# Machine Learning for Identifying Risk in Financial Statements: A Survey

ELIAS ZAVITSANOS, Institute of Informatics and Telecommunications, NCSR Demokritos, Athena, Greece
EIRINI SPYROPOULOU, Institute of Informatics and Telecommunications, National Centre for Scientific Research-Demokritos, Athena, Greece
GEORGE GIANNAKOPOULOS, Institute of Informatics and Telecommunications, National Centre for Scientific Research-Demokritos, Athena, Greece
GEORGIOS PALIOURAS, Institute of Informatics and Telecommunications, National Centre for Scientific Research-Demokritos, Athens, Greece

The work herein reviews the scientific literature on Machine Learning approaches for financial risk assessment using financial reports. We identify two prominent use cases that constitute fundamental risk factors for a company, namely misstatement detection and financial distress prediction. We further categorize the related work along four dimensions that can help highlight the peculiarities and challenges of the domain. Specifically, we group the related work based on (a) the input features used by each method, (b) the sources providing the labels of the data, (c) the evaluation approaches used to confirm the validity of the methods, and (d) the machine learning methods themselves. This categorization facilitates a technical overview of risk detection methods, revealing common patterns, methodologies, significant challenges, and opportunities for further research in the field.

CCS Concepts: • **Computing methodologies** → **Machine learning**; • **Information systems** → **Clustering and classification**; **Information extraction**; • **Applied computing** → **Decision analysis**; • **General and reference** → **Surveys and overviews;**

Additional Key Words and Phrases: Risk assessment, misstatement detection, financial distress, bankruptcy prediction, fraud detection, financial reports, financial statements, machine learning, data mining, auditing

**ACM Reference Format:**
Elias Zavitsanos, Eirini Spyropoulou, George Giannakopoulos, and Georgios Paliouras. 2025. Machine Learning for Identifying Risk in Financial Statements: A Survey. *ACM Comput. Surv.* 57, 9, Article 220 (April 2025), 37 pages. https://doi.org/10.1145/3723157

Authors' Contact Information: Elias Zavitsanos, Institute of Informatics and Telecommunications, NCSR Demokritos, Athena, Attica, Greece; e-mail: izavits@iit.demokritos.gr; Eirini Spyropoulou, Institute of Informatics and Telecommunications, National Centre for Scientific Research-Demokritos, Athena, Attica, Greece; e-mail: espyropoulou@iit.demokritos.gr; George Giannakopoulos, Institute of Informatics and Telecommunications, National Centre for Scientific Research-Demokritos, Athena, Attica, Greece; e-mail: ggianna@iit.demokritos.gr; Georgios Paliouras, Institute of Informatics and Telecommunications, National Centre for Scientific Research-Demokritos, Athens, Attica, Greece; e-mail: paliourg@iit.demokritos.gr.


ACM 0360-0300/2025/04-ART220
https://doi.org/10.1145/3723157

## 1 Introduction

Financial auditing is a long and time-consuming process that aims at ensuring that the information in a company's financial report is correct. It is an objective examination and evaluation of an organization's financial statements to ensure that the financial records are a fair and accurate representation of the transactions they claim to represent.[1] One of the crucial steps of financial auditing is risk assessment, i.e., the process of identifying risks and rating the likelihood and impact of a risk event [84]. Risk assessment appropriately informs the audit plan [98], which is generated at the beginning of the audit, aiming at ensuring that appropriate attention is devoted to essential aspects and potential problems are promptly identified. During this process, auditors examine many risk factors, which usually come from industry rules, news feeds, and their own judgment based on the contents of a financial report.

Various computational tools support or augment the human judgment of risk based on the financial report. Lately, **machine learning (ML)** methods have supported some of these tools, which, in turn, require sufficient and appropriate data. Since the financial reports of publicly traded companies are openly available, academic research in this domain has focused on use cases for which (human-assigned, ground truth) labels can also be obtained for these public reports. Good examples of such labels are known misstatements, financial distress cases, or bankruptcy events that are obtainable in widely available databases. Consequently, such historical data has been used to train ML algorithms that give a probability of a report containing a misstatement or revealing future financial distress.

In this work, we review the literature on ML methods for risk assessment using financial reports. In particular, we provide an overview of the two prominent use cases in the literature: misstatement detection and financial distress/bankruptcy prediction. Recent survey articles provide an overview of the scientific work for each of the two use cases [15, 99, 132, 188]. However, as with older surveys [17, 196, 212], the main focus is on a specific use case and under specific perspectives. The work in [99], for example, reports on efforts that deal with the identification of fraud, which is one type of misstatement, namely intentional misstatement. The recent survey of [15] extensively reviews trends over the years and challenges in various dimensions but also focuses on fraud detection methods. Moreover, the survey of bankruptcy prediction methods in [132] focuses on methods that use data from a particular country, and the work in [188] provides a basic categorization of distress prediction methods based only on the ML method used. Finally, bibliometric studies, such as the work in [66, 195], and [197], although helpful in conducting a quantitative examination of the academic output and addressing current trends in the field, they do not reflect the full spectrum of challenges and open issues faced in the domain.

In contrast to these surveys, this integrative survey contributes the following. We review the related work of misstatement detection and distress prediction under the common umbrella of risk detection, focusing on four axes that reveal the peculiarities of the financial domain and highlight the challenges that need to be addressed. Considering that the principal input data for risk assessment emerges from financial reports, we group the related work based on the features extracted from the reports, the labeling sources, the evaluation approach, and the actual ML methods used. The review provides two complementary viewpoints. The first focuses on the user, presenting the two use cases in order to connect the ML-based approaches to the characteristics of the domain that directly affect their performance. The second provides a more technical overview, revealing issues, trends, research opportunities, and challenges that still need to be addressed.

The rest of the article is organized as follows: Section 2 provides an overview of the financial reports, the primary data source for all methods. It provides basic information about reports,

[1] https://www.investopedia.com/terms/a/audit.asp

highlighting their fundamental financial statements and the leading financial indices typically used in risk assessment applications. Then, Section 3 provides the main categorization of misstatement detection approaches (Section 3.1) and distress prediction approaches (Section 3.2) along four dimensions: the input features, labeling sources, evaluation approach, and proposed ML method. Based on this categorization, Section 4 provides a *technical overview* of the reviewed methods, highlighting aspects that need attention and providing suggestions for further research. Finally, Section 5 concludes the article with the most significant results and lists future work and related directions to financial report analysis for the interested reader.

## 2 Financial Reports

A large part of the literature on ML for financial risk identification has focused on analyzing financial reports as they constitute public information. In this article, we focus on three cases of risk: the risk of a report containing a misstatement, the risk related to the company being under financial distress, and the risk of a company's bankruptcy. In this survey, we explore distress and bankruptcy prediction together because (a) they are conceptually connected, since the former event can lead to the latter, and (b) there is no clear definition of financial distress, due to the fact that different countries have different accounting procedures and legal rules.

Most of the literature uses annual reports, especially in the 10-K form[2] used in the US as input for these three risk cases [181]. Annual reports contain definitions and descriptions related to the company. They also include discussions and disclosures, using domain-specific language, that should provide a fair and credible representation of the company's financial position. The nature of the domain dictates a specific terminology which is provided in the glossary Table 1.

Moreover, financial reports should reflect risk factors and the management's point of view regarding the future of the company. Financial reports combine textual and numeric data from several consolidated statements of the company, including the following:

- Income statement: The income statement lists revenues, profits, earnings, costs, and expenses. It contains attributes that reflect the total revenue, the cost of revenue, the gross profit, income tax expenses, net incomes from operations, and more. Such indices constitute factors of annual growth. They facilitate the comparison of the company's financial status to previous years and the examination of the consistency and growth of the company's operations and financial trends over the years.
- Balance sheet: As the name suggests, the balance sheet must balance up. The main idea is that assets should equal liabilities plus equity. Thus, the balance sheet is often divided into sections that describe what a company owns and owes. It contains attributes that reflect the company's assets, the cash and cash equivalents, the investments, the short, current, and long-term debt, and more. In other words, the balance sheet reflects the health of the company. Initial comparisons between total assets and liabilities can quickly reflect the company's financial health.
- Cash flow statement: The cash flow statement shows the actual cash being paid by the company and to the company. It can be broken down into operating cash flow, investing cash flow, and financing cash flow. The operating cash flow shows how much cash the company generates by operating its business. The investing cash flow shows the cash that the company spent on property, plant, and equipment or on investing in other companies and bonds. Finally, the financing cash flow reflects cash paid for bonds, dividends, and similar items. Therefore, the cash flow statement contains attributes that reflect the changes in receivables, capital expenditures, investments, net borrowings, dividends paid, and more.

[2]https://www.sec.gov/files/form10-k.pdf

Table 1. Glossary of Terms used in Financial Reports and Statements in Alphabetical Order

| Term | Definition |
|---|---|
| Accounts receivable | Money owed to the company by clients or services performed. |
| Accruals | Provisions for liabilities for which an invoice has not been received. |
| Assets | Total resources with monetary value owned by a company. |
| Balance sheet | Overview of the company's assets, liabilities, and shareholders' equity. |
| Capital expenditure | Money spent on a major project or asset. |
| Cash flow statement | Overview of how the cash flows throughout a company. |
| Consolidated statements | The overall financial statements of any entity with multiple divisions. |
| Cost of sales | Cost of acquired and resold merchandise. |
| Current assets | Assets that are expected to be turned into cash, sold, or consumed the coming year. |
| Current liabilities | Amount to be paid within one year for salaries and other debts. |
| Depreciation | Amount of expense allocated for certain types of assets that lose their value over time. |
| Dividend | A way of distributing the profits to the shareholders. |
| Equity | The amount of one's practice's total assets they actually own. |
| Expenses | The costs associated with providing services over a period of time. |
| Financial statements | Written records that convey the financial activities of a company. |
| Fiscal year | A one-year period that companies use for financial planning and budgeting. |
| Fixed assets | Long-term assets not expected to be turned into cash or consumed the coming year. |
| Income statement | Overview of revenues, expenses, net income, and earnings at a specific range of time. |
| Intangible assets | A class of non-physical, long-lived assets. |
| Long-term liabilities | Amounts owed for debts that will not become due for at least one year. |
| Net accounts receivable | Total accounts receivable minus an estimate for uncollectibles. |
| Net income | The difference between total revenue and total expenses. |
| Net operating revenue | Revenue generated by things that are not directly related to the services offered. |
| Operating revenue | Revenue generated from the day-to-day operations. |
| Profit | The amount of money after paying operating expenses, taxes, and other expenses. |
| Revenue | Money collected or expected to be collected for providing services. |
| Tangible assets | A class of physical assets that can be touched and seen. |
| Total assets | The sum of current and fixed assets. |
| Total liabilities | The sum of current and long-term liabilities. |
| Total revenue | The sum of operating and non-operating revenue. |
| Uncollectibles | An account that cannot be collected. |

An example of a financial report is depicted in Figure 1, where a few sample pages are shown to highlight the contents of such a report. Specifically, reports may be lengthy, consisting of dozens or even hundreds of pages. They have many sections, that are called items, aiming at illustrating different aspects of the status of the company, such as future risks, the auditor's opinion, the management's opinion, and consolidated statements. In this example figure, the first page of an annual report is shown (Figure 1(a)), providing basic information about the company. Then, the table of contents is depicted (Figure 1(b)), illustrating the length of the annual report, which in this case is more than one hundred pages. Additionally, a sample of the balance sheet is presented (Figure 1(c)), as an example of a consolidated financial statement, and last, a sample of a textual narrative is shown (Figure 1(d)), reflecting the management's discussion and analysis.

All the above statements provide a snapshot of the financial activity of a company for a specific period. They provide a basis for computing rates of return and evaluating a company's capital structure. Thus, they constitute valuable information for auditors and other interested parties, such as investors and analysts.

UNITED STATES
SECURITIES AND EXCHANGE COMMISSION
Washington, D.C. 20549
FORM 10-K

(Mark One)
☒ ANNUAL REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934
For the fiscal year ended December 31, 2021
OR
☐ TRANSITION REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934
For the transition period from _________ to _________
Commission File Number: 001-34756

**Tesla, Inc.**
(Exact name of registrant as specified in its charter)

Delaware
(State or other jurisdiction of incorporation or organization)

91-2197729
(I.R.S. Employer Identification No.)

13101 Tesla Road
Austin, Texas
(Address of principal executive offices)

78725
(Zip Code)

(512) 516-8177
(Registrant's telephone number, including area code)

Securities registered pursuant to Section 12(b) of the Act:

| Title of each class | Trading Symbol(s) | Name of each exchange on which registered |
|---|---|---|
| Common stock | TSLA | The Nasdaq Global Select Market |

(a) First page



(b) Table of contents

**Consolidated Balance Sheets**
**(in millions, except per share data)**

| | December 31, 2021 | December 31, 2020 |
|---|---|---|
| **Assets** | | |
| Current assets | | |
| Cash and cash equivalents | $ 17,576 | $ 19,384 |
| Short-term marketable securities | 131 | — |
| Accounts receivable, net | 1,913 | 1,886 |
| Inventory | 5,757 | 4,101 |
| Prepaid expenses and other current assets | 1,723 | 1,346 |
| Total current assets | 27,100 | 26,717 |
| Operating lease vehicles, net | 4,511 | 3,091 |
| Solar energy systems, net | 5,765 | 5,979 |
| Property, plant and equipment, net | 18,884 | 12,747 |
| Operating lease right-of-use assets | 2,016 | 1,558 |
| Digital assets, net | 1,260 | — |
| Intangible assets, net | 257 | 313 |
| Goodwill | 200 | 207 |
| Other non-current assets | 2,138 | 1,536 |
| **Total assets** | $ 62,131 | $ 52,148 |
| **Liabilities** | | |
| Current liabilities | | |
| Accounts payable | $ 10,025 | $ 6,051 |
| Accrued liabilities and other | 5,719 | 3,855 |
| Deferred revenue | 1,447 | 1,458 |
| Customer deposits | 925 | 752 |
| Current portion of debt and finance leases | 1,589 | 2,132 |
| Total current liabilities | 19,705 | 14,248 |
| Debt and finance leases, net of current portion | 5,245 | 9,556 |
| Deferred revenue, net of current portion | 2,052 | 1,284 |
| Other long-term liabilities | 3,546 | 3,330 |
| **Total liabilities** | 30,548 | 28,418 |
| Commitments and contingencies (Note 15) | | |
| Redeemable noncontrolling interests in subsidiaries | 568 | 604 |
| Convertible senior notes (Note 11) | — | 51 |

(c) Consolidated balance sheet

**ITEM 7. MANAGEMENT'S DISCUSSION AND ANALYSIS OF FINANCIAL CONDITION AND RESULTS OF OPERATIONS**

*The following discussion and analysis should be read in conjunction with the consolidated financial statements and the related notes included elsewhere in this Annual Report on Form 10-K. For discussion related to changes in financial condition and the results of operations for fiscal year 2019-related items, refer to Part II, Item 7. Management's Discussion and Analysis of Financial Condition and Results of Operations in our Annual Report on Form 10-K for fiscal year 2020, which was filed with the Securities and Exchange Commission on February 8, 2021.*

**Overview and 2021 Highlights**

Our mission is to accelerate the world's transition to sustainable energy. We design, develop, manufacture, lease and sell high-performance fully electric vehicles, solar energy generation systems and energy storage products. We also offer maintenance, installation, operation, financial and other services related to our products. Additionally, we are increasingly focused on products and services based on artificial intelligence, robotics and automation.

In 2021, we produced 930,422 vehicles and delivered 936,222 vehicles. We are currently focused on increasing vehicle production and capacity, improving and developing battery technologies, improving our FSD capabilities, increasing the affordability and efficiency of our vehicles and expanding our global infrastructure.

In 2021, we deployed 3.99 GWh of energy storage products and 345 megawatts of solar energy systems. We are currently focused on ramping production of energy storage products, improving our Solar Roof installation capability and efficiency, and increasing market share of retrofit and new build solar energy systems.

In 2021, we recognized total revenues of $53.82 billion, representing a 71% increase compared to the prior year. We continue to ramp production, build new manufacturing capacity and expand our operations to enable increased deliveries and deployments of our products and further revenue growth.

In 2021, our net income attributable to common stockholders was $5.52 billion, representing a favorable change of $4.80 billion, compared to the prior year. We continue to focus on improving our profitability through production and operational efficiencies.

We ended 2021 with $17.58 billion in cash and cash equivalents, representing a decrease of $1.81 billion from the end of 2020. Our cash flows provided by operating activities during 2021 was $11.50 billion, representing an increase of $5.55 billion compared to $5.94 billion during 2020, and capital expenditures amounted to $6.48 billion during 2021, compared to $3.16 billion during 2020. Sustained growth has allowed our business to generally fund itself, but we will continue investing in a number of capital-intensive projects in upcoming periods.

**Management Opportunities, Challenges and Risks and 2022 Outlook**

*Impact of COVID-19 Pandemic*

Beginning in the first quarter of 2021, there has been a trend in many parts of the world of increasing availability and administration of vaccines against COVID-19, as well as an easing of restrictions on social, business, travel and government activities and functions. On the other hand, infection rates and regulations continue to fluctuate in various regions and there are ongoing global impacts resulting from the pandemic, including challenges and increases in costs for logistics and supply chains, such as increased port congestion, intermittent supplier delays and a shortfall of semiconductor supply. We have also previously been affected by temporary manufacturing closures, employment and compensation adjustments, and impediments to administrative activities supporting our product deliveries and deployments.

Ultimately, we cannot predict the duration of the COVID-19 pandemic. We will continue to monitor macroeconomic conditions to remain flexible and to optimize and evolve our business as appropriate, and we will have to accurately project demand and infrastructure requirements globally and deploy our production, workforce and other resources accordingly.

31

(d) MD&A section

Fig. 1. Example of a financial annual report. This is a sample collage of pages showing (a) the first page, (b) the table of contents, (c) a balance sheet, and (d) a discussion in free text.

Additional indicator variables can be defined based on the raw financial indices that appear in the financial statements. Variables and ratios that reflect earnings, accruals, cash flows for financing activities, cash changes, and cash expenditure variations are commonly used in the literature for financial risk assessment.

Such quantitative features that are based on financial indices or more complex variables are used as variables in ML models for risk assessment. Research reveals that certain types of firms are more prone to misstatement risks, such as those with more significant debt and covenant restrictions [35, 71]. Such firms are more likely to have a higher amount of debt used to finance their assets, and manipulation of earnings is more prevalent in cases where significant debt covenant restraints exist. There is also evidence that earnings management, and particularly components of executive compensation directly tied to company performance, are connected to increased risk of misstating financial results [175]. Manipulation of earnings is generally performed through day-to-day inflation of company accounts and leads to artificially inflating company revenues, which are reflected in the financial statements [166]. Finally, other firm attributes reflected in the financial statements, such as the use of leasing [56] and high expectations for market performance [190], are also associated with high financial risk.

Similarly, numerical variables mentioned above that are directly extracted from the annual reports and calculated ratios of those have also been found to be good predictors for financial failure and bankruptcy. Very early work in the field [22] analyzed thirty financial ratios based on cash flow, net income, debt, total assets, liquid assets, and turnover. According to that work, the cash-flow to total debt ratio had the strongest ability to predict failure. Also strong predictors were the net-income to total assets ratio, total debt to total assets, working capital to total assets, current ratio, and no-credit interval. Starting with such ratios, the research at [10] identified 22 financial variables, including working capital to total assets, retained earnings to total assets, earnings before interest and taxes to total assets, market value equity to book value of total debt, and sales to total assets. The two studies above are the most referenced, leading to subsequent work on distress prediction with more recent approaches [207].

Although much of the assessment of company health and risks has been performed using quantitative data from financial reports, studies have shown that reports' narratives and disclosures provide value-relevant information regarding a company's future [189]. Concerning how annual reports are written and how they disclose information, there are clear indications that the language deployed by management engaged in falsifying firm performance is discernibly different from risk-free reports [166]. Additionally, substantial evidence indicates how our choice of words can reveal our inner intentions [103, 167, 237]. The **Management Discussion and Analysis** (**MD&A**) section of the annual reports is the one that has gained the most attention and has been used in the majority of analyses. There are efforts that: (a) study the tone of the language used in this section and reveal specific market reaction to positive or negative disclosures [80]; (b) try to identify the relationship between the readability of annual reports and financial performance [152]; (c) build on these latter studies trying to confirm that readability and language have incremental power in predicting misstatement risks [129] and bankruptcy outcomes [154]. Although the MD&A section has a monopoly in risk prediction using text data, there are other sections of the 10-K reports that are likely to contain narratives that deal with risk and uncertainty as reported in [166], such as Item 1 Business, and Item 7A Quantitative and Qualitative Disclosures about Risk.

## 3 Machine Learning for Risk Identification

This section analyzes two prominent use cases of ML for identifying risk in financial reports. We focus on the task of misstatement detection and the task of distress/bankruptcy prediction in Sections 3.1 and 3.2, respectively.

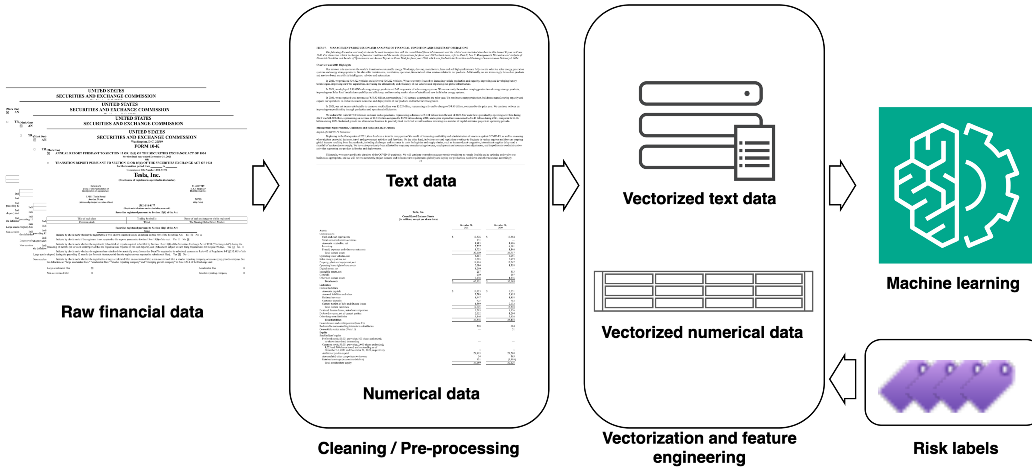


Fig. 2. The generic approach to risk identification. Starting with raw data, a pre-processing step is required to extract textual or numerical information or both. Then, a feature engineering step is needed to select features or construct new ones from the raw financial indices or the text data. This step may also include information from external sources, such as market data. The vectorized output of this step is used to feed the machine learning algorithms.

In both case studies, as already mentioned, the majority of the related work relies on features that are extracted from the financial reports of the companies. The process of preparing the data and applying ML for risk identification is depicted in Figure 2. The first step involves the collection of historical financial reports. Starting from such raw data, a cleaning or pre-processing step is needed to extract the necessary information. Depending on the type of input each method considers, specific text parts of the report or numerical values from the financial statements are extracted. Then, a feature selection or engineering step is required to select or construct features from the text data, the numerical data, or both. When numerical data are used, much work relies on external data providers that have already performed this step and offer numerical variables that comprise several financial ratios and market data. Finally, the vectorized data that reflect, ideally, the financial status of each company, along with the corresponding labels about the type of risk, can be used to train machine learning algorithms.

### 3.1 Detecting Misstatements in Financial Reports

A misstatement is an anomaly in the financial report that needs correction or requires a document amendment. It can imply disclosure issues, readability issues, or even more severe cases that need attention. Misstatements can be intentional or unintentional.

Most of the work on ML for misstatement detection treats the problem as a binary classification, i.e., misstatement versus no misstatement. However, what constitutes a misstatement differs across the literature. Much work deals with the identification of intentional misstatements, typically referred to as fraud detection. The most common examples of financial statement fraud are management fraud, expenses manipulation, assets overstatement, and assets understatement. However, some approaches deal with unintentional misstatements [133], and some do not distinguish between the two types of misstatement [26, 56]. Finally, some efforts attempt to distinguish intentional and unintentional misstatements [104, 128, 175, 182].

Much literature focuses on comparing existing approaches, such as in [128]. A comparative study of predictive models is also presented in [70] for the problem of misstatement versus non-misstatement classification. The authors point out that excluding unintentional misstatements from the data samples and focusing only on intentional misstatements may lead to incomplete predictive models. This is because unintentional misstatements are more common and should not be excluded. Furthermore, they point out that the evolving nature of financial fraud has introduced a sophisticated level of hiding fraudulent activities through actions that appear unintentional.

Besides the type of misstatement, misstatement detection approaches are differentiated mainly by their focus on different input features. In particular, whether they use quantitative features, derived from financial analysis or textual features. Moreover, different approaches also differ according to the source used to label the training examples, the type of evaluation adopted, and the machine learning method used. The following paragraphs briefly describe different approaches along these four dimensions, explaining related terminology.

**Input features:** There are two main commonly used types of input feature: financial and textual. Financial features may include indices and variables found in the financial statements, as well as ratios calculated from the variables of the financial statements. Those features usually come from proprietary databases like COMPUSTAT[3] which uses professional tools for extracting information from financial statements. Alternatively, they are extracted directly from the financial statements using custom tools. On the other hand, textual features consider the tokens that appear in the text or other linguistic information.

The features that are commonly used to train the classifiers are financial [11, 19, 42, 58, 97, 102, 106, 118, 125, 130, 131, 148, 163, 173, 179, 185] with the work in [56] and [41] being the most influential in the literature. Both of these articles provide a comprehensive analysis of financial indices, variables, and ratios. They also provide insights into the usefulness of different features in detecting misstatements that have inspired many misstatement detection approaches. Other features may concern operational characteristics, such as the turnover of top management or past auditors [1, 81], which are usually combined with financial features. CEO characteristics and auditor information have been used in recent work to enhance the feature representation of financial ratios with such traits that could help in misstatement detection [89, 117, 176, 177, 225].

There are some approaches to misstatement detection that go beyond financial features. Most work in this category has focused on the text of the MD&A section of financial reports. This section provides qualitative information and reflects the management's opinion and analysis of the reported financial results. Based on this information, the work in [186] selects the best 200 words from a corpus of MD&A sections to use as features for classification. Those words are assumed to be good predictors for misstatements, while in [113], the emphasis is on the use of "pleasant words", which are assumed to be used more commonly in fraudulent disclosures. Typically, those linguistic features come from pre-determined lists of words that have been associated with intentional misstatements. Additional features that have been used in the literature include readability measures, such as the average length of words and sentences, lexical diversity, and sentence complexity [151]. Similarly, the authors in [91, 93] use word lists that assign sentiment polarity (positive, negative, and neutral) to words, while others rely on lists of words expressing anger, anxiety, and other negative aspects, as provided by the Linguistic Inquiry and Word Count dictionary [135, 167]. On the other hand, the work in [82] does not rely on predefined word predictors. It uses word tokens with their corresponding TF-IDF values to train classifiers and predict misstatements. Last, in a recent approach [28], the authors use BERT to encode the first and the last part of the management discussion section of the annual report separately and create an ensemble of predictions from the two BERT models.

[3] https://www.marketplace.spglobal.com/en/datasets/compustat-fundamentals-(8)

Finally, a few cases combine financial and textual features, such as the work in [100], which has shown that it is possible to improve the predictive performance of the learned model by combining the two types of features. In the era of deep learning with deep neural networks, a more recent approach [52] has combined the two types of features by concatenating the neural encoding of the documents with the financial features before the final classification layer. Similarly, the work in [224] has combined both financial and textual information to train recurrent neural networks, such as LSTMs, GRUs, as well as **convolutional neural networks** (**CNNs**).

**Labeling sources:** Various databases have been used to label the training examples in the datasets throughout the related literature. In datasets containing reports of US companies, the most common labeling databases used are the US **Government Accountability Office**[4] (**GAO**), the US **Securities and Exchange Commission** (**SEC**) **Accounting and Auditing Enforcement Releases**[5] (**AAER**), and the **Audit Analytics**[6] (**AA**) database.

GAO is a compilation of restatement announcements that cover misstatements associated mainly with financial misrepresentation. AAER is a designation assigned by the SEC to administrative proceedings or litigation releases and covers intentional misstatements. Finally, AA tracks financial restatements in public filings from EDGAR,[7] the SEC's Electronic Data Gathering, Analysis, and Retrieval system. A comparison of these databases can be found in [122].

AAER labels have been mainly used in cases where the focus is on identifying intentional misstatements [1, 20, 41, 52, 56, 81, 82, 97, 100, 146, 181]. On the other hand, GAO is less widely used [128]. AA is an interesting source, containing a comprehensive set of various types of restatement cases, with more information than the other databases. However, it is also seldomly used [70]. In cases where the data sample comes from countries other than the US, alternative labeling sources have been used, such as the Athens Stock Exchange and taxation databases in [130, 131, 179], the Chinese Stock Exchange in [173, 224, 225], the Istanbul Stock Exchange in [16, 102], or other exchanges and organizations based on different regions, such as the work in [118] that focuses on Iran, the work in [117] that studies firms in Taiwan, the work in [101] that examines Bosnian firms, and the work in [105] that focuses on Poland.

A horizontal observation that emerges naturally is that using different labeling mechanisms and different datasets in the literature makes it difficult to compare existing methods and draw meaningful conclusions.

**Evaluation framework:** One of the main characteristics of misstatement detection as a classification task is the rarity of misstatements compared to the total number of financial reports. Although there is such a severe class imbalance, it is often assumed that the *a-priori* probability of a misstatement is equal to that of an accurate financial report. Much work in the literature adopts this assumption and uses balanced datasets for evaluation [58, 81, 89, 90, 97, 100, 101, 105, 113, 120, 125, 130, 149, 163, 173, 176, 177, 179, 185, 204].

A common and straightforward approach to creating such balanced datasets is to select as many negative cases (i.e., legitimate financial reports) as the positive examples available (i.e., misstatements), usually coming from the same year and industry. The difficulty in finding positive examples may sometimes lead to particularly small datasets comprising only a few dozen of examples for each class [97, 101, 120, 125, 130, 176, 177, 204]. In more extreme cases, this procedure leads to tiny datasets containing less than 50 samples [105, 163, 179]. If such models were deployed

[4] https://www.gao.gov/
[5] https://sites.google.com/usc.edu/aaerdataset/buy-the-data?authuser=0
[6] https://www.auditanalytics.com
[7] https://www.sec.goc/edgar.shtml

in the real world, where the class distribution is heavily imbalanced, their performance might differ significantly from that reported based on a balanced test set.

Admittedly, some efforts take a more realistic approach and consider the class imbalance of the domain to some degree, usually with a positive-to-negative ratio between 1:3 and 1:6 [19, 52, 102, 106, 117, 118, 128, 131, 142, 146, 203].

Independent of whether and how class imbalance is handled, many approaches use unweighted accuracy as an evaluation measure, treating all classes equally and producing misleading conclusions. As an exception to this rule, Craja et al. [52] also report results using AUC (Area Under the ROC Curve), sensitivity, and specificity, focusing on the positive misstatement class that is of greatest interest. Methods that consider an even stronger class imbalance [1, 11, 20, 41, 56, 70, 93, 181, 186, 224, 225], utilize several measures beyond accuracy, such as AUC [1, 70, 181, 224, 225], **true negative rate** (**TNR**) [70], Precision, and Recall [1, 11, 41, 102, 186].

An additional problem in how misstatement classification is evaluated is using random splits and cross-validation, ignoring the dimension of time in the data. Such an approach is very different from the real-world setting, where predictions need to happen on future financial reports based on past data - a strong time-related requirement.

Recent work [20, 230] has started considering these issues. They suggest that misstatement detection should be evaluated as an information retrieval task where results are ranked based on the predicted probability of containing a misstatement. This is aligned with the use case where a person checks the results of the systems and selects to assess some reports for misstatement. Additionally, Bao et al. [20] have used ROC-AUC (area under the receiver operating characteristic curve), **Normalized Discounted Cumulative Gain** (**NDCG**), and Precision on a small portion of the test set to evaluate various ML methods. They have also considered the time dimension to maintain the chronological order of the data when taking training and testing splits. Furthermore, Zavitsanos et al. [230] evaluate classifiers in terms of R-precision, and at the same time, they consider both the chronological order of the data and the fact that misstatements take time to be identified in the real world.

**Machine learning methods:** Most of the novelty of related literature is in the engineering of useful features for the problem. In terms of ML methods, many traditional algorithms have been used and typically different classifiers are tested on the same data.

Kirkos et al. [130] use decision trees, neural networks, and Bayesian belief networks, with the latter achieving the best results. At the same time, the work in [131] also examines the performance of **Support Vector Machines** (**SVMs**) and **K-Nearest Neighbors** (**KNN**). SVMs have also been used in [41]. In another study [148], logistic regression, neural networks, and classification trees were considered, with logistic regression achieving higher accuracy than the others. In more recent work [19, 173], regression trees and neural networks have been shown to achieve the best results. Even more recent approaches that use gradient boosting trees and XGBoost, such as the work in [11, 224, 225], also experiment with traditional methods (SVM, KNN, logistic regression). Such models are usually considered baselines and compared to more state-of-the-art approaches, such as LSTMs, CNNS, and GRUs [117, 224].

In general, different comparative studies of various ML methods [11, 16, 89, 102, 146, 181, 225] have produced different conclusions with reduced generalizability. There are several reasons for these inconsistent results: first, the datasets that are used are different and relatively small; second, the features that are used are different; and third, there is no standard common evaluation framework that adequately captures the particular characteristics of the domain.

Regarding textual features, only a few studies apply **natural language processing** (**NLP**) techniques to represent the entire textual content of the crucial sections of the financial reports. Even in these cases, text is treated as a bag of words, turned into a feature vector for traditional machine learning methods, such as SVM in [82], and Naive Bayes in [113]. To the best of our knowledge, only a few methods [52, 224] use deep learning models for text analysis in financial statements. The authors in [52] use a hierarchical attention network to model word sequences, and they concatenate the document encoding with quantitative financial features before the final classification layer. Similar approaches are considered in [224], where recurrent neural networks are used and compared against RusBoost and simpler classification models.

In the current era of **Large Language Models** (**LLMs**), approaches based on as GPT, FinBERT, and FinGPT are increasingly applied to financial tasks, due to their ability to process complex textual financial data. LLMs can be leveraged to analyze financial disclosures, earnings calls, and company filings that may suggest misstatements [143]. FinBERT [13], for example, which is a domain-specific variant of BERT, has shown promising results in financial sentiment analysis, which can indirectly help to alert for potential irregularities in textual data [172]. On the other hand, models like FinGPT [226] integrate financial news, reports, and other media to provide context-aware analyses that can detect potential fraud or errors in financial statements [169]. More recently, the authors in [28] used LLMs, particularly BERT, to encode parts of the management's opinion text. To address the limitation of BERT regarding the length of the input sequence, the authors used two BERT models to encode the first 512 tokens of the text and the last 512 tokens of the text and combined their predictions in an ensemble. Finally, FinChain-BERT [227] demonstrated a successful application of BERT models in detecting financial fraud and effectively handling complex financial text information.

### 3.2 Financial Distress/Bankruptcy Prediction

Financial distress is a situation where a company is unable to continue generating cash flows from its business, maintaining its profitability, or meeting its maturing obligations as they fall due [112]. When a firm faces such a situation, a bankruptcy event becomes probable. Consequently, prediction of financial distress and bankruptcy are essential components of risk assessment, providing early signals of financial risks for the company and the interested parties.

We explore distress prediction and bankruptcy prediction together for two reasons. First, they are conceptually connected. The first event can lead to the other. Second, there is no clear definition of financial distress because different countries have different accounting procedures and legal rules [210]. Thus, no unified rule applies globally and explicitly states whether a company is in financial distress. For these reasons, the most common labels used for training examples in this problem are either distress predictions of auditors or bankruptcy labels coming from various labeling sources [161]. Consequently, in this survey, we find studies that either use labels of distress for their predictors or use bankruptcy labels to predict events of financial default, depending on the labeling source and their perspective on distress [48, 112].

Similarly to the misstatement detection use case, we can categorize the related work based on the input features, the labeling sources, the evaluation framework, and the machine learning methods. In the following paragraphs, we overview the approaches along those axes. Again, we assume that financial reports are used as input data, containing valuable information about the tasks of distress prediction and bankruptcy prediction.

**Input features:** Regarding input features, again, we find two main types that are commonly used: financial indices, ratios, and textual features.

As far as financial features are concerned, much work relies on the pioneering work of Altman that describes financial ratios for corporate bankruptcy prediction [10] and on the work of Beaver [22] that has identified significant differences in several financial variables for bankrupt and non-bankrupt firms.

Much of the subsequent literature aimed at expanding the set of financial variables that could be good predictors for distress or bankruptcy prediction [4, 7, 25, 32, 44, 47, 49, 50, 62, 63, 77, 79, 112, 116, 124, 139, 140, 147, 157, 160, 170, 178, 180, 187, 202, 210, 216, 228, 233], with the most widely used ratios being the following: net income to total assets; working capital to total assets; returning earnings to total assets; sales to total assets; total debt to total assets; current assets to total assets; net income to net worth; total or current liabilities to total assets; cash to total assets; cash flow operations to total assets or total debt; current assets or inventory to sales; and operating income to total assets [24].

Another part of the literature aims at improving the prediction models by combining financial variables with other sources of information about the company, such as (a) market features [38, 43, 61, 64, 144, 183, 213], (b) corporate governance, like the board structure, ownership structure, leadership personnel, and others [9, 78, 121, 144], (c) characteristics of the country that the study takes place [64, 114], such as that country-level data on the economic and business environment and energy efficiency policies.

Additionally, it has been observed that the language used in the textual sections of financial statements (and in other documents or media) to describe business operations and reports is correlated with financial distress [12, 54, 138, 211]. Hence, some methods use text features from financial reports, treating text again as a bag of words, ignoring word sequence, and counting word frequency. For example, Laughran et al. [136] count the frequency of the word "ethic" and phrases like "corporate responsibility" in annual business reports to determine if the company is related to "sin" stocks. They find that managers of firms that discuss these topics are more likely to be involved in unethical situations, which can be an indicator of the future financial health of the company. Moreover, Bodnaruk et al. in [29] conclude that the frequency of negative words in an annual report is more likely to identify inflection points. In their regression models, the percentage of negative words in the 10-K documents helps indicate when a large company might suddenly slip into being financially constrained, and they observe that the more managers believe the company will face a more uncertain future, the more the text of the 10-K document will reflect this negative outlook. These results have been confirmed in a more recent work [86] where the authors also concluded that the more negative words in the annual reports, the higher the probability of financial distress.

Finally, there is evidence that textual disclosures may help in distress prediction when combined with financial variables [136, 151, 154]). However, only a few methods combine both types of features [2, 37, 119, 156, 161, 162, 236]. They report improvement over using financial features alone, concluding that deep learning methods, especially those that treat text as sequences of words, can effectively capture the most relevant features from the textual sections to complement the numeric data.

**Labeling sources:** A wide variety of labeling sources and datasets are used in the task of distress prediction, mainly because there is no unified definition of whether a firm faces distress issues. The work in [39] states that financial distress is characterized by insufficiency of liquidity, equity, and debt default. Similarly, in [83], distress is defined as a severe liquidity problem that cannot be resolved without a large-scale restructuring of the firm. According to [65], besides the above characteristics, distress also includes the situation of having a negative net asset value, which means that total liabilities exceed total assets. In [191], financial distress is defined as any of the following: (a) business failure, which means that a company cannot pay the debt even after liquidation; (b) legal bankruptcy, which means that a company has filed some form of application and declared bankruptcy; (c) technical bankruptcy, which means that a company cannot repay principal and interest; and (d) accounting bankruptcy, which means that the net assets are negative. Most recent studies, though, that collect data from developed countries focus on predicting bankruptcy, which is the ultimate form of financial distress.

A significant part of the literature uses data about US companies, coming from SEC, COMPUSTAT and CRSP [32, 61, 63, 154, 156, 171, 183, 213], which contain bankruptcy indicators, as well as from the Department of Treasury [36] and the **Federal Deposit Insurance Corporation** (**FDIC**) [194]. Studies of non-US companies, such as Hungarian [132], Danish [161], Norwegian [50], use worldwide data and labels from Bankscope, a database of micro-level banking information in the banking sector, including data that can be used for bank bankruptcy prediction [114].

Regarding European companies, data often comes from Bureau Van Dijk,[8] a private database that captures company information about compliance, supplier risk, corporate credit risk, and others that can be used for distress and bankruptcy prediction [46, 55, 64, 67, 68, 184, 208].

A third part of the related work is focused on Chinese and Taiwanese companies, using data and labels from the **China Stock Market and Accounting Research Database** (**CSMAR**), a research-oriented database focusing on Chinese finance and economy [88, 141, 210], or from the **Taiwan Economic Journal** (**TEJ**) database that provides data about corporate governance, credit risk and compliance along with financial distress-related labels [46, 47, 110, 112, 144].

It is worth stressing again that different rules and mechanisms apply in different countries regarding what is considered distress and bankruptcy. Various data sources and labels that, in some cases, indicate distress but do not, when applied to other countries, prevent the direct comparison of different methods.

**Evaluation framework:** As in the case of misstatement detection, a common characteristic of this use case is that the number of companies that face financial distress is much lower than the number of healthy companies. Although the imbalanced distribution of classes in the data should be taken into account, it seems that much of the existing work relies on balanced datasets, usually by performing negative sampling based on industry and asset size [210]. In such cases where we have a rebalanced dataset, the majority of work [5, 8, 25, 68, 72, 112, 132, 157, 170, 171, 184, 209, 210, 219] uses the measure of accuracy to evaluate the performance of the method.

Data rebalancing, through subsampling of the negative class (i.e., healthy companies), usually results in small sample sizes due to the small proportion of distressed companies [5, 87, 88, 112, 116, 121, 141, 144, 154, 170, 183, 210, 219]. Thus, rebalanced datasets do not reflect the real world, where the number of healthy companies is much bigger than those facing financial distress. Some efforts experiment with slightly imbalanced data, having a positive-to-negative ratio of 1:2 and 1:3 [3, 31, 32, 53, 57], which is still far from the real-world scenario. Another interesting case is the work in [32] and [53], which uses an imbalanced dataset in favor of the positive examples, in the sense that the set of the distressed samples is three times larger than the non-distressed ones. This is an analogy that is even further from the real-world situation.

Moving to methods that consider a more realistic class imbalance of the data, the work in [7] uses a dataset where the positive examples account only for 3% of the data and provides results in terms of failing likelihood probability, having different methods to produce different rankings of firms that are likely to fail. Similarly, the work in [55] uses a dataset with severe class imbalance and presents results in terms of mean bankruptcy probability, where a decision threshold is needed to classify the companies as risky or non-risky. Moreover, the work in [208] uses a large dataset of 63 K examples with only 256 positive cases. However, that work evaluates results in terms of accuracy, which is inappropriate for highly imbalanced datasets. Finally,

[8] https://www.bvdinfo.com/en-gb/

aiming at a more meaningful measure that does not simply average over the two classes, some work [9, 44, 64, 119, 124, 156, 161, 178, 202, 233, 236] provides results in terms of ROC-AUC.

**Machine learning methods:** Going beyond initial work on statistical methods proposed in this area [10, 22], many ML algorithms have been applied to distress or bankruptcy prediction, using the financial ratios and indices related to the use case.

Starting with supervised methods, a large part of the related work has applied regression models and especially logistic regression [63, 67, 87, 114, 121, 171, 183, 184]. Early work using artificial neural networks in this domain [17, 127, 155] had already suggested a better performance than regression and discriminant analysis [209]. In that direction, the work in [31, 48, 144] uses neural networks and multilayer perceptrons for the task of distress prediction.

On a different perspective, Min and Lee in [165], as well as Shin et al. [198], present **support vector machines** (**SVMs**) as good candidates for this task, assuming careful optimization of hyper-parameters. Similarly, SVM-based approaches were also used in [31, 63, 76, 144, 154, 171, 206].

Bagging and ensembles of tree-based estimators have also been used in [68] and [63] respectively, as well as ensembles of other estimators [88, 145] and Adaboost [115, 214]. Tree-based approaches, like classification trees, have been used in [144] and [114], with the latter presenting results using **k-nearest neighbor** (**KNN**) classifiers. In the spirit of ensemble methods, the work in [40, 51, 223, 238] trains fast classifiers for bankruptcy prediction and bank failures in the Eurozone, using gradient descent and tree ensemble learning via XGBoost.

Deep neural methods have also been applied in this domain. A hybrid approach of deep belief networks with SVM was used in [134] for financial distress prediction, while Hosaka [107] interestingly converted financial data to images to use a convolutional neural network for bankruptcy prediction. Moreover, the work in [2] and [108] relies on deep neural networks, either recurrent ones or using GoogleNet. Other recent work [53, 156, 161, 236] combines both financial and textual features under a deep learning framework to predict financial distress events. They use deep neural networks to represent the textual data, consisting of disclosures from the annual reports, and concatenate the text representation with the financial features before the final classification layer.

Focusing on NLP, the work in [205] utilizes deep learning based on the BERT encoder to show the potential value of text for predicting the bankruptcy of small businesses. The authors in [168] perform text analysis following a dictionary-based sentiment approach to predict corporate bankruptcy, showing that the textual features built on the 10-K filings significantly improve the classification performance of the prediction models on all segments, including small and medium enterprises. In parallel with textual data, the work in [150] proposes an end-to-end architecture that combines financial variables with annual reports to predict bankruptcy. The financial variables are modeled as time-series inputs in a recursive neural network, while the annual reports are handled separately by an NLP module. The module consists of an extractive text summarizer, the output of which is given to a BERT-based encoder to produce a dense representation of the text. The output document embedding is concatenated with the output of the recursive neural network and fed to the final classification layer.

More recent foundation models have also be used in the task of bankruptcy prediction. LLMs can analyze a combination of financial ratios, market data, and qualitative information (like management commentary or market sentiment) to extract relevant features for predictive tasks. Models such as BloombergGPT [222] and FinGPT [226] have been used for tasks like credit scoring and company viability assessments, leveraging domain-specific pre-training on financial corpora to enhance accuracy [143, 169]. Finally, LLMs have also been used for data augmentation in financial tasks. The authors in [199] have used the Llamma-3 model to generate realistic 10-K text sections for bankruptcy cases, by utilizing indicative bankruptcy phrases as seed text in the input prompt, while also asking the model to mimic the MD&A and the Auditor's Opinion narratives. Moreover, they used the LLM in a zero-shot classification setting, where they asked the model to directly predict the bankruptcy label based on the text of the auditor's opinion found in the 10-K report. The authors noticed that the model may be biased toward bankruptcy test cases existing in the data used for pre-training. To address this bias, they proposed a Named Entity Recognition step that filters out all the people and company names in the texts before using them as input to the LLM. They observed a drop in performance, indicating the possible challenges of using LLMs in retrospective evaluation scenarios and the need to handle look-ahead bias.

Unsupervised methods, like clustering [7, 31, 55] and **self-organizing maps** (**SOM**) [3, 7, 208], have also been used to address the lack of labeled data. The work in [3], in particular, uses a combination of genetic programming and SOM, while the work in [57] relies on clustering for identifying less representative samples (outliers) and then trains classification trees on the remaining data. Descriptive and unsupervised methods, though, do not aim at predicting a target variable, but they mainly focus on the intrinsic structure, relations, and interconnections. They aim at filtering or grouping the data to reduce overall cost and processing time and help identify outliers. In such cases, high-dimensional data are usually projected onto a 2-dimensional space to be more easily analyzed and interpreted by domain experts [193]. Finally, a few efforts that are based on isolation forests and auto-encoders have also been presented [157, 239].

Similar to the case of misstatement detection, comparative studies of methods for distress or bankruptcy prediction [50, 69, 78, 116, 124, 192, 202, 236] suffer from the problem that the methodologies are not directly comparable due to the use of different datasets, different features, and different evaluation approaches.

## 4 Technical Appraisal of Related Literature

This section provides an overview of the approaches presented in Section 3 under a technical perspective. Based on the categorization of the approaches presented in the previous section, we review both misstatement detection and distress prediction together since they share common characteristics and challenges, aiming at highlighting the significant conclusions that can drive future research. In particular, we examine the following aspects of each study: dataset size and sources, features used, evaluation approach, data labels, and data characteristics.

**Dataset size and sources:** There is a wide variety concerning the number of samples used between datasets, ranging from a few dozen to thousands of samples. Much work uses datasets with 100-500 samples, and more than half of the related work uses datasets with less than 1,500 samples, as shown in Figure 3. Especially in cases where the dataset size is of a few dozen or a few hundred samples (i.e., 100-500), it is not easy to draw robust conclusions that reflect the generalization potential of the proposed methods.

An additional challenge is the need for standardized and publicly available datasets. This is a significant issue due to proprietary restrictions, privacy concerns, and the sensitivity of financial data. Many datasets are owned by financial institutions or regulatory bodies, and access is limited to protect corporate and individual confidentiality. To fill this gap, the authors in [153] released a corpus of financial reports for US companies with textual disclosures and pre-trained word embeddings. This data can be combined with raw indices and numbers extracted from financial statements and made available by the authors of [231, 232]. Moreover, the work in [14] uses the financial reports provided in [153] and enhances them with financial ratios from COMPUSTAT and bankruptcy labels to further facilitate and accelerate related research, while in the recent work of [199], this dataset was further enriched to include the auditors' opinions, providing an additional source of audit-related information.

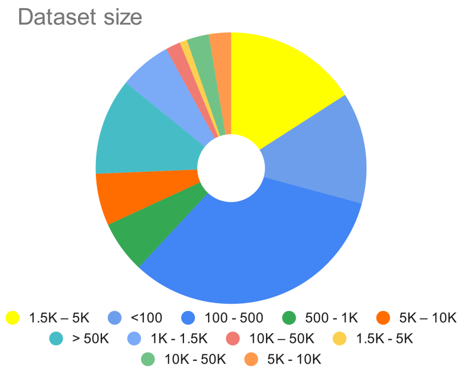


Fig. 3. Distribution of dataset size used in the reviewed articles. More than half of the articles use less than 1500 samples.

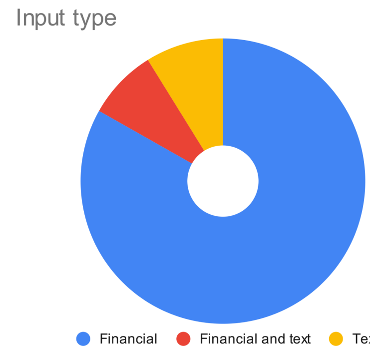


Fig. 4. Distribution of types of features used in the reviewed articles. The majority uses quantitative financial features.

Financial reports constitute the primary source of information, since they provide free text that reflects the financial status of the company, future risks, and the manager's opinion. They also provide numerical indices in the form of financial statements. However, much work and preprocessing is required to extract the financial features from those statements, involving heuristics to locate the required information or feature engineering to construct the necessary ratios and variables. Besides that, there are parts of the financial reports that are underused. Most of the related work uses the manager's narrative when considering textual information. However, there is evidence that other parts can also be helpful, such as the first section that presents the risks of the company, as well as the section containing the auditor's opinion. For example, the recent work [30] identifies links between issues reflected in the auditor's opinion and future bankruptcy events.

**Features used:** Irrespective of the data size, the majority of the work uses quantitative financial features to train the proposed algorithms for risk detection (Figure 4). In particular, much work relies on the identification of the most appropriate features [20, 41, 45, 56, 77, 79, 109, 173, 238] that serve as good predictors for each use case, which are then used to train classifiers by ML methods. Such features are usually a combination of different "raw" indices found in the financial statements that reflect risk aspects. However, recent work provides evidence that even individual raw financial features alone can be good predictors for the same use cases [20]. Table 2 shows the financial features that are more frequently used in the financial literature, including raw indices, ratios, and market variables. One caveat with such features is that they require either much manual work to be extracted or the development of custom tools to identify related tables in financial statements and extract information from them. Moreover, these features typically require the combination of financial figures from various external sources to the reports. This is why most related work relies on proprietary databases to produce the training datasets.

Besides financial features, some methods also incorporate the text found in disclosures of the annual reports. They either use text features alone, or in combination with financial features. Two factors that contributed to the increased use of text are its availability in the reports and the intuition that the language used may relate to financial risk [12, 29, 54, 86, 113, 136, 138, 151, 211]. Related methods focus on sentiment analysis, using predefined lists of negative and positive words, linguistic features, and external lexica. On this basis, it is surprising that natural language processing with deep learning methods has not been widely used yet, with the exception of a few recent approaches [2, 52, 156, 157, 161, 224, 236].

Table 2. Financial Features that are Typically Used in the Literature

| | | |
|---|---|---|
| Current assets, total | Account payable | Assets, total |
| Common/ordinary equity, total | Cash & short investments | Cost of goods sold |
| Common shares outstanding | Debt in current liabilities, total | Inventories, total |
| Long-term debt issuance | Long-term debt, total | Net income |
| Property & equipment, total | Depreciation & amortization | Receivables, total |
| Income bf extraordinary items | Investment and advances, other | Retained earnings |
| Short-term inv/ments, total | Current liabilities, total | Liabilities, total |
| Preferred stock, total | Sales/turnover | Sale of common/preferred stock |
| Income taxes payable | Income taxes, total | Interest and expense, total |
| Price close, annual, fiscal | WC accruals | RSST accruals |
| Change in receivables | Change in inventory | Change in cash sales |
| Soft assets percentage | Change in cash margin | Change in return on assets |
| actual issuance | Book to market | Depreciation index |
| Retained earnings over assets | Earnings before interest | Cash Flow Earnings Difference |
| Change in employees | Change in order backlog | Return on pension plan assets |
| Leverage | Market-adjusted stock return | Earnings to price |
| Net income/total assets | Current ratio | Quick ratio |
| Sales/total assets | Working capital/total assets | Cash/total assets |
| Total debt/total assets | Current assets/total assets | Net income/net worth |
| Earnings bf int. & tax/total assets | Inventory/sales | Net income/sales |
| Total liabilities/total assets | Current assets/sales | Total debt/net worth |
| Market equity value/Total debt value | Net worth/total assets | Capital/assets |
| Cash flow operations/total assets | Operating income/total assets | Sales/inventory |

In addition, in the era of deep learning, besides deep neural network architectures that focus on text processing and creation of corresponding dense representations and embeddings, there are similar architectures that handle categorical and numerical features [95, 96, 111], which remain unexplored in the financial domain and particularly in the context of risk assessment.

One obstacle in adopting deep learning methods may be the requirement for interpretable classification decisions. Specifically, interpretability is the degree to which a human can understand the cause of the decision of the model [164], or in other words, the degree to which a human can consistently predict the model's result [126]. The higher the interpretability of a machine learning model, the easier it is to comprehend why specific predictions have been made. Interpretability in this domain is critically important for justifying decisions supported by an automated system for risk assessment. Most approaches that rely on data mining techniques, especially using numeric data, seem to lack interpretation, explanation, and understanding from the practitioners' viewpoint [18, 207]. Methods that rely on traditional classification techniques, such as decision trees, are, in general, explainable in the sense that the classification model can be visualized, and one can verify the decisions. However, in more complex algorithms, the justification of the system's decisions needs more attention. As an ML model's performance increases, the model's explainability generally decreases [21], and when dealing with tabular data, the techniques that are mainly applied to address this issue rely on **Local Interpretable Model-agnostic Explanations** (**LIME**) and **Shapley Additive exPlanations** (**SHAP**) [234]. This is also a factor that encouraged the use of textual data and text features in this domain, as such features can help the user justify their opinion given a statement or annual report, e.g., by identifying "red-flag" sentences in the MD&A sections of the reports [52]. Figure 5 shows an example of a document containing a misstatement. On the left (Figure 5(a)), the numerical features that contribute to the decision of the model are identified using SHAP values, while on the right (Figure 5(b)), the "red-flag" sentences

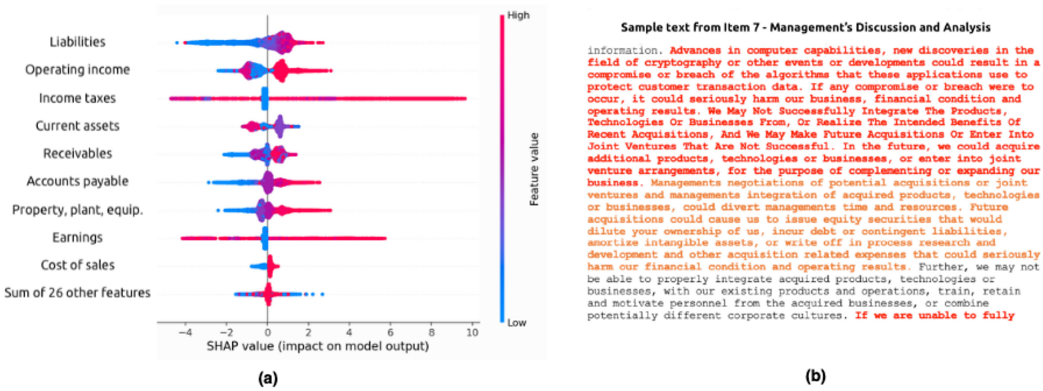


Fig. 5. Interpretation of an instance with material misstatement. (a) The SHAP values are used to identify the importance of features, when variables derived from the financial statements are used as predictors. (b) Parts of the narrative section have been identified as "red-flags" using the weights of the trained deep neural network, then the MD&A section is used as input. Images inspired from [234] and [52].

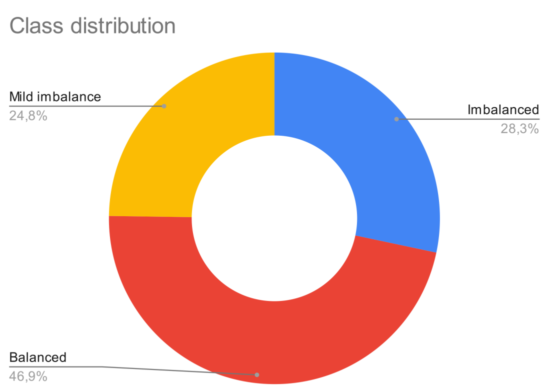


Fig. 6. Distribution of classes in the datasets. Almost 50% of the literature uses balanced datasets, while less than 30% of the related work considers datasets with large class imbalance. Mild imbalance refers to negative-to-positive ratios below 10:1.

of the MD&A section have been highlighted using the attention weights of a hierarchical deep neural network.

**Evaluation approach:** As mentioned in Section 3, the typical case is to have many examples of regular or risk-free financial reports and only a few reports that contain a risk of misstatement or reflect a risk of financial distress. This class imbalance must be taken carefully into account when estimating the performance of the trained models. However, much work uses either balanced or mildly imbalanced datasets to train and test the algorithms (Figure 6).

Additionally, even when class imbalance is considered, standard measures such as model accuracy are widely used for evaluation (Figure 7), measuring the percentage of the correctly classified instances. This measure seems to dominate, especially in cases that use small datasets containing a few dozen or up to a few hundred instances (Figure 8). In cases where there is a severe class imbalance between the classes, this measure can lead to misleading conclusions because it reflects the ability of the model to correctly predict the dominating negative class, which is of far lower interest than the positive one. Thus, reporting per class precision/recall or sensitivity/specificity

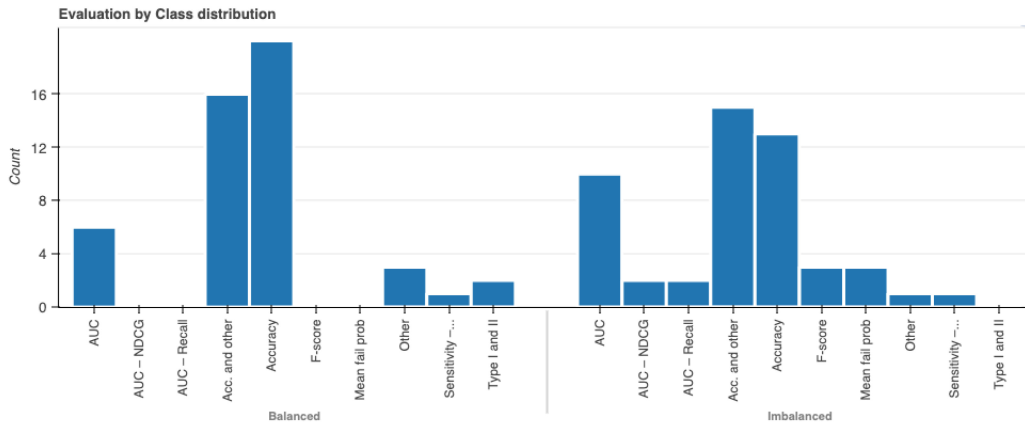


Fig. 7. Distribution of evaluation methods used in balanced or imbalanced datasets. Accuracy or accuracy plus additional measures, such as precision and recall (grouped under the category “Acc. and other”) prevail in both cases. The next most commonly used measure is the area under the ROC curve (AUC).

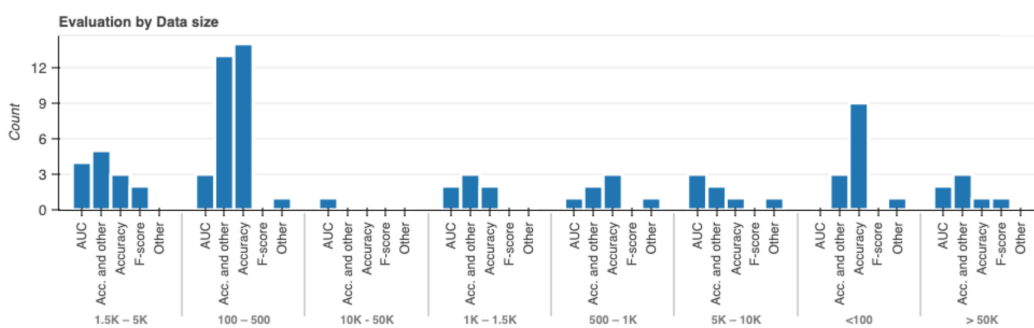


Fig. 8. Distribution of evaluation methods by dataset size. Accuracy prevails noticeably when small datasets are used. In larger datasets, additional measures are also used, either alone or reported together with accuracy.

better estimates the classifier’s performance for this task, especially since the majority of the related work adopts a supervised classification approach (Figure 9). Evidently, some previous work has followed this approach, as shown in Figure 10. In addition, recent efforts have evaluated the learned classifiers as rankers, based on the probability of the positive class and using information retrieval measures such as AUC, NDCG, Precision at k and R-Precision [20, 230]. This way of evaluating the real applicability and usefulness of the models is appropriate when a user (domain expert) can make the best use of their available time by going through a ranked list of predictions.

However, using the area under the ROC curve (AUC) as an evaluation measure in highly imbalanced data may give an overly optimistic assessment of the model’s performance. For instance, the authors in [20] report AUC between 62% and 72.5%, while at the same time the Precision ranges from 0.85% to 4.48%. This bias arises because the ROC curve’s **false positive rate** (**FPR**) can become very small when the number of actual negative examples is large. As a result, even a large number of false positives would only lead to a small FPR, leading to a potentially high AUC that does not reflect the practical reality of using the model. Therefore, in realistic evaluation settings that consider the high class imbalance observed in this domain, measures such as the area under the precision - recall curve or other ranking measures should be preferred.

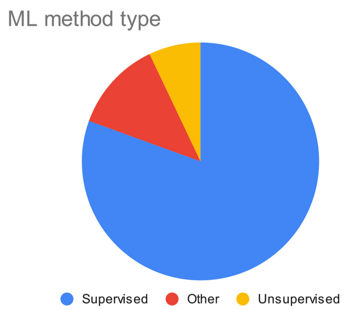


Fig. 9. Distribution of ML method type used in the reviewed articles. Most efforts use supervised learning methods.

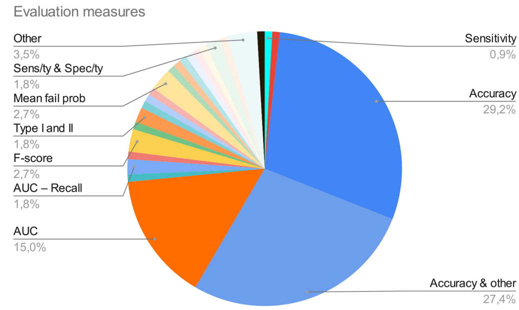


Fig. 10. Distribution of evaluation measures. A large part of the work uses Accuracy.

On the other hand, some approaches, usually proposed by domain experts, introduce new specialized evaluation measures, e.g., in terms of the impact of risk on CEO/CFO turnover and economic costs, as shown in Table 3 of Appendix A. The work in [104], for instance, searches for positive correlations between CFO and CEO turnover rates for firms with misstatements, while the work in [86] explores how financial distress relates to dependent variables, such as dividend payments, percent loan loss provisions and **return on assets** (**ROA**).

Another important aspect of evaluating predictive models related to risk is the assumption on *decision cost*. Most existing work assumes that the cost of misclassifying a true positive is the same as that of misclassifying a true negative. In other words, the cost of misclassifying a financial statement as not risky when it is, is the same as the cost of misclassifying a financial statement as risky when it is not. However, this is not true in the real world, making the use of cost-sensitive evaluation and learning approaches particularly interesting.

A further important parameter that needs to be taken into account in the evaluation process is the dimension of *time*. Many methods reviewed so far use random splits and cross-validation for training and testing, ignoring the chronological order of data. This is very different from the real-world setting, where predictions need to happen on future financial reports. As a result, the performance estimates using cross-validation are likely to be too optimistic when the trained models are deployed in practice. When the chronological order of data is maintained, one needs to also take into account that the predictive power of features (especially financial, which reflect a snapshot of the financial status of a company at a specific period) may deteriorate when used to make predictions for test sets with a large temporal distance from the training set [230]. Hence, if one is after a good estimate of the performance of the model in a real-world situation, one needs to carefully select the training and test sets, taking into account that the financial environment changes.

The *evolution* of the financial environment means that the characteristics of the domain in terms of financial features and the language used in the reports are also evolving and changing. For instance, the authors in [59] noticed that there are significant changes between recent financial documents and documents written two decades ago, in terms of the vocabulary and the style of the language used. The work in [91] suggests that the quantitative financial numbers alone may not change significantly when a company is behaving abnormally, but the writing style that is used to communicate information changes. This position is not accepted by everyone though. A study presented in [33], for instance, concludes that identifying abnormal behavior is a hard task on its own, given the wide use of standardized boilerplate language for the reports.

In summary, we need to start building methods that will be able to evolve together with the financial environment, that will exploit rich representations of data of the domain, and that will use transfer learning techniques to adapt quickly to new challenges of the domain.

Domain-specific language models have been shown to perform well in the context of deep learning, such as BERT [60] models, as reported in [13, 34, 59]. Specifically, in that work, the BERT model was trained in the financial domain, using annual reports from the SEC, and was evaluated on related tasks, such as financial sentiment analysis and indexing. In these cases, the models were able to capture the domain-specific language and outperformed models trained on general texts. However, all these observations require further validation in more financial applications.

Other efforts use more recent financial variants of LLMs, such as FinGPT, BloombergGPT, and Fin-chain BERT [222, 226, 227] to facilitate textual analysis in financial applications. LLMs in this domain may offer great value in handling financial risk, by providing textual and contextual understanding. They may interpret nuances in language that other models may miss. They can combine textual and numerical data to offer a more holistic analysis, and they can be adapted or fine-tuned to financial datasets to specialize in this domain.

On the other hand, methods that rely on LLMs should consider that both misstatements and bankruptcy are rare events, requiring additional techniques such as oversampling or careful generation of synthetic data that mimic the minority class. Additionally, although LLMs can generate plausible explanations, they remain a black box, and the explainability of the context they generate is limited. Therefore, their use may pose potential challenges in terms of regulation and governance. Currently, LLMs play an auxiliary role, aiding in text analysis or sentiment analysis, which can then be fed into existing models that handle financial variables or text data. This indicates that LLMs still have the potential to evolve and move beyond their current role as powerful tools for augmenting existing methods [235].

**Data labels:** A common problem in training and evaluating predictive models is the *subjectivity of annotations/labels* in the datasets. In both use cases of risk assessment studied in this article, subjectivity may arise due to varying personal views or different legal and financial settings. In particular, regarding distress prediction, as noted in Section 3.2, there is no commonly accepted definition of distress across countries, and thus, different criteria may be used in different countries. A company in one country may be considered to face financial distress, while a company with similar characteristics and financial status in another country may be considered to be healthy. Additionally, since financial distress is a dynamic situation that goes through various phases, from mild to serious distress resulting in default events, it is hard to identify a single point in time at which a company starts being in distress. In the majority of cases, financial distress is defined strictly as a situation that clearly indicates financial difficulty, such as bankruptcy, or entering a situation of **special treatment (ST)** by a stock exchange. Such a strict definition facilitates research sampling and data selection but excludes the majority of cases that domain experts would consider distress situations.

Another important characteristic of the data, which mainly applies to the case of misstatement detection, is that the training set of negative examples may also contain positive ones (i.e., misstatements). When the dataset is created, it is assumed that all the cases that have not been identified as misstatements are negative examples. However, this is not always the case. The negative will often include examples that are positive but have not been identified as positive (i.e., misstatements) at the time. This is a very important aspect that is not widely considered. The work in [20] deals with this issue to some degree, introducing a two-year gap between training and testing datasets. The aim is at excluding the training years that may have the largest number of unidentified misstatements. The work in [230] improves on this aspect using the restatement date to label misstatements correctly at each training period. A more general treatment of this problem would be to use methods that learn from positive and unlabeled examples (PU Learning) [23].

An additional issue, encountered specifically in the work on misstatement detection, concerns the handling of serial misstatements, especially when considering text features as input. As observed in [230] and [28], in some firms, misstatements span over multiple consecutive years, which can be due to accounting errors that propagate through the years or even because the same professional body is responsible for filing the financial reports. In such cases, it is important to investigate if the trained model is actually learning the misstatement behavior, or merely learning the writing patterns of serial misstatements and essentially overfits on these serial cases. It is important to handle such situations by including a qualitative evaluation, or by performing different training and test splits of the data to exclude serial positive cases that appear multiple times in the training set.

Throughout the review, we observed that different labeling sources are often used, even to label the same annual reports. This happens especially in the case of misstatement detection, where different labeling sources provide different kinds of misstatements [122, 230]. Depending on the labelling source, the prediction task is made either easier or harder. For instance, it is easier to detect intentional misstatements than unintentional ones, which may contain accounting errors that are hard to detect in financial reports. On the other hand, in the case of distress prediction, we usually see datasets that either use distress labels as defined by different countries, or even use bankruptcy labels to detect financial distress. Overall it seems that there is a lack of good benchmark datasets that could facilitate model development, evaluation, and comparison, and as Table 3 shows, the datasets differ greatly across the various studies in the literature. This is also due to the fact that very few annotated datasets are available openly because they require custom extraction tools to gather financial variables, while the labeling of data requires much manual effort that is usually performed by specific financial data providers. Additionally, the financial domain is a sensitive area in terms of data privacy, and stakeholders, investors, company owners, and researchers in the field are reluctant to share information and data that may lead to the identification of an individual. However, we need to invest in efforts that will produce publicly available labeled datasets to support the use of ML methods and accelerate progress in the field.

The various challenges highlighted in this section in developing predictive models for financial misstatement and distress detection using ML remain largely unaddressed by current literature. Significant effort has been invested in describing and addressing the above challenges, but methods that consider all the characteristics and peculiarities of the financial domain are still missing.

### 4.1 Misstatement Detection Versus Bankruptcy Prediction

Throughout this survey, we viewed misstatement detection and bankruptcy prediction as the two most prominent use cases of risk assessment, mainly due to the amount of work they involve. From the technical appraisal of the related literature, it is evident that there are key overlaps between these two cases in terms of features used (e.g., financial ratios and textual parts from financial reports) and challenges (e.g., class imbalance, evaluation frameworks, available data). However, these two tasks remain quite distinct and their differences arise from the nature of the problems, the data used, and the goals of the predictions.

By its definition, financial misstatement detection aims at identifying whether a company's financial statements have been intentionally or unintentionally misstated. This translates to detecting fraudulent or erroneous reporting practices. On the other hand, the goal of bankruptcy prediction is to predict whether a company is likely to declare bankruptcy or face severe financial distress in the future, which involves assessing the company's financial health and, therefore, may require a more complex analysis based on historical data for each company.

By its nature, misstatement detection is often framed as a binary classification task (i.e., "misstatement" vs. "not-misstatement") or multi-class (i.e., "fraudulent", "not-fraudulent", or "error"). It focuses on anomalies or patterns that deviate from expected norms and typically involves understanding intentional manipulation or accounting errors, which may include accounting practices or irregularities that may lead to material misstatements. On the other hand, bankruptcy prediction can be framed either as a classification or a regression problem. It focuses on predicting the likelihood of bankruptcy within a specific time horizon (e.g., one year, five years) and it is based on long-term trends in financial health and risk of insolvency. It is, therefore, more suited to time-series-based approaches or methods that examine historical data per company. In some cases, it is more straightforward than misstatement detection, as it looks at objective metrics of financial decline rather than deceptive intent.

Both methods rely mainly on financial reports and financial statements as data sources. However, the task of misstatement detection can also benefit from other data sources, such as restatement announcements, accounting ratios, and management statements. In contrast, bankruptcy prediction may benefit more by leveraging financial ratios, macroeconomic indicators, market data, and auditor opinions. Therefore, misstatement detection often relies on features related to red flags, such as accounting ratios, sudden changes in revenue or expenses, audit outcomes, or anomalies in cash flow patterns, and behavioral indicators, such as management changes or audit firm switches. On the other hand, bankruptcy prediction relies on features that capture financial decline over time, such as debt-to-equity ratio, interest coverage ratio, working capital trends, operating cash flows, and market conditions. In both cases, the data challenges are similar. Both misstatements and bankruptcy are rare events. However, the data for misstatement detection may also contain label noise that should be considered at the time of training or subtle manipulations that require feature engineering or domain-specific expertise. On the other hand, the patterns leading to bankruptcy may be more evident than those in misstatements.

Regarding evaluation measures, in both tasks, handling false positives and false negatives separately is critical, due to class imbalance. Although standard measures such as accuracy, precision, recall, and F1 score are often used, ranking measures or precision-recall curves that leverage probabilistic predictions need to be incorporated.

Finally, misstatement detection and bankruptcy prediction are of relevance to different stakeholders. Semi-automated misstatement detection can help regulatory bodies (e.g., SEC) for auditing and enforcement or internal audit departments for risk mitigation. Semi-automated bankruptcy prediction may be helpful for credit risk assessment by banks and lenders, investment decisions by stakeholders, or for supporting early-warning systems for corporate management.

To enable all the above, availability and access to publicly available datasets is required. However, as already mentioned, there is a limited amount of data publicly available. Commercial datasets like Bloomberg, COMPUSTAT, or CRSP are behind paywalls, making them inaccessible to many researchers. Both financial misstatements and bankruptcies are relatively rare and perhaps not adequately recorded. Additionally, historical data might not capture changes in accounting regulations or financial practices. Moreover, sharing detailed financial data may breach confidentiality agreements or legal restrictions, limiting dataset availability. Besides the efforts that are already being made toward compiling large financial datasets [14, 153, 199, 231, 232], there are alternatives and approaches for researchers to practice their methods. Platforms like Kaggle,[9] UCI Machine Learning Repository,[10] or SEC filings[11] are helpful sources for initial studies. Small datasets from small or medium markets are also a good starting point [220]. In addition, researchers can

[9]https://www.kaggle.com/datasets
[10]https://archive.ics.uci.edu/datasets
[11]https://www.sec.gov/data-research

create synthetic datasets by simulating financial ratios and firm-level data. Techniques such as Generative Adversarial Networks [94] can help produce realistic synthetic data while preserving privacy. Moreover, researchers can participate in challenges or hackathons hosted by financial institutions, which sometimes provide limited access to real-world datasets. Using existing financial LLMs like FinBERT or FinGPT, which have been pre-trained on publicly available financial text data, may also be helpful in the initial steps, since such models may provide contextual insights even without direct access to proprietary datasets. These alternatives enable researchers to begin working in the domain, while navigating the challenges of limited access to proprietary datasets.

## 5 Conclusion

This survey article presented a review of the scientific literature of ML approaches for risk identification in financial reports. We identified two main use cases that heavily relate to risk assessment; misstatement detection and financial distress prediction. Both cases affect at the same time the company under consideration, the auditor, and other interested parties.

The review provided a categorization of the approaches along four principal dimensions: the input features, the labeling sources, the evaluation method, and the ML methods used. This categorization aimed at acquainting the reader with the main aspects that affect the performance of the ML methods, as well as to identify the main characteristics of the domain and the challenging problems that should be considered.

The aforementioned categorization helped also to group together the various gaps and challenges that appear in both use cases, in order to study their technical aspects and suggest possible future directions. The main takeaways are that although there is much work in the field with ML methods, there are important challenges that require attention. Future work should consider the following:

— the lack of standardized datasets in the field and the difficulties of compiling and distributing open source data;
— the difficulties in choosing between different labeling sources and their implications on the task per se;
— the need to model the problem under a semi-supervised (PU) learning setting to deal with unreliable labels and annotations;
— the need to account for the chronological order of data and to evaluate the derived models realistically, considering at the same time the severe class imbalance in the data;
— the need to use appropriate evaluation measures that focus on the correct class and truly reflect the performance of the learned models and they are not influenced by the large number of negative examples or the severe class imbalance;
— the need to combine various types of features to improve the predictive performance of the models and to incorporate text data;
— the need to use other parts of the financial reports that have not been sufficiently used and may constitute good predictors for risk assessment;
— the need to use modern ML methods and approaches that can evolve together with the changes of the financial domain and have the ability to model both textual data and numeric or tabular data with modern neural representations;
— the need to incorporate explainability techniques to enhance the involvement of the users, to explore unintended consequences, and to facilitate fairness of AI.

One should also be aware of numerous efforts that underline the value of financial report analysis and can constitute interesting sources of research benchmarks and indicative results in the future. Such efforts include workshops, such as the "Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation" [73]. The corresponding research community works on a number of identified challenges, through shared tasks: (i) financial narrative summarization [74, 75, 229]; (ii) financial document structure extraction [27]; (iii) financial document causality detection [159]. Since the related domain has been already overviewed by targeted survey articles, we would also like to point the interested reader directly to such publications [123, 158, 218] that can possibly help in the task of risk assessment by providing summarized and structured information.

All these endeavors, together with the investment of main stakeholders in AI research, form the very challenging and very promising ecosystem of financial document analysis. This survey aspires to constitute a starting point for researchers who want to delve deeper into this ecosystem of opportunities, providing a stepping stone toward research on ML for financial document intelligence.

## Acknowledgments

We would like to acknowledge the financial support of Qualco SA. The opinions of the authors expressed herein do not necessarily state or reflect those of Qualco SA.

## References

[1] Ahmed Abbasi, Conan Albrecht, Anthony Vance, and James Hansen. 2012. Metafraud: A meta-learning framework for detecting financial fraud. *Mis Quarterly* 36, 4 (2012), 1293–1327.

[2] Asad Abdi, Siti Mariyam Shamsuddin, Shafaatunnur Hasan, and Jalil Piran. 2019. Deep learning-based sentiment classification of evaluative text based on Multi-feature fusion. *Information Processing & Management* 56, 4 (2019), 1245–1259. DOI: https://doi.org/10.1016/j.ipm.2019.02.018

[3] H. A Abdou. 2009. Genetic programming for credit scoring: The case of Egyptian public sector banks. *Expert Systems with Applications* 36, 9 (2009), 11402–11417.

[4] Mohammad Abdullah. 2021. The implication of machine learning for financial solvency prediction: An empirical analysis on public listed companies of Bangladesh. *Journal of Asian Business and Economic Studies* 28, 4 (2024/04/17 2021), 303–320. DOI: https://doi.org/10.1108/JABES-11-2020-0128

[5] Debi Acharjya and Rathi Raj. 2021. An extensive study of statistical, rough, and hybridized rough computing in bankruptcy prediction. *Multimedia Tools and Applications* 80, 28 (2021), 35387–35413. DOI: https://doi.org/10.1007/s11042-020-10167-2

[6] Nurul Alam, Junbin Gao, and Stewart Jones. 2021. Corporate failure prediction: An evaluation of deep learning vs discrete hazard models. *Journal of International Financial Markets, Institutions and Money* 75 (2021), 101455. DOI: https://doi.org/10.1016/j.intfin.2021.101455

[7] P. Alam, D. Booth, K. Lee, and T. Thordarson. 2000. The use of fuzzy clustering algorithm and self-organizing neural networks for identifying potentially failing banks: An experimental study. *Expert Systems with Applications* 18, 3 (2000), 185–199.

[8] H. Aljawazneh, A. M. Mora, P. García-Sánchez, and P. A. Castillo-Valdivieso. 2021. Comparing the performance of deep learning methods to predict companies' financial failure. *IEEE Access* 9 (2021), 97010–97038. DOI: https://doi.org/10.1109/ACCESS.2021.3093461

[9] Nawaf Almaskati, Ron Bird, Danny Yeung, and Yue Lu. 2021. A horse race of models and estimation methods for predicting bankruptcy. *Advances in Accounting* 52 (2021), 100513. DOI: https://doi.org/10.1016/j.adiac.2021.100513

[10] Edward I. Altman. 1968. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance* 23, 4 (1968), 589–609.

[11] B. An and Y. Suh. 2020. Identifying financial statement fraud with decision rules obtained from modified random Forest. *Data Technologies and Applications* 54, 2 (2020), 235–255. DOI: https://doi.org/10.1108/DTA-11-2019-0208

[12] W. Antweiler and M. Frank. 2004. Is all that talk just noise? The information content of internet stock message boards. *Journal of Finance* 59, 3 (2004), 1259–1294.

[13] Dogu Araci. 2019. FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. *ArXiv abs/1908.10063* (2019): n. pag.

[14] Henri Arno, Klaas Mulier, Joke Baeck, and Thomas Demeester. 2023. From numbers to words: Multi-modal bankruptcy prediction using the ECL dataset. In *Proceedings of the 6th Workshop on Financial Technology and Natural Language Processing*. Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, Hsin-Hsi Chen, Hiroki Sakaji, and Kiyoshi Izumi (Eds.), Association for Computational Linguistics, Bali, Indonesia, 11–21. DOI: https://doi.org/10.18653/v1/2023.finnlp-2.2

[15] Matin N. Ashtiani and Bijan Raahemi. 2021. Intelligent fraud detection in financial statements using machine learning and data mining: A systematic literature review. *IEEE Access* 0, 0 (2021), 1–1. DOI: https://doi.org/10.1109/ACCESS.2021.3096799. Early access.

[16] H. A. Ata and I. H. Seyrek. 2009. The use of data mining techniques in detecting fraudulent financial statements: An application on manufacturing firms. *The Journal of Faculty Economics and Administrative Science* 14, 2 (2009), 157–170.
[17] A. F. Atiya. 2001. Bankruptcy prediction for credit risk using neural networks: A survey and new results. *IEEE Transactions on Neural Networks* 12, 4 (2001), 929–935.
[18] B. Baesens, T. V. Gestel, S. Viaene, M. Stepanova, J. Suykens, and J. Vanthienen. 2003. Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society* 54, 6 (2003), 627–635.
[19] B. Bai, J. Yen, and X. Yang. 2008. False financial statements: Characteristics of China's listed companies and CART detecting approach. *International Journal of Information Technology and Decision Making* 7, 2 (2008), 339–359.
[20] Yang Bao, Bin Ke, Bin Li, Y. Julia Yu, and Jie Zhang. 2020. Detecting accounting fraud in publicly traded US firms using a machine learning approach. *Journal of Accounting Research* 58, 1 (2020), 199–235.
[21] G. Baryannis, S. Dani, and G. Antoniou. 2019. Predicting supply chain risks using machine learning: The tradeoff between performance and interpretability. *Future Generation Computer Systems* 101 (2019), 993–1004.
[22] William H. Beaver. 1966. Financial ratios as predictors of failure. *Journal of Accounting Research* 4 (1966), 71–111. Retrieved from http://www.jstor.org/stable/2490171
[23] J. Bekker and J. Davis. 2020. Learning from positive and unlabeled data: A survey. *Machine Learning* 109, 4 (2020), 719–760.
[24] J. L. Bellovary, D. E. Giacomino, and M. D. Akers. 2007. A review of bankruptcy prediction studies: 1930 to present. *The Journal of Financial Education* 33 (2007), 1–42.
[25] Sami Ben Jabeur, Nicolae Stef, and Pedro Carmona. 2022. Bankruptcy prediction using the XGBoost algorithm and variable importance feature engineering. *Computational Economics* 61, 2 (2022), 715–741. DOI: https://doi.org/10.1007/s10614-021-10227-1
[26] M.D. Beneish. 1999. The detection of earnings manipulation. *Financial Analysis Journal* 55, 5 (1999), 24–36.
[27] Najah-Imane Bentabet, Rémi Juge, Ismail El Maarouf, Virginie Mouilleron, Dialekti Valsamou-Stanislawski, and Mahmoud El-Haj. 2020. The financial document structure extraction shared task (FinToc 2020). In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation.* COLING, Barcelona, Spain (Online), 13–22. Retrieved from https://www.aclweb.org/anthology/2020.fnp-1.2
[28] Indranil Bhattacharya and Ana Mickovic. 2024. Accounting fraud detection using contextual language learning. *International Journal of Accounting Information Systems* 53 (2024), 100682. DOI: https://doi.org/10.1016/j.accinf.2024.100682
[29] A. Bodnaruk, T. Loughran, and B. McDonald. 2015. Using 10-K text to gauge financial constraints. *Journal of Financial and Quantitative Analysis* 50, 5 (2015), 623–646.
[30] Konstantinos Bougiatiotis, Elias Zavitsanos, and Georgios Paliouras. 2023. Identifying going concern issues in auditor opinions: Link to bankruptcy events. In *Proceedings of the 2023 IEEE International Conference on Big Data.* IEEE, IEEE, Sorrento, Italy, 2805–2813.
[31] M. A. Boyacioglu, Y. Kara, and O. K. Baykan. 2009. Predicting bank financial failures using neural networks, support vector machines and multivariate statistical methods: A comparative analysis in the sample of savings deposit insurance fund (SDIF) transferred banks in Turkey. *Expert Systems with Applications* 36, 2 (2009), 3355–3366.
[32] Xavier Bredart, Vanessa Vella, and Joseph Bonello. 2018. Machine learning models for predicting financial distress. *Journal of Research in Economics* 2, 2 (2018), 174–185. DOI: https://doi.org/10.24954/jore.2018.22
[33] S. Brown and J. Tucker. 2011. Large sample evidence on firms year-over-year MD&A modifications. *Journal of Accounting Research* 49, 2 (2011), 309–346.
[34] Vinicio Desola, Kevin Hanna, and Pri Nonis. 2019. Finbert: Pre-trained model on sec filings for financial natural language tasks. DOI: 10.13140/RG.2.2.19153.89442
[35] N. Burns and S. Kedia. 2006. The impact of performance-based compensation on misreporting. *Journal of Financial Economics* 79, 1 (2006), 35–67.
[36] Raffaella Calabrese, Marta Degl'Innocenti, and Silvia Angela Osmetti. 2017. The effectiveness of TARP-CPP on the US banking industry: A new copula-based approach. *European Journal of Operational Research* 256, 3 (2017), 1029–1037. DOI: https://doi.org/10.1016/j.ejor.2016.07.046
[37] J. L. Campbell, H. Chen, and D. S. Dhaliwal. 2014. The information content of mandatory risk factor disclosures in corporate filings. *Review of Accounting Studies* 19, 1 (2014), 396–455. DOI: https://doi.org/10.1007/s11142-013-9258-3
[38] J. Y. Campbell, J. Hilscher, and J. Szilagyi. 2008. In search of distress risk. *The Journal of Finance* 63, 6 (2008), 2899–2939.
[39] D. R. Carminchael. 1972. *The Auditor's Reporting Obligation.* AICPA, NY, USA.
[40] P. Carmona, F. Climent, and A. Momparler. 2019. Predicting failure in the U.S banking sector: An extreme gradient boosting approach. *International Review of Economics and Finance* 61, C (2019), 304–323.
[41] M. Cecchini, H. Aytug, G. J. Koehler, and P. Pathak. 2010. Detecting management fraud in public companies. *Management Science* 56, 7 (2010), 1146–1160.

[42] W. Chai, B. K. Hoogs, and B. T. Verschueren. 2006. Fuzzy ranking of financial statements for fraud detection. In *Proceeding of the IEEE International Conference on Fuzzy System*. IEEE, Vancouver, BC, 152–158.
[43] D. Karthik Chandra, V. Ravi, and I. Bose. 2009. Failure prediction of dotcom companies using hybrid intelligent techniques. *Expert Systems with Applications* 36, 3, Part 1 (2009), 4830–4837. DOI: https://doi.org/10.1016/j.eswa.2008.05.047
[44] Chris Charalambous, Spiros Martzoukos, and Zenon Taoushianis. 2022. Estimating corporate bankruptcy forecasting models by maximizing discriminatory power. *Review of Quantitative Finance and Accounting* 58, 1 (2022), 297–328. DOI: https://doi.org/10.1007/s11156-021-00995-0
[45] A. Chaudhuri and K. De. 2011. Fuzzy support vector machine for bankruptcy prediction. *Applied Soft Computing* 11, 2 (2011), 2472–2486.
[46] Mu-Yen Chen. 2011. Predicting corporate financial distress based on integration of decision tree classification and logistic regression. *Expert Systems with Applications* 38, 9 (2011), 11261–11272. DOI: https://doi.org/10.1016/j.eswa.2011.02.173
[47] S. Chen and Z. D. Shen. 2020. Financial distress prediction using hybrid machine learning techniques. *Asian Journal of Economics, Business and Accounting* 16, 2 (2020), 1–12.
[48] W.-S. Chen and Y.-K. Du. 2009. Using neural networks and data mining techniques for the financial distress prediction model. *Expert Systems with Applications* 36 (2009), 4075–4086.
[49] S. Cho, H. Hong, and B.-C. Ha. 2010. A hybrid approach based on the combination of variable selection using decision trees and case-based reasoning using the Mahalanobis distance for bankruptcy prediction. *Expert Systems and Applications* 37, 4 (2010), 3482–3488.
[50] Claudiu Clement. 2020. Machine learning in bankruptcy prediction. *Journal of Public Administration, Finance and Law* 17, 17 (2020), 20.
[51] F. Climent, A. Momparler, and P. Carmona. 2019. Anticipating bank distress in the Eurozone: An extreme gradient boosting approach. *Journal of Business Research* 101, C (2019), 885–896.
[52] P. Craja, A. Kim, and S. Lessmann. 2020. Deep learning for detecting financial statement fraud. *Decision Support Systems* 139 (2020), 113421.
[53] Fang Da, Chao Peng, Haomin Wang, and Tie Li. 2022. A risk detection framework of Chinese high-tech firms using wide & deep learning model based on text disclosure. *Procedia Computer Science* 199 (2022), 262–268. DOI;https://doi.org/10.1016/j.procs.2022.01.032
[54] S. R. Das and M. Y. Chen. 2007. Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management Science* 53, 9 (2007), 1375–1388.
[55] Javier De Andrés, Pedro Lorca, Francisco Javier de Cos Juez, and Fernando Sánchez-Lasheras. 2011. Bankruptcy forecasting: A hybrid approach using Fuzzy c-means clustering and Multivariate Adaptive Regression Splines (MARS). *Expert Systems with Applications* 38, 3 (2011), 1866–1875. DOI: https://doi.org/10.1016/j.eswa.2010.07.117
[56] P. M. Dechow, W. Ge, C. R. Larson, and R. G. Sloan. 2011. Predicting material accounting misstatements. *Contemporary Accounting Research* 28, 1 (2011), 17–82.
[57] Z. Defu, S. C. H. Leung, and Y. Zhimei. 2008. A decision tree scoring model based on genetic algorithm and k-means algorithm. In *Proceedings of the 3rd International Conference on Convergence and Hybrid Information Technology* . IEEE, Busan, Korea (South), 1043–1047.
[58] Q. Deng and G. Mei. 2009. Combining self-organizing map and K-means clustering for detecting fraudulent financial statements. In *Proceedings of the IEEE International Conference on Granular Computing*. IEEE, Nanchang, China, 126–131.
[59] Vinicio Desola, Kevin Hanna, and Pri Nonis. 2019. FinBERT: Pre-trained Model on SEC Filings for Financial Natural Language Tasks. DOI: https://doi.org/10.13140/RG.2.2.19153.89442
[60] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL-HLT* 1, 2019 (2018), 4171–4186.
[61] A. Adam Ding, Shaonan Tian, Yan Yu, and Hui Guo. 2012. A class of discrete transformation survival models with application to default probability prediction. *J. Amer. Statist. Assoc.* 107, 499 (2012), 990–1003. DOI: https://doi.org/10.1080/01621459.2012.682806
[62] M. Divsalar, H. Roodsaz, F. Vahdatinia, G. Norouzzadeh, and A.H. Behrooz. 2012. A robust data-mining approach to bankruptcy prediction. *Forecast* 31, 6 (2012), 504–523.
[63] I. Bose D. K. Chandra, V. Ravi. 2009. Failure prediction of dotcom companies using hybrid intelligent techniques. *Expert Systems and Applications* 36, 3 (2009), 4830–4837.
[64] Michalis Doumpos, Kostas Andriosopoulos, Emilios Galariotis, Georgia Makridou, and Constantin Zopounidis. 2017. Corporate failure prediction in the European energy sector: A multicriteria approach and the effect of country characteristics. *European Journal of Operational Research* 262, 1 (2017), 347–360. DOI: https://doi.org/10.1016/j.ejor.2017.04.024

[65] M. Doumpos and C. Zopounidis. 1998. A multicriteria discrimination method for the prediction of financial distress: The case of greece. *Multinational Finance Journal* 3, 2 (1998), 71–101.

[66] Robertas Damaševičius Dovilė Kuizinienė, Tomas Krilavičius and Rytis Maskeliūnas. 2022. Systematic review of financial distress identification using artificial intelligence methods. *Applied Artificial Intelligence* 36, 1 (2022), 2138124. DOI: https://doi.org/10.1080/08839514.2022.2138124

[67] Philippe Du Jardin. 2015. Bankruptcy prediction using terminal failure processes. *European Journal of Operational Research* 242, 1 (2015), 286–303.

[68] Philippe du Jardin. 2016. A two-stage classification technique for bankruptcy prediction. *European Journal of Operational Research* 254, 1 (2016), 236–252.

[69] D. Duarte and F. Barboza. 2020. Forecasting financial distress with machine learning – a review. *Future Studies Research Journal: Trends and Strategies* 12, 3 (2020), 528–574.

[70] I. Dutta, S. Dutta, and B. Raahemi. 2017. Detecting financial restatements using data mining techniques. *Expert Systems with Applications* 90 (2017), 374–393.

[71] J. Efendi, A. Srivastava, and E.P. Swanson. 2007. Why do corporate managers misstate financial statements? The role of option compensation and other factors. *Journal of Financial Economics* 85, 3 (2007), 667–708.

[72] M. El-Bannany, M. Sreedharan, and A.M. Khedr. 2020. A robust deep learning model for financial distress prediction. *International Journal of Advanced Computer Science and Applications* 11, 2 (2020), 170–175.

[73] Dr Mahmoud El-Haj, Dr Vasiliki Athanasakou, Dr Sira Ferradans, Dr Catherine Salzedo, Dr Ans Elhag, Dr Houda Bouamor, Dr Marina Litvak, Dr Paul Rayson, Dr George Giannakopoulos, and Nikiforos Pittaras (Eds.), 2020. *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*. COLING, Barcelona, Spain (Online).

[74] Mahmoud El-Haj, Ahmed AbuRa'ed, Marina Litvak, Nikiforos Pittaras, and George Giannakopoulos. 2020. The financial narrative summarisation shared task (FNS 2020). In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*. COLING, Barcelona, Spain (Online), 1–12. Retrieved from https://www.aclweb.org/anthology/2020.fnp-1.1

[75] M. El-Haj, N. Zmandar, P. Rayson, A. AbuRa'ed, M. Litvak, N. Pittaras, G. Giannakopoulos, A. Kosmopoulos, B. Carbajo-Coronado, and A. Moreno-Sandoval. 2022. The financial narrative summarisation shared task (FNS 2022). In *Proceedings of the 4th Financial Narrative Processing Workshop, LREC 2022*. LREC, European Language Resources Association, Marseille, France, 43–52.

[76] B. E. Erdogan. 2013. Prediction of bankruptcy using support vector machines: An application to bank bankruptcy. *Journal of Statistical Computation and Simulation* 83, 8 (2013), 1543–1555.

[77] F.-C. Li F.-L. Chen. 2010. Combination of feature selection approaches with SVM in credit scoring. *Expert Systems and Applications* 37, 7 (2010), 4902–4909.

[78] Hossam Faris, Ruba Abukhurma, Waref Almanaseer, Mohammed Saadeh, Antonio M. Mora, Pedro A. Castillo, and Ibrahim Aljarah. 2020. Improving financial bankruptcy prediction in a highly imbalanced class distribution using oversampling and ensemble learning: A case from the Spanish market. *Prog. in Artif. Intell.* 9, 1 (2020), 31–53. DOI: https://doi.org/10.1007/s13748-019-00197-9

[79] A. Feki, A. B. Ishak, and S. Feki. 2012. Feature selection using Bayesian and multiclass support vector machines approaches: application to bank risk prediction. *Expert Systems and Applications* 39, 3 (2012), 3087–3099.

[80] R. Feldman, S. Govindaraj, J. Livnat, and B. Segal. 2010. Management's tone change, post earnings announcement drift and accruals. *Review of Accounting Studies* 15, 4 (2010), 915–953.

[81] E. H. Feroz, T. M. Kwon, V. Pastena, and K. J. Park. 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 and Management* 9, 3 (2000), 145–157.

[82] Marcia Fissette and Theo de Vries. 2017. Text mining to detect indications of fraud in annual reports worldwide. In *Benelearn 2017: Proceedings of the 26th Benelux Conference on Machine Learning, Technische Universiteit Eindhoven*. S.N., Eindhoven, 69–71.

[83] G. Foster. 1986. *Financial Statement Analysis*. Prentice Hall, NJ, USA.

[84] Mark L. Frigo and Richard J. Anderson. 2009. Strategic risk assessment. A first step for improving risk management and governance. *Strategic Finance* 1, 1 (2009), 25–33.

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

[86] Priyank Gandhi, Tim Loughran, and Bill McDonald. 2018. Using annual report sentiment as a proxy for financial distress in U.S. Banks (September 25, 2018). Available at SSRN: https://ssrn.com/abstract=2905225 or http://dx.doi.org/10.2139/ssrn.2905225

[87] V. Garcia, A. I. Marques, J. S. Sanchez, and H. J. Ochoa-Dominguez. 2019. Dissimilarity-based linear models for corporate bankruptcy prediction. *Computational Economics* 53, 3 (2019), 1019–1031.

[88] Ruibin Geng, Indranil Bose, and Xi Chen. 2015. Prediction of financial distress: An empirical study of listed Chinese companies using data mining. *European Journal of Operational Research* 241, 1 (2015), 236–247. DOI: https://doi.org/10.1016/j.ejor.2014.08.016

[89] Adrian Gepp, Kuldeep Kumar, and Sukanto Bhattacharya. 2021. Lifting the numbers game: Identifying key input variables and a best-performing model to detect financial statement fraud. *Accounting & Finance* 61, 3 (2021), 4601–4638. DOI: https://doi.org/10.1111/acfi.12742. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/acfi.12742

[90] F. H. Glancy and S. B. Yadav. 2011. A computational model for financial reporting fraud detection. *Decision Support Systems* 50, 3 (2011), 595–601.

[91] S. Goel and J. Gangolly. 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 (2012), 75–89.

[92] S. Goel, J. Gangolly, S. Faerman, and O. Uzuner. 2010. Can linguistic predictors detect fraudulent financial filings? *Journal of Emerging Technologies in Accounting* 7, 1 (2010), 25–46.

[93] S. Goel and O. Uzuner. 2016. Do sentiments matter in fraud detection? Estimating semantic orientation of annual reports. *Intelligent Systems in Accounting, Finance and Management* 23, 3 (2016), 215–239.

[94] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Communications of the ACM* 63, 11 (2020), 139–144. DOI: https://doi.org/10.1145/3422622

[95] Yura Gorishniy, Ivan Rubachev, and Artem Babenko. 2022. On embeddings for numerical features in tabular deep learning. *ArXiv abs/2203.05556* (2022), n. pag.

[96] Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. 2021. Revisiting deep learning models for tabular data. In *Proceedings of the 35th International Conference on Neural Information Processing Systems (NIPS'21)*. Curran Associates Inc., Red Hook, NY, USA, Article 1447, 18932–18943.

[97] B. P. Green and J. H. Choi. 1997. Assessing the risk of management fraud through neural network technology. *Auditing* 16, 1 (1997), 14–28.

[98] K. Gupta. 2005. *Contemporary Auditing*. McGraw Hill, US.

[99] S. Gupta and K. Mehta. 2021. Data mining-based financial statement fraud detection: Systematic literature review and meta-analysis to estimate data sample mapping of fraudulent companies against non-fraudulent companies. *Global Business Review* 0, 0 (2021), 1–24.

[100] P. Hajek and R. Henriques. 2017. Mining corporate annual reports for intelligent detection of financial statement fraud – A comparative study of machine learning methods. *Knowledge-Based Systems* 128 (2017), 139–152.

[101] Sanel Halilbegovic, Nedim Čelebić, Ermin Cero, Elvisa Buljubasic, and Anida Mekic. 2020. Application of Beneish M-score model on small and medium enterprises in Federation of Bosnia and Herzegovina Application of Beneish M-score model on small and medium enterprises | 147. *Eastern Journal of European Studies* 11, 1 (2020), 146.

[102] S. Hamal and O. Senvar. 2021. Comparing performances and effectiveness of machine learning classifiers in detecting financial accounting fraud for Turkish SMEs. *International Journal of Computational Intelligence Systems* 14, 1 (2021), 769–782.

[103] J. Hancock, L. Curry, S. Goorha, and Woodworth M. 2008. On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Process* 45, 1 (2008), 1–23.

[104] K. M. Hennes, A. J. Leone, and B. P. Miller. 2008. The importance of distinguishing errors from irregularities in restatement research: The case of restatements and CEO/CFO turnover. *The Accounting Review* 83, 6 (2008), 1487–1519.

[105] A. Holda. 2020. Using the Beneish M-score model: Evidence from non-financial companies listed on the Warsaw Stock Exchange. *Investment Management and Financial Innovations* 17, 4 (2020), 389–401. DOI: https://doi.org/10.21511/imfi.17(4).2020.33

[106] B. Hoogs, T. Kiehl, C. Lacomb, and D. Senturk. 2007. A genetic algorithm approach to detecting temporal patterns indicative of financial statement fraud. *Intelligent Systems in Accounting, Finance and Management* 15, 1–2 (2007), 41–56.

[107] T. Hosaka. 2019. Bankruptcy prediction using imaged financial ratios and convolutional neural networks. *Expert Systems with Applications* 117 (2019), 287–299.

[108] Tadaaki Hosaka. 2019. Bankruptcy prediction using imaged financial ratios and convolutional neural networks. *Expert Systems with Applications* 117 (2019), 287–299. DOI: https://doi.org/10.1016/j.eswa.2018.09.039

[109] S. Huang, R. Tsaih, and F. Yu. 2014. Topological pattern discovery and feature extraction for fraudulent financial reporting. *Expert Systems with Applications* 41, 9 (2014), 4360–4372.

[110] Shi-Ming Huang, Chih-Fong Tsai, David C. Yen, and Yin-Lin Cheng. 2008. A hybrid financial analysis model for business failure prediction. *Expert Systems with Applications* 35, 3 (2008), 1034–1040. DOI: https://doi.org/10.1016/j.eswa.2007.08.040

[111] X. Huang, A. Khetan, M. Cvitkovic, and Z. Karnin. 2020. TabTransformer: Tabular DataModeling Using Contextual Embeddings. arXiv:2012.06678.

[112] Yu Pei Huang and Meng Feng Yen. 2019. A new perspective of performance comparison among machine learning algorithms for financial distress prediction. *Applied Soft Computing Journal* 83 (2019), 105663. DOI: https://doi.org/10.1016/j.asoc.2019.105663
[113] S. L. Humpherys, K. C. Moffitt, M. B. Burns, J. K. Burgoon, and W. F. Felix. 2011. Identification of fraudulent financial statements using linguistic credibility analysis. *Decision Support Systems* 50, 3 (2011), 585–594.
[114] Christos Ioannidis, Fotios Pasiouras, and Constantin Zopounidis. 2010. Assessing bank soundness with classification techniques. *Omega* 38, 5 (2010), 345–357. DOI: https://doi.org/10.1016/j.omega.2009.10.009
[115] Heo J and Yang JY. 2014. AdaBoost based bankruptcy forecasting of Korean construction companies. *Applied Soft Computing* 24 (2014), 494–499.
[116] Sami Ben Jabeur, Cheima Gharib, Salma Mefteh-Wali, and Wissal Ben Arfi. 2021. CatBoost model and artificial intelligence techniques for corporate failure prediction. *Technological Forecasting and Social Change* 166 (2021), 120658. DOI: https://doi.org/10.1016/j.techfore.2021.120658
[117] Chyan-Long Jan. 2021. Detection of financial statement fraud using deep learning for sustainable development of capital markets under information asymmetry. *Sustainability* 13, 17 (2021), 20. DOI: https://doi.org/10.3390/su13179879
[118] A. Javadian Kootanaee, A. A. Poor Aghajan, and M. Hosseini Shirvani. 2021. A hybrid model based on machine learning and genetic algorithm for detecting fraud in financial statements. *Journal of Optimization in Industrial Engineering* 14, 2 (2021), 169–186. DOI: https://doi.org/10.22094/joie.2020.1877455.1685
[119] Cuiqing Jiang, Ximei Lyu, Yufei Yuan, Zhao Wang, and Yong Ding. 2022. Mining semantic features in current reports for financial distress prediction: Empirical evidence from unlisted public firms in China. *International Journal of Forecasting* 38, 3 (2022), 1086–1099. DOI: https://doi.org/10.1016/j.ijforecast.2021.06.011
[120] Kathleen A. Kaminski, T. Sterling Wetzel, and Liming Guan. 2004. Can financial ratios detect fraudulent financial reporting? *Managerial Auditing* 19, 1 (2004), 15–28.
[121] Sunaina Kanojia and Shasta Gupta. 2023. Bankruptcy in Indian context: Perspectives from corporate governance. *Journal of Management & Governance* 27, 2 (2023), 505–545. DOI: https://doi.org/10.1007/s10997-022-09630-
[122] J. M. Karpoff, A. Koester, D. S. Lee, and G. S. Martin. 2012. A Critical Analysis of Databases used in Financial Misconduct Research (Working paper).
[123] Colm Kearney and Sha Liu. 2014. Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis* 33 (2014), 171–185.
[124] Maria Sultana Keya, Himu Akter, Md. Atiqur Rahman, Md. Mahbobur Rahman, Minhaz Uddin Emon, and Md. Sabab Zulfiker. 2021. Comparison of different machine learning algorithms for detecting bankruptcy. In *Proceedings of the 2021 6th International Conference on Inventive Computation Technologies*. IEEE, Coimbatore, India, 705–712. DOI: https://doi.org/10.1109/ICICT50816.2021.9358587
[125] T. R. Kiehl, B. K. Hoogs, and C. A. LaComb. 2005. Evolving multi-variate time series patterns for the discrimination of fraudulent financial filings. In *Proceedings of the Genetic and Evolutionary Computation Conference*. ACM, Washington, DC, USA, 8.
[126] Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. 2016. Examples are not enough, learn to criticize! Criticism for Interpretability. In *Proceedings of the Advances in Neural Information Processing Systems*. D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Eds.), Vol. 29, Curran Associates, Inc., Barcelona, Spain. Retrieved from https://proceedings.neurips.cc/paper/2016/file/5680522b8e2bb01943234bce7bf84534-Paper.pdf
[127] J. W. Kim, H. R. Weistroffer, and R. T. Redmond. 1993. Expert systems for bond rating: A comparative analysis of statistical, rule-based and neural network systems. *Expert Systems* 10, 3 (1993), 167–172.
[128] Y. Kim, B. Baik, and S. Cho. 2016. Detecting financial misstatements with fraud intention using multi-class cost-sensitive learning. *Expert Systems with Applications* 62, 4 (2016), 32–43.
[129] Kin Lo, Felipe Ramos, and Rafael Rogo. 2017. Earnings management and annual report readability. *Journal of Accounting and Economics* 63, 1 (2017), 1–25. https://doi.org/10.1016/j.jacceco.2016.09.002
[130] E. Kirkos, C. Spathis, and Y. Manolopoulos. 2007. Data mining techniques for the detection of fraudulent financial statements. *Expert Systems and Applications* 32 (2007), 995–1003.
[131] S. Kotsiantis, E. Koumanakos, D. Tzelepis, and V. Tampakas. 2006. Forecasting fraudulent financial statements using data mining. *Computational Intelligence* 3, 2 (2006), 104–110.
[132] Tamás Kristóf and Miklós Virág. 2020. A comprehensive review of corporate bankruptcy prediction in hungary. *Journal of Risk and Financial Management* 13, 2 (2020), 35. DOI: https://doi.org/10.3390/jrfm13020035
[133] B. Louise Hayes. 2014. Identifying Unintentional Error in Restatement Disclosures (January 1, 2014). Available at SSRN: https://ssrn.com/abstract=2269086 or http://dx.doi.org/10.2139/ssrn.2269086
[134] Zineb Lanbouri and Said Achchab. 2015. A hybrid deep belief network approach for financial distress prediction. In *Proceedings of the 2015 10th International Conference on Intelligent Systems: Theories and Applications*. IEEE, IEEE, Rabat, Morocco, 1–6.
[135] D. Larcker and A. Zakolyukina. 2012. Detecting deceptive discussion in conference calls. *Journal of Accounting Research* 50, 2 (2012), 495–540.

[136] T. Laughran, B. McDonald, and H. Yun. 2009. A wolf in sheep's clothing: The use of ethics-related terms in 10-K reports. *Journal of Business Ethics* 89 (2009), 39–49.

[137] M. Lenard, A. Watkins, and P. Alam. 2007. Effective use of integrated decision making: An advanced technology model for evaluating fraud in service-based computer and technology firms. *Journal of Emerging Technologies in Accounting* 4, 1 (2007), 123–137.

[138] F. Li. 2008. Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and Economics* 45, 2–3 (2008), 221–247.

[139] H. Li and J. Sun. 2011. Predicting business failure using support vector machines with straightforward wrapper: A re-sampling study. *Expert Systems and Applications* 38, 10 (2011), 12747–12756.

[140] H. Li and J. Sun. 2011. Principal component case-based reasoning ensemble for business failure prediction. *Information Management* 48, 6 (2011), 220–227.

[141] Hui Li, Jie Sun, and Bo-Liang Sun. 2009. Financial distress prediction based on OR-CBR in the principle of k-nearest neighbors. *Expert Systems with Applications* 36, 1 (2009), 643–659. DOI: https://doi.org/10.1016/j.eswa.2007.09.038

[142] X. Li, W. Xu, and X. Tian. 2014. How to protect investors? A GA-based DWD approach for financial statement fraud detection. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics* IEEE, San Diego, CA, USA, 3548–3554.

[143] Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. 2023. Large language models in finance: A survey. In *Proceedings of the Fourth ACM International Conference on AI in Finance (ICAIF'23)*. Association for Computing Machinery, New York, NY, USA, 374–382. https://doi.org/10.1145/3604237.3626869

[144] Deron Liang, Chia-Chi Lu, Chih-Fong Tsai, and Guan-An Shih. 2016. Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study. *European Journal of Operational Research* 252, 2 (2016), 561–572. DOI: https://doi.org/10.1016/j.ejor.2016.01.012

[145] D. Liang, C. Tsai, A. Dai, and W. Eberle. 2018. A novel classifier ensemble approach for financial distress prediction. *Knowledge and Information Systems* 54, 2 (2018), 437–4612.

[146] C. Lin, A. Chiu, S. Huang, and D. Yen. 2015. Detecting the financial statement fraud: The analysis of the differences between data mining techniques and experts' judgments. *Knowledge-Based Systems* 89 (2015), 459–470.

[147] F. Lin, D. Liang, and E. Chen. 2011. Financial ratio selection for business crisis prediction. *Expert Systems and Applications* 38, 12 (2011), 15094–15102.

[148] Fen-May Liou. 2008. Fraudulent financial reporting detection and business failure prediction models: A comparison. *Managerial Auditing Journal* 23, 7 (2008), 650–622.

[149] C. Liu, Y. Chan, S. Hasnain, A. Kazmi, and H. Fu. 2015. Financial fraud detection model: Based on random forest. *Journal of Economics and Finance* 7, 7 (2015), 178–188.

[150] Gianfranco Lombardo, Andrea Bertogalli, Sergio Consoli, and Diego Reforgiato Recupero. 2024. Natural language processing and deep learning for bankruptcy prediction: An end-to-end architecture. *IEEE Access* 12 (2024), 151075–151091. DOI: https://doi.org/10.1109/ACCESS.2024.3476621

[151] T. Loughran and B. McDonald. 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance* 66, 1 (2011), 35–65.

[152] T. Loughran and B. McDonald. 2014. Measuring readability in financial disclosures. *Journal of Finance* 69, 4 (2014), 1643–71.

[153] Lefteris Loukas, Manos Fergadiotis, Ion Androutsopoulos, and Prodromos Malakasiotis. 2021. EDGAR-CORPUS: Billions of tokens make the world go round. In *Proceedings of the 3rd Workshop on Economics and Natural Language Processing*. Association for Computational Linguistics, Punta Cana, Dominican Republic, 13–18. DOI: https://doi.org/10.18653/v1/2021.econlp-1.2

[154] Cecchini M, Aytug H, Koehler GJ, and Patha kP. 2010. Making words work: Using financial text as a predictor of financial events. *Decision Support Systems* 50, 1 (2010), 164–175.

[155] J. J. Maher and T. K. Sen. 1997. Predicting bond ratings using neural networks: A comparison with logistic regression. *Intelligent Systems in Accounting, Finance and Management* 6, 1 (1997), 59–72.

[156] Feng Mai, Shaonan Tian, Chihoon Lee, and Ling Ma. 2019. Deep learning models for bankruptcy prediction using textual disclosures. *European Journal of Operational Research* 274, 2 (2019), 743–758. DOI: https://doi.org/10.1016/j.ejor.2018.10.024

[157] Mohamed Wiem Mkaouer Makram Soui, Salima Smiti and Ridha Ejbali. 2020. Bankruptcy prediction using stacked auto-encoders. *Applied Artificial Intelligence* 34, 1 (2020), 80–100. DOI: https://doi.org/10.1080/08839514.2019.1691849

[158] Xiliu Man, Tong Luo, and Jianwu Lin. 2019. Financial sentiment analysis (fsa): A survey. In *Proceedings of the 2019 IEEE International Conference on Industrial Cyber Physical Systems*. IEEE, Taipei, Taiwan, 617–622.

[159] Dominique Mariko, Hanna Abi-Akl, Estelle Labidurie, Stephane Durfort, Hugues De Mazancourt, and Mahmoud El-Haj. 2020. The financial document causality detection shared task (FinCausal 2020). In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*. COLING, Barcelona, Spain (Online), 23–32.

[160] A. Martin, V. Gayathri, G. Saranya, P. Gayathri, and P. Venkatesan. 2011. A hybrid model for bankruptcy prediction using genetic algorithm, fuzzy c-means and MARS. *International Journal of Soft Computing* 2, 1 (2011), 12–23.

[161] Rastin Matin, Casper Hansen, Christian Hansen, and Pia Mølgaard. 2019. Predicting distresses using deep learning of text segments in annual reports. *Expert Systems with Applications* 132 (2019), 199–208. DOI: https://doi.org/10.1016/j.eswa.2019.04.071

[162] W. J. Mayew, M. Sethuraman, and M. Venkatachalam. 2015. MD&A disclosure and the firm's ability to continue as a going concern. *Accounting Review* 90, 4 (2015), 1621–1651.

[163] Mohammad Hussein Ranjbar Mehdi Rad, Ali Amiri and Hojatollah Salari. 2021. Predictability of financial statements fraud-risk using Benford's Law. *Cogent Economics & Finance* 9, 1 (2021), 1889756. DOI: https://doi.org/10.1080/23322039.2021.1889756

[164] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38. DOI: https://doi.org/10.1016/j.artint.2018.07.007

[165] J. H. Min and Y. C. Lee. 2005. Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert Systems with Applications* 28, 4 (2005), 603–614.

[166] S. Minhas and A. Hussain. 2016. From spin to swindle: Identifying falsification in financial text. *Cognitive Computation* 8, 4 (2016), 729–745.

[167] M. Newman, J., D. Berry, and J. Richards. 2003. Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin* 29, 5 (2003), 665–675.

[168] Ba-Hung Nguyen and Van-Nam Huynh. 2022. Textual analysis and corporate bankruptcy: A financial dictionary-based sentiment approach. *Journal of the Operational Research Society* 73, 1 (2022), 102–121. DOI: https://doi.org/10.1080/01605682.2020.178

[169] Yuqi Nie, Yaxuan Kong, Xiaowen Dong, John M. Mulvey, H. Vincent Poor, Qingsong Wen, and Stefan Zohren. 2024. A Survey of Large Language Models for Financial Applications: Progress, Prospects and Challenges. *ArXiv abs/2406.11903* (2024): n. pag.

[170] Triasesiarta Nur and Rosinta Ria Panggabean. 2020. Accuracy of financial distress model prediction: The implementation of artificial neural network, logistic regression, and discriminant analysis. In *Proceedings of the 1st Borobudur International Symposium on Humanities, Economics and Social Sciences.* Atlantis Press, Indonesia, 402–406. DOI: https://doi.org/10.2991/assehr.k.200529.084

[171] David L. Olson, Dursun Delen, and Yanyan Meng. 2012. Comparative analysis of data mining methods for bankruptcy prediction. *Decision Support Systems* 52, 2 (2012), 464–473. DOI: https://doi.org/10.1016/j.dss.2011.10.007

[172] Ahmet Murat Ozbayoglu, Mehmet Ugur Gudelek, and Omer Berat Sezer. 2020. Deep learning for financial applications: A survey. *Applied Soft Computing* 93 (2020), 106384.

[173] Ravisankar P., Ravi V., Raghava RG, and Bose L. 2011. Detection of financial statement fraud and feature selection using data mining techniques. *Decision Support Systems* 50, 2 (2011), 491–500.

[174] P. F. Pai, M. F. Hsu, and M. C. Wang. 2011. A support vector machine-based model for detecting top management fraud. *Knowledge-Based Systems* 24, 2 (2011), 314–321.

[175] Z. V. Palmrose and S. Scholz. 2004. The circumstances and legal consequences of Non-GAAP reporting: Evidence from restatements. *Contemporary Accounting Research* 21, 1 (2004), 139–180.

[176] Mário Papík and Lenka Papíková. 2020. Detection models for unintentional financial restatements. *ournal of Business Economics and Management* 21, 1 (2020), 64–86. DOI: https://doi.org/10.3846/jbem.2019.10179

[177] Mário Papík and Lenka Papíková. 2021. Application of selected data mining techniques in unintentional accounting error detection. *Equilibrium. Quarterly Journal of Economics and Economic Policy* 16, 1 (2021), 185–201. DOI: https://doi.org/10.24136/eq.2021.007

[178] Michal Pavlicko, Marek Durica, and Jaroslav Mazanec. 2021. Ensemble model of the financial distress prediction in visegrad group countries. *Mathematics* 9, 16 (2021), 26. DOI: https://doi.org/10.3390/math9161886

[179] M. Pazarskis, G. Lazos, A. Koutoupis, and G. Drogalas. 2021. Preventing the unpleasant: Fraudulent financial statement detection using financial ratios. *Journal of Operational Risk* 17, 1 (2021), 33–50.

[180] Guido Perboli and Ehsan Arabnezhad. 2021. A machine learning-based DSS for mid and long-term company crisis prediction. *Expert Systems with Applications* 174 (2021), 114758. DOI: https://doi.org/10.1016/j.eswa.2021.114758

[181] J. Perols. 2011. Financial statement fraud detection: An analysis of statistical and machine learning algorithms. *Auditing: A Journal of Practice and Theory* 30, 2 (2011), 19–50.

[182] M. Plumlee and T. L. Yohn. 2010. An analysis of the underlying causes attributed to restatements. *Accounting Horizons* 24, 1 (2010), 41–64.

[183] I. M. Premachandra, Gurmeet Singh Bhabra, and Toshiyuki Sueyoshi. 2009. DEA as a tool for bankruptcy assessment: A comparative study with logistic regression technique. *European Journal of Operational Research* 193, 2 (2009), 412–424. DOI: https://doi.org/10.1016/j.ejor.2007.11.036

[184] Maria Psillaki, Ioannis E. Tsolas, and Dimitris Margaritis. 2010. Evaluation of credit risk based on firm performance. *European Journal of Operational Research* 201, 3 (2010), 873–881. DOI: https://doi.org/10.1016/j.ejor.2009.03.032

[185] A. W. Pupokusumo, B. L. Handoko, Willy, Ricky, and Hendra E. 2022. Benford's law as a tool in detecting financial statement fraud. *Journal of Theoretical and Applied Information Technology* 100, 14 (2022), 5300–5305.
[186] L. Purda and D. Skillicorn. 2015. Accounting variables, deception, and a bag of words: Assessing the tools of fraud detection. *Contemporary Accounting Research* 32, 3 (2015), 1193–1223.
[187] Hongyi Qian, Baohui Wang, Minghe Yuan, Songfeng Gao, and You Song. 2022. Financial distress prediction using a corrected feature selection measure and gradient boosted decision tree. *Expert Systems with Applications* 190 (2022), 116202. DOI: https://doi.org/10.1016/j.eswa.2021.116202
[188] Y. Qu, P. Quan, M. Lei, and Y. Shi. 2019. Review of bankruptcy prediction using machine learning and deep learning techniques. *Procedia Comouter Science* 162, 4 (2019), 895–899.
[189] Balakrishnan R., Qiu X.Y., and Srinivasan P. 2010. On the predictive ability of narrative disclosures in annual reports. *European Journal of Operational Research* 202, 3 (2010), 789–801.
[190] S. Richardson, I. Tuna, and M. Wu. 2002. Predicting earnings management: The case of earnings restatements. *Journal of Accounting Research* 22, 1 (2002), 59–82.
[191] S. A. Ross, R. W. Westerfield, and J. F. Jaffe. 1999. *Corporate Finance*. Homewood IL, IL, USA.
[192] Yazan F. Roumani, Joseph K. Nwankpa, and Mohan Tanniru. 2020. Predicting firm failure in the software industry. *Artificial Intelligence Review* 53, 6 (2020), 4161–4182. DOI: https://doi.org/10.1007/s10462-019-09789-2
[193] I. Sadgali, N. Sael, and F. Benabbou. 2019. Performance of machine learning techniques in the detection of financial frauds. *Procedia Computer Science* 148 (2019), 45–54. DOI: https://doi.org/10.1016/j.procs.2019.01.007. THE SECOND INTERNATIONAL CONFERENCE ON INTELLIGENT COMPUTING IN DATA SCIENCES, ICDS2018.
[194] Carlos Serrano-Cinca and Begoña Gutiérrez-Nieto. 2013. Partial least square discriminant analysis for bankruptcy prediction. *Decision Support Systems* 54, 3 (2013), 1245–1255. DOI: https://doi.org/10.1016/j.dss.2012.11.015
[195] T. Shahana, Vilvanathan Lavanya, and Aamir Rashid Bhat. 2023. State of the art in financial statement fraud detection: A systematic review. *Technological Forecasting and Social Change* 192 (2023), 122527. DOI: https://doi.org/10.1016/j.techfore.2023.122527
[196] A. Sharma and P. K. Panigrahi. 2012. A review of financial accounting fraud detection based on data mining techniques. *International Journal of Computer Applications* 39, 1 (2012), 37–47.
[197] Sakshi Sharma and Bhag Singh Bodla. 2024. A bibliometric analysis of corporate bankruptcy prediction research: Trends and insights. *Journal of Informatics Education and Research* 4, 2 (2024), 175–190.
[198] K. S. Shin, T. S. Lee, and H. Kim. 2005. An application of support vector machines in bankruptcy prediction model. *Expert Systems with Applications* 28, 1 (2005), 127–135.
[199] Andreas Sideras, Konstantinos Bougiatiotis, Elias Zavitsanos, Georgios Paliouras, and George Vouros. 2024. Bankruptcy prediction: Data augmentation, LLMs and the need for auditor's opinion. In *Proceedings of the 5th International Conference on AI in Finance*. ACM, New York, USA, 453–460.
[200] D. Skillicorn and L. Purda. 2012. Detecting fraud in financial reports. In *Proceedings of the European Intelligence and Security Informatics Conference*. IEEE, Odense, Denmark Denmark, 7–13.
[201] Matthew Smith and Francisco Alvarez. 2022. Predicting firm-level bankruptcy in the spanish economy using extreme gradient boosting. *Computational Economics* 59, 1 (2022), 263–295. DOI: https://doi.org/10.1007/s10614-020-10078-
[202] Salima Smiti and Makram Soui. 2020. Bankruptcy prediction using deep learning approach based on borderline SMOTE. *Information Systems Frontiers* 22, 5 (2020), 1067–1083. DOI: https://doi.org/10.1007/s10796-020-10031-
[203] Xin-Ping Song, Zhi-Hua Hu, Jian-Guo Du, and Zhao-Han Sheng. 2014. Application of machine learning methods to risk assessment of financial statement fraud: Evidence from China. *Journal of Forecasting* 33, 8 (2014), 611–626,.
[204] Ch Spathis, Michael Doumpos, and Constantine Zopounidis. 2002. Detecting falsified financial statements: A comparative study using multicriteria analysis and multivariate statistical techniques. *Accounting Review* 11, 3 (2002), 509–535.
[205] Matthew Stevenson, Christophe Mues, and Cristián Bravo. 2021. The value of text for small business default prediction: A Deep Learning approach. *European Journal of Operational Research* 295, 2 (2021), 758–771. DOI: https://doi.org/10.1016/j.ejor.2021.03.008
[206] J. Sun and H. Li. 2012. Financial distress prediction using support vector machines: Ensemble vs. individual. *Applied Soft Computing* 12, 8 (2012), 2254–2265.
[207] J. Sun, H. Li, Q. Huang, and K. He. 2014. Predicting financial distress and corporate failure: A review from the state of the art definitions, modeling, sampling and featuring approaches. *Knowledge-Based Systems* 57, 4 (2014), 41–56.
[208] Fernando Sánchez-Lasheras, Javier de Andrés, Pedro Lorca, and Francisco Javier de Cos Juez. 2012. A hybrid device for the solution of sampling bias problems in the forecasting of firms' bankruptcy. *Expert Systems with Applications* 39, 8 (2012), 7512–7523. DOI: https://doi.org/10.1016/j.eswa.2012.01.135
[209] Eric Séverin and David Veganzones. 2021. Can earnings management information improve bankruptcy prediction models? *Annals of Operations Research* 306, 1 (2021), 247–272. DOI: https://doi.org/10.1007/s10479-021-04183-

[210] Xiaobo Tang, Shixuan Li, Mingliang Tan, and Wenxuan Shi. 2020. Incorporating textual and management factors into financial distress prediction: A comparative study of machine learning methods. *Journal of Forecasting* 39, 5 (2020), 769–787. DOI: https://doi.org/10.1002/for.2661

[211] P. C. Tetlock. 2007. Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance* 62, 3 (2007), 1139–1168.

[212] S. Thiruvadi and S. C. Patel. 2011. Survey of data mining techniques used in fraud detection and prevention. *Information Technology Journal* 10, 4 (2011), 710–716.

[213] S. Tian, Y. Yu, and H. Guo. 2008. Variable selection and corporate bankruptcy forecasts. *Journal of Banking and Finance* 52 (2008), 89–100.

[214] F. H. Tunio, Y. Ding, A. N. Agha, K. Agha, and H. Panhwar. 2021. Financial distress prediction using and bagging in Pakistan stock exchange. *Journal of Asian Finance* 1, 1 (2021), 665–673.

[215] K. UlagaPriya and S. Pushpa. 2021. A comprehensive study on ensemble-based imbalanced data classification methods for bankruptcy data. In *Proceedings of the 2021 6th International Conference on Inventive Computation Technologies* . IEEE, Coimbatore, India, 800–804. DOI: https://doi.org/10.1109/ICICT50816.2021.9358744

[216] David Veganzones, Eric Séverin, and Souhir Chlibi. 2023. Influence of earnings management on forecasting corporate failure. *International Journal of Forecasting* 39, 1 (2023), 123–143. DOI: https://doi.org/10.1016/j.ijforecast.2021.09.006

[217] S. Virdhagriswaran and G. Dakin. 2006. Camouflaged fraud detection in domains with complex relationships. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, USA, 941–947.

[218] Chuan-Ju Wang, Ming-Feng Tsai, Tse Liu, and Chin-Ting Chang. 2013. Financial sentiment analysis for risk prediction. In *Proceedings of the 6th International Joint Conference on Natural Language Processing*. Asian Federation of Natural Language Processing, Nagoya, Japan, 802–808.

[219] Nanxi Wang. 2017. Bankruptcy prediction using machine learning. *Journal of Mathematical Finance* 07, 04 (2017), 908–918. DOI: https://doi.org/10.4236/jmf.2017.74049

[220] Xinlin Wang, Zs'ofia Kraussl, and Mats Brorsson. 2024. Datasets for advanced bankruptcy prediction: A survey and taxonomy. *Papers 2411.01928, arXiv.org*.

[221] D. Whiting, J. Hansen, B. McDonald, C. Albrecht, and W. Albrecht. 2012. Machine learning methods for detecting patterns of management fraud. *Computational Intelligence* 28 (2012), 505–527.

[222] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. BloombergGPT: A large language model for finance. *ArXiv abs/2303.17564* (2023), n. pag.

[223] Y. Xia, C. Liu, Y. Li, and N. Liu. 2016. A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. *Expert Systems with Applications* 78 (2016), 225–241.

[224] Wu Xiuguo and Du Shengyong. 2022. An analysis on financial statement fraud detection for chinese listed companies using deep learning. *IEEE Access* 10 (2022), 22516–22532. DOI: https://doi.org/10.1109/ACCESS.2022.3153478

[225] Xin Xu, Feng Xiong, and Zhe An. 2023. Using machine learning to predict corporate fraud: Evidence based on the GONE framework. *Journal of Business Ethics* 186, 1 (2023), 137–158. DOI: https://doi.org/10.1007/s10551-022-05120-2

[226] Hongyang Yang, Xiao-Yang Liu, Christina Dan Wang. 2023. FinGPT: Open-source financial large language models (June 9, 2023). *FinLLM at IJCAI 2023*, Available at SSRN: https://ssrn.com/abstract=4489826 or http://dx.doi.org/10.2139/ssrn.4489826

[227] Xinze Yang, Chunkai Zhang, Yizhi Sun, Kairui Pang, Luru Jing, Shiyun Wa, and Chunli Lv. 2023. FinChain-BERT: A high-accuracy automatic fraud detection model based on NLP methods for financial scenarios. *Information* 14, 9 (2023), 26. DOI: https://doi.org/10.3390/info14090499

[228] G. Ji Z. Yang, W. You. 2011. Using partial least squares and support vector machines for bankruptcy prediction. *Expert Systems and Applications* 38, 7 (2011), 8336–8342.

[229] Elias Zavitsanos, Aris Kosmopoulos, George Giannakopoulos, Marina Litvak, Blanca Carbajo-Coronado, Antonio Moreno-Sandoval, and Mo El-Haj. 2023. The financial narrative summarisation shared task (FNS 2023). In *Proceedings of the 2023 IEEE International Conference on Big Data*. IEEE, IEEE, Sorrento, Italy, 2890–2896.

[230] Elias Zavitsanos, Dimitris Mavroeidis, Konstantinos Bougiatiotis, Eirini Spyropoulou, Lefteris Loukas, and Georgios Paliouras. 2021. Financial misstatement detection: A realistic evaluation. In *Proceedings of the 2nd ACM International Conference on AI in Finance*. ACM, USA, online, 0.

[231] Elias Zavitsanos, Dimitris Mavroeidis, Eirini Spyropoulou, Manos Fergadiotis, and Paliouras Georgios. 2024. ENTRANT: A large financial dataset for table understanding. *Nature Scientific Data* 11 (2024), 876. DOI: https://doi.org/10.1038/s41597-024-03605-5

[232] E. Zavitsanos, D. Mavroeidis, E. Spyropoulou, M. Fergadiotis, and G. Paliouras. 2024. A Large Financial Dataset for Table Understanding [Data set]. DOI: https://doi.org/10.5281/zenodo.10667088

[233] Yuri Zelenkov and Nikita Volodarskiy. 2021. Bankruptcy prediction on the base of the unbalanced data using multi-objective selection of classifiers. *Expert Systems with Applications* 185 (2021), 115559. DOI: https://doi.org/10.1016/j.eswa.2021.115559
[234] C. (Abigail) Zhang, S. Cho, and M. Vasarhelyi. 2022. Explainable artificial intelligence (XAI) in auditing. *International Journal of Accounting Information Systems* 1 (2022), 58. DOI: https://doi.org/10.2139/ssrn.3981918
[235] H. Zhao, Z. Liu, Z. Wu, Y. Li, T. Yang, P. Shu, S. Xu, H. Dai, L. Zhao, G. Mai, N. Liu, and T. Liu. 2024. Revolutionizing finance with LLMs: An overview of applications and insights. *ArXiv, abs/2401.11641.*
[236] Shuping Zhao, Kai Xu, Zhao Wang, Changyong Liang, Wenxing Lu, and Bo Chen. 2022. Financial distress prediction by combining sentiment tone features. *Economic Modelling* 106 (2022), 105709. DOI: https://doi.org/10.1016/j.econmod.2021.105709
[237] L. Zhou, J. Burgoon, J. Nunamaker, and D. Twitchell. 2004. Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communication. *Group Decission and Negotiation* 13, 1 (2004), 81–106.
[238] M. Zieba, S.K. Tomczak, and J.M. Tomczak. 2016. Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction. *Expert Systems with Applications* 58 (2016), 93–101.
[239] Martin Zoričák, Peter Gnip, Peter Drotár, and Vladimír Gazda. 2020. Bankruptcy prediction for small- and medium-sized companies using severely imbalanced datasets. *Economic Modelling* 84 (2020), 165–176. DOI: https://doi.org/10.1016/j.econmod.2019.04.003

## Appendix

### A Grouping of Related Work

Table 3 provides a grouping of the related work based on dataset size, input features, evaluation method and type of ML method used.

Table 3. Grouping of Related Work based on Dataset Size, Input Features, Evaluation Measure, and Type of ML Method

| Publication | Data size | Input type | Evaluation | Type of method |
|---|---|---|---|---|
| [217] | 1700 samples, 49 pos | Fin. variables | Sensitivity | Supervised |
| [125] | 27 firms, 10 pos | Fin. variables | TP rate | Other |
| [131] | 164 firms, 41 pos | Fin. variables | Accuracy | Supervised |
| [58] | 100 samples | Fin. variables | Accuracy | Unsupervised |
| [166] | 408 samples, 102 pos | Fin. variables | Acc. TP, TN, Kappa | Supervised |
| [100] | 622 samples, 311 pos | Text and fin. variables | Acc. AUC, F1, TP, TN | Supervised |
| [200] | 5697 samples, 1127 pos | Text | Accuracy | Supervised |
| [221] | 228 firms, 114 pos | Fin. variables | AUC | Supervised |
| [204] | 75 samples, 38 pos | Fin. variables | Acc. Recall | Supervised |
| [120] | 158 samples, 79 pos | Fin. variables | Accuracy | Other |
| [106] | 390 firms, 51 pos | Fin. variables | Accuracy | Other |
| [148] | 1515 firms | Fin. variables | Accuracy | Supervised |
| [19] | 148 firms, 24 pos | Fin. variables | Accuracy | Supervised |
| [90] | 100 samples | Text | Accuracy | Supervised |
| [174] | 75 samples, 25 pos | Fin. variables | Accuracy | Supervised |
| [130] | 76 samples, 38 pos | Fin. variables | Accuracy | Supervised |
| [113] | 202 firms, 101 pos | Text | Accuracy | Supervised |
| [1] | 9000 firms, 815 pos | Fin. variables | AUC | Other |
| [146] | 576 firms, 129 pos | Fin. variables | Accuracy | Supervised |
| [70] | 64000 samples, 788 pos | Fin. variables | Acc., AUC, Specificity | Supervised |
| [41] | 3319 firms, 132 pos | Fin. variables | AUC, Recall | Supervised |
| [104] | 630 samples | Fin. variables | Turnover measure | Other |
| [56] | 2190 samples, 896 pos | Fin. variables | F-score | Other |

(Continued)

Table 3. Continued

| Publication | Data size | Input type | Evaluation | Type of method |
|---|---|---|---|---|
| [161] | 278000 samples, 8033 pos | Text and fin. variables | AUC, Log score | Supervised |
| [92] | 1027 samples, 405 pos | Text | Accuracy | Supervised |
| [128] | 3000 samples, 788 pos | Fin. variables | Accuracy | Supervised |
| [137] | 312 firms | Fin. variables | Accuracy | Supervised |
| [97] | 181 samples, 86 pos | Fin. variables | Acc., Recall | Supervised |
| [81] | 132 firms | Fin. variables | Acc., Error rate | Supervised |
| [181] | 15900 samples | Fin. variables | ERC | Supervised |
| [85] | 398 firms, 199 pos | Fin. variables | Accuracy | Supervised |
| [186] | 6025 samples, 1130 pos | Text | Acc., Recall | Supervised |
| [109] | 144 firms, 72 pos | Fin. variables | Type I and II error | Supervised |
| [173] | 202 firms, 101 pos | Fin. variables | Acc., Sens/ty, Spec/ty | Supervised |
| [91] | 7100 samples, 405 pos | Text | p-value | Other |
| [154] | 122 firms, 61 pos | Text | Type I and II error | Supervised |
| [86] | 7200 samples | Text | Economic losses | Supervised |
| [52] | 1064 samples, 201 pos | Text and fin. variables | Acc., AUC, Sens/ty, Spec/ty | Supervised |
| [112] | 64 firms, 32 pos | Fin. variables | Accuracy | Other |
| [32] | 461 neg, 1387 pos | Fin. variables | Accuracy | Supervised |
| [219] | 32 neg, 32 pos | Fin. variables | Accuracy | Supervised |
| [210] | 212 neg, 212 pos | Fin. variables | Accuracy, AUC | Supervised |
| [170] | 290 neg, 131 pos | Fin. variables | Accuracy, P, R | Supervised |
| [156] | 1% positives | Text and fin. variables | Accuracy, AUC | Supervised |
| [64] | 1.29% positives | Fin. variables | AUROC | Supervised |
| [55] | 59K samples, 138 pos | Fin. variables | Mean bankrupt prob. | Unsupervised |
| [208] | 63K neg, 256 pos | Fin. variables | Accuracy | Unsupervised |
| [7] | 248 neg, 8 pos | Fin. variables | Failing prob. | Unsupervised |
| [3] | 841 neg, 420 pos | Fin. variables | Missclas. cost | Unsupervised |
| [31] | 44 neg, 21 pos | Fin. variables | Accuracy | Unsupervised |
| [20] | 1% positives | Raw indices | AUC, NDCG, Precision@K | Supervised |
| [230] | 1.1% positives | Text and fin. variables | R-precision | Supervised |
| [47] | 786 neg, 262 pos | Fin. variables | AIC value | Supervised |
| [63] | 120 neg, 120 pos | Fin. variables | ROC | Supervised |
| [49] | 500 neg, 500 pos | Fin. variables | Class. ratio | Supervised |
| [62] | 71 neg, 65 pos | Fin. variables | Acc., Sens/ty, Spec/ty | Supervised |
| [139] | 153 samples | Fin. variables | Accuracy | Supervised |
| [140] | 135 samples | Fin. variables | Accuracy | Supervised |
| [147] | 120 neg, 120 pos | Fin. variables | Accuracy | Supervised |
| [228] | 64 neg, 56 pos | Fin. variables | Sens/ty, Spec/ty | Supervised |
| [144] | 239 neg, 239 pos | Fin. variables | Acc., Type I, II error | Supervised |
| [114] | 629 samples | Fin. variables | Accuracy | Supervised |
| [162] | 262 samples | Text | AUC | Other |
| [36] | 1.3% positives | Fin. variables | Failure prob. | Other |
| [194] | 7973 neg, 320 pos | Fin. variables | Acc., F-score, Type I,II | Supervised |
| [87] | 250 samples | Fin. and risk variables | Accuracy | Supervised |
| [171] | 624 neg, 697 pos | Fin. variables | Accuracy | Supervised |
| [76] | 42 samples | Fin. variables | Acc., Sens/ty, Spec/ty | Supervised |
| [203] | 440 neg, 110 pos | Fin. variables | Accuracy | Supervised |
| [142] | 45 neg, 12 pos | Fin. variables | Accuracy | Supervised |

(Continued)

Table 3. Continued

| Publication | Data size | Input type | Evaluation | Type of method |
|---|---|---|---|---|
| [149] | 160 neg, 138 pos | Fin. variables | Accuracy | Supervised |
| [225] | 35.9K neg, 4.4K pos | Fin. variables | AUC, NDCG, P, R | Supervised |
| [224] | 4886 neg, 244 pos | Text and fin. variables | AUC, Acc., P, R | Supervised |
| [185] | 1:1 ratio | Fin. variables | Other | Other |
| [179] | 23 neg, 23 pos | Fin. variables | Acc., Sens/ty, Spec/ty | Supervised |
| [163] | 25 neg, 25 pos | Fin. variables | Chi-square | Other |
| [176] | 200 neg, 200 pos | Fin., categorical | Acc, P, R, F1 | Supervised |
| [177] | 200 neg, 200 pos | Fin., categorical | Acc, P, R, F1 | Supervised |
| [89] | 464 neg, 464 pos | Fin., categorical | Other | Supervised |
| [102] | 1384 neg, 321 pos | Fin. variables | Acc, Precision | Supervised |
| [118] | 135 neg, 45 pos | Fin. variables | TPR, FPR | Other |
| [117] | 102 neg, 51 pos | Text and fin. variables | Acc, P, R, F1 | Supervised |
| [101] | 68 firms | Fin. variables | Other | Other |
| [105] | 8 firms | Fin. variables | Accuracy | Other |
| [11] | 31.6K neg, 1.5K pos | Fin. variables | Acc. P, R, F1 | Supervised |
| [4] | 244 firms | Fin. variables | AUC, PRAUC, Acc. | Supervised |
| [8] | 2859 samples | Fin., categorical | Acc., P, R, Type I, II | Supervised |
| [44] | 11K samples, 422 pos | Fin. variables | AUC | Supervised |
| [25] | 1850 firms | Fin. variables | Accuracy | Supervised |
| [53] | 218 neg, 820 pos | Text and fin. variables | Acc., AUC, R | Supervised |
| [116] | 133 firms | Fin. variables | Acc., AUC | Supervised |
| [187] | 4366 firms | Fin. variables | AUC | Supervised |
| [233] | 2.4K neg, 456 pos | Fin. variables | AUC | Supervised |
| [124] | 5K-10K neg, 200-515 pos | Fin. variables | AUC, P, R, F1 | Supervised |
| [178] | 558K neg, 109K pos | Fin. variables | AUC, F1 | Supervised |
| [180] | 3000 firms | Fin. variables | AUC, P, R, F1 | Supervised |
| [236] | 1360 neg, 67 pos | Text and fin. variables | AUC | Supervised |
| [9] | 1651 firms, 194 pos | Fin., categorical | Acc., AUC, P, R | Supervised |
| [78] | 2860 firms, 62 pos | Fin., categorical | Acc., Type I, II | Supervised |
| [215] | 1000 firms | Fin. variables | AUC, P, R, F1 | Supervised |
| [216] | 4000 firms | Fin. variables | AUC | Supervised |
| [119] | 1128 neg, 69 pos | Text and fin. variables | AUC, other | Supervised |
| [202] | 5K-10K neg, 200-515 pos | Fin. variables | AUC, Sens/ty, Spec/ty | Supervised |
| [121] | 68 neg, 68 pos | Fin., categorical | Other | Supervised |
| [192] | 408 firms, 43 pos | Fin. variables | Acc., P, R, F1 | Supervised |
| [2] | 640K firms | Text | P, R, F1 | Supervised |
| [6] | 503K neg, 138K pos | Fin. variables | Acc., Spec/ty, Sens/ty | Supervised |
| [239] | 1.2K-5.8K neg, 14-30 pos | Fin. variables | Other | Unsupervised |
| [157] | 6725 neg, 274 pos | Fin. variables | Acc., Spec/ty, Sens/ty | Unsupervised |
| [108] | 2168 firms | Fin. variables | F1 | Supervised |
| [201] | 58K neg, 6057 pos | Fin. variables | Spec/ty, Sens/ty | Supervised |
| [209] | 4000 firms | Fin. variables | Acc., Type I, II | Other |
| [5] | 200 firms | Fin. variables | Acc., P, R | Supervised |
| [28] | 30.6K neg, 289 pos | Text and fin. variables | AUC, NDCG | Supervised |
| [150] | 5688 neg, 502 pos | Text and fin. variables | Accuracy | Supervised |

Received 27 October 2022; revised 19 November 2024; accepted 7 March 2025