-	0.053	0.565
	Saved	-	33,440.10	1,631.24	-	-

The table reports absolute and relative costs. The costs are based on the classification outcome of the holdout subsets of the samples f1-f12, weighted according to the empirically estimated costs from section 5.2.4 ($0.52 million for per wrongfully classified non-fraudulent observation, $5.23 million per wrongfully classified fraudulent case for the cost ratio of 1:10 and $1.6 million for the cost ratio 1:3). The applied detection model consisted of the combined feature vector and an SVM classifier in accordance with the previous results. The naïve strategy classifies all observations as non-fraudulent whereas the surveillance strategy investigates every observation. The absolute cost-saving (saved) potential is calculated by deducting the error costs of the model from the costs of the naïve and the surveillance strategy respectively for each cost ratio (for example $677.09 million - $35,360.30 million = $34,682.94 million for the cost ratio of 1:3 and the surveillance strategy as a benchmark).

Table 35 – Appendix E: Determination of absolute cost-saving potential