Przegląd Statystyczny. Statistical Review, 2022, vol. 69, 3, 1–22 DOI: 10.5604/01.3001.0016.1165
Received: 26.09.2022, revised: 8.11.2022, accepted: 28.11.2022

From business to clinical trials: a systematic review of the literature on fraud detection methods to be used in central statistical monitoring

Maciej Fronc[a], Michał Jakubczyk[b]

Abstract. Data-driven decisions can be suboptimal when the data are distorted by fraudulent behaviour. Fraud is a common occurrence in finance or other related industries, where large datasets are handled and motivation for financial gain may be high. In order to detect and the prevent fraud, quantitative methods are used. Fraud, however, is also committed in other circumstances, e.g. during clinical trials. The article aims to verify which analytical fraud--detection methods used in finance may be adopted in the field of clinical trials. We systematically reviewed papers published over the last five years in two databases (Scopus and the Web of Science) in the field of economics, finance, management and business in general. We considered a broad scope of data mining techniques including artificial intelligence algorithms. As a result, 37 quantitative methods were identified with the potential of being fit for application in clinical trials. The methods were grouped into three categories: pre--processing techniques, supervised learning and unsupervised learning. Our findings may enhance the future use of fraud-detection methods in clinical trials.

Keywords: fraud detection, clinical trials, finance, data mining, big data
JEL: C00, C38, C55

1. Introduction

The understanding of the term 'fraud' changes across different fields of study. It can have various potential meanings depending on the scope of the investigated issues, as claimed by Gupta (2013), and West and Bhattacharya (2016). In this paper, however, we define fraud as an intentional manipulation of data (e.g. fabrication, falsification or deletion) or misconduct in the data production process caused by personal motivation of a fraudster or their carelessness.

Fraud is a common problem observed not only in the financial sector but also in everyday life and can expose both transaction parties to huge losses (Al-Hashedi & Magalingam, 2021). According to the US Federal Trade Commission (2022),

[a] SGH Warsaw School of Economics, Institute of Econometrics, Decision Analysis and Support Unit, GSK, Central Monitoring and Data Analytics, al. Niepodległości 162, 02-554 Warszawa, e-mail: mf85106@doktorant.sgh.waw.pl, ORCID: https://orcid.org/0000-0002-4874-950X.
[b] SGH Warsaw School of Economics, Institute of Econometrics, Decision Analysis and Support Unit, al. Niepodległości 162, 02-554 Warszawa, e-mail: mjakubc@sgh.waw.pl, ORCID: https://orcid.org/0000-0002-0006-6769.

2.8 million fraud cases of different type were registered in the USA in 2021 only, and the total generated financial loss exceeded $5.8 billion. *PwC's Global Economic Crime and Fraude Survey 2022* (PwC, 2022) showed that 46% of the surveyed companies reported fraud in the last 24 months.

The rapid development of information systems and the progressing digitalisation of data necessitate the development of methods allowing large datasets to be handled properly. Al-Hashedi and Magalingam (2021) explained how the technological revolution extended the opportunities for committing fraud. The transfer of money and related activities became easier as digital technologies developed, which, in turn, has made banking activities vulnerable to deception. As a result, a significant increase of fraudulent schemes in finance is observed. The vast majority of data on financial activities are stored in databases, which causes the volume of such data to increase. Therefore, the application of relevant analytical tools is necessary to mitigate the risk associated with fraud. Moreover, such tools facilitate the decision-making process as to where the resources of an organisation should be allocated most efficiently (Zhou & Kapoor, 2011).

Financial data is relatively common and brings a significant portion of information, which allows us to distinguish activities which are fraudulent from the non-fraudulent ones with the use of analytical tools based on statistics. Registered transactions constitute the main source of financial data. They come from sectors such as banking, the bond market, the securities market, transaction systems, financial statements, etc. (Zhang et al., 2022).

Apart from the financial sector, other fields are also vulnerable to fraud – like clinical trials (CTs). According to the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (2016), a CT is 'any investigation in human subjects intended to discover or verify the clinical, pharmacological and/or other pharmacodynamic effects of an investigational product(s)' in order to provide its safety towards the patient as the final recipient. Within the pharmaceutical industry, it is this stage that is the most complex in the drug development process. The execution of a single study produces a tremendous amount of data that are potentially exposed to manipulation. Therefore, keeping the collected data under control is a must and can be achieved through the centralised monitoring of CTs. This monitoring process involves a remote evaluation of the clinical data resulting from a study in order to maintain the high quality of the investigated product (Kirkwood et al., 2013).

Fraud in the context of CTs might be caused by the researcher's carelessness (e.g. making mistakes in the data in the documentation), ambitions or expediency (e.g. enrolling as many patients as possible for extra financial gain even though they had

already been enrolled elsewhere). In all the cases, the obtained results and the outcome of the entire study can be affected by the researcher's motivation. Fraud in CTs might result not only in a huge financial loss but it can also undermine the credibility of the trial sponsor (Gupta, 2013; Sakamoto & Buyse, 2016). It is additionally worth noting that the history of investigating fraud in finance is far longer than that of researching fraud in CTs. What is more, this subject is relatively uncommon among researchers, as evidenced by their lack of awareness of the phenomenon (Kirkwood et al., 2013). The implementation of well-developed fraud detection techniques aligned with business objectives increases the efficiency of an organisation. However, the adoption of an approach aiming to minimise the losses caused by fraud is recommended rather than focusing on statistical measures such as likelihood (Höppner et al., 2022). Undertaking preventive measures is better than reacting to failures. Fraud detection in CTs can be further developed through the implementation of the solutions already applied in the field of finance. The aim of our study is to identify the analytical concepts that can be tailored to the specific nature of CTs and applied to detect fraud among clinical data.

The aforementioned goal provides a better insight into the methods currently applied in fraud detection. The methods were extracted and compared through a systematic review of the available literature. The review focused on papers discussing fraud detection based on quantitative methods such as data mining, machine learning, artificial intelligence or econometrics. These methods were evaluated not only in terms of their statistical performance, but also in terms of constraints when considering their potential use. Unsupervised techniques prove more useful when no labelled data is involved and when the outstanding cases are detected across the whole dataset. The outcome obtained by means of unsupervised techniques was compared with the original data labels that were intentionally omitted in the analysis. On the other hand, the performance of supervised techniques may be improved through the use of pre-processing algorithms that can handle imbalanced data. These detection methods ensured a satisfactory level of accuracy (ranging from 60% to nearly 100%) in the context of real-world decision-making problems, which altogether resulted in a more efficient resource allocation.

The paper has a following structure: Section 2 presents the adopted approach within the systematic review, the results of the review are shown in Section 3 (with an additional tabular summary of the identified methods, and final conclusions are presented in Section 4).

2. Methods

2.1. Literature search and methods extraction

We searched two databases: Scopus and the Web of Science. In the search, we used keywords related to fraud or manipulation detection (as a problem to solve) and statistics, data mining, machine learning, artificial intelligence and econometrics (i.e. the kinds of methods that we are interested in). We limited the search to papers published in the years 2018–2022. The search focused on journals in the area of business, management, accounting, decision sciences, economics, econometrics and finance. The search in the Web of Science differed slightly from that in Scopus due to the fact that a different classification was adopted in the former data base. We therefore focused on the following subject areas: business, economics, operations research, management sciences and mathematical methods in social sciences. Eventually, the searches in both databases produced similar results. Considering the manageability of the review and the fact that new fraud detection methods are constantly developed, we believe that such a narrowing was warranted. Specific keywords used in the queries are listed in the appendix.

The papers selected during the search were subsequently analysed, and we included papers which met all the following criteria:

- research or review based on quantitative methods applicable to fraud detection in the aforementioned areas;
- research based on real-life data or analysis-ready datasets;
- papers focusing on detecting committed frauds;
- articles describing algorithms that can be applied to solve related problems;
- articles including fraud detection methods applicable to labeled and unlabeled data;
- articles including methods that handle imbalanced data.

We excluded papers that met at least one of the criteria below:

- duplicated papers;
- articles that do not propose any particular methods of fraud detection.

At the second stage of the selection those articles were excluded that:

- considered fraud in terms of the behavioural aspects of fraudsters' motivation;
- involved methods with a narrow range of application.

2.2. Systematisation of the papers

We extracted information on fraud detection methods from the selected papers. These methods were then categorised according to their use into data pre-processing techniques and supervised and unsupervised methods of fraud detection.

Pre-processing techniques refer to operations performed on an initial dataset in order to prepare raw data for a proper analysis by reducing their inherent complexity. This data modification was necessary to obtain a dataset which would improve the performance of an applied algorithm and provide higher quality results. In our paper, techniques of this kind focused on resampling and attribute selection, which proves useful in handling skewed data.

Supervised learning algorithms use labelled data (training sets) as the basis for patterns recognition among a newly implemented dataset (test set). In our work, the labels represented the particular classes that the new data points were assigned to. For this approach it was crucial to know in advance which predefined classes the training datapoints belong to as the starting point for the learning process.

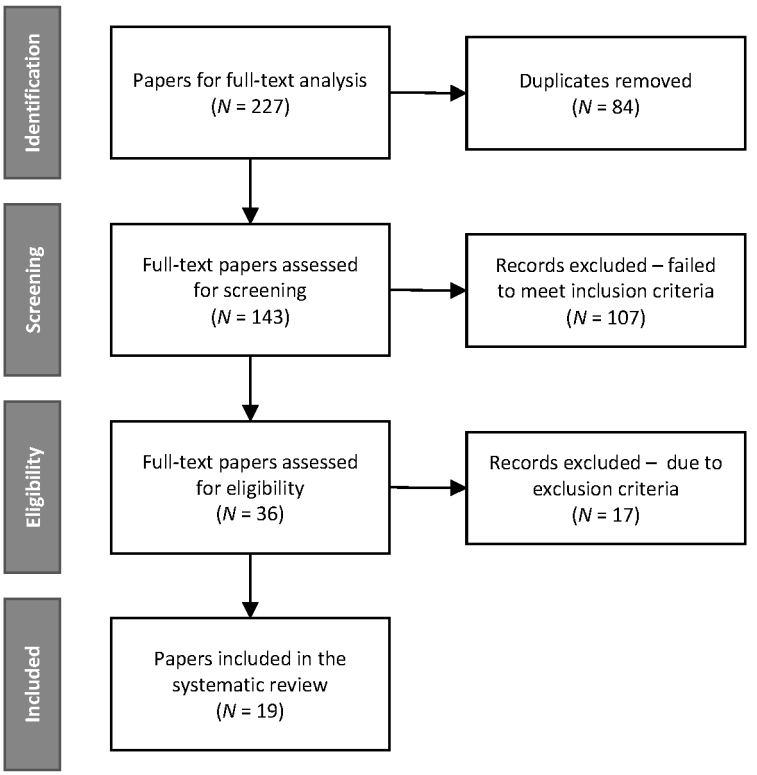
Unsupervised learning enables splitting an initial dataset into subgroups without any *a priori* information about their categories. In contrast to supervised learning, unsupervised learning focuses on the differences and similarities between datapoints in space which are created by their attributes. These kinds of algorithms aim to disclose the hidden patterns from the information included in the processed data.

The proposed division takes into consideration the data characteristics and fraud specificity. In the next section, each of the extracted methods is discussed. A toolbox presented in Section 3 contains the methods that are already used in the central monitoring of CTs, or may potentially be applied as an extension to the current methodology.

3. Results

The systematic literature review allowed the compilation of a total of 19 papers which met the selection criteria. These papers presented 37 quantitative techniques applicable for direct fraud detection or as supportive tools. Figure illustrates the number of publications, *N*, obtained at each stage of the search process. The identified methods were divided into three categories, as specified in the Methods section: pre-processing, supervised learning and unsupervised learning. Unsupervised algorithms (15 techniques) proved to be the most frequently occurring category compared to supervised learning (12) and pre-processing (10). The prevalence of unsupervised techniques might result from the specificity of the available data which were mostly unlabelled (due to the lack of prior knowledge as to the fraudulent activity among the collected data). All these methods are summarised in Table at the end of this section.

Figure. Stages of the search process for a systematic review



Source: authors' work.

3.1. Pre-processing methods

Ekin et al. (2021) devised a classification of financial frauds committed in the healthcare sector. They studied overpayment issues based on data on healthcare payment claims. The data included billings for the services of physician assistants and interventional pain management providers. The authors focused on coping with imbalanced data, which is typical for problems of this kind; fraud cases in most instances form a minority in the whole dataset. Therefore, this subset should be oversampled to avoid disproportion among the data. Ekin et al. used oversampling of the informative minority data points and undersampling of the non-fraudulent cases. The following oversampling algorithms were used: synthetic minority oversampling technique (SMOTE), majority weighted minority oversampling technique (MWMOTE), and random walk oversampling (RWO). All of these techniques involve generating synthetic data based on minority classes in a dataset. The performance of these techniques depends on the imbalance ratio which refers to

a fraction of the minority class in the whole dataset. The study demonstrated that RWO is the most efficient sampling method among all the investigated techniques in terms of the performance measured by AUC (0.84).

As regards undersampling algorithms, only one was applied, i.e. random undersampling (RU). This technique involved removing random cases out of the majority class. RU performed poorer in terms of model quality than the oversampling methods. The exception was the com-putational time which was the shortest for RU, which resulted from the fact that it was the smallest dataset analysed.

Kamalov (2020) investigated kernel density estimation (KDE) to verify the issue of handling imbalanced data. He tested this algorithm on simulated data. KDE is a technique which estimates the unknown probability density distribution based on a sample (Botev et al., 2010). This enables the generation of new datapoints according to the distribution of the minority class in order to remove the data imbalance. KDE is flexible because of the possibility to use different kernel functions, which allows the customisation of the sampling process to be done. This method is popular among researchers and well-investigated, which makes it an attractive solution to implement. Kamalov compared the KDE performance to other pre-processing techniques such as random oversampling (ROS), SMOTE, the adaptive synthetic sampling approach (ADASYN) and NearMiss. ROS creates new data points by simply resampling the minority class. ADASYN works similarly to SMOTE, but it generates more data points at the edge of the minority class. NearMiss undersamples the majority class, which causes loss of information. Each of these algorithms were tested respectively to three imbalance ratios (70, 80 and 90%). Regardless of the performance measure, NearMiss worked the least efficiently, whereas the effectiveness of the performance of the rest of the algorithms was similar. However, KDE performed best while measured by the G-mean and *F*1-score. According to the AUC measured, KDE was the best at an 80% imbalance ratio. For the rest of the ratio values, KDE was the second best.

Przekop (2020) proposed a solution to cope with cases handling too many variables in the form of feature engineering (FE). His experiment was based on real-life data which came from new customers' bank applications. Przekop indicated two different approaches within FE. The first involved using a combination of variables instead of a single one. This allowed the dimensionality of an investigated issue to be reduced as unique combinations of variables were indicated by means of a decision-tree-based algorithm. The second approach involved the segmentation of the population into homogenous peer groups. This made it possible to describe each of the group by a set of variables that were specific to a given segment. The proposed approaches improved the predictive power of the fraud detection models by

determining the specific relationships between variables. However, both approaches required expert knowledge about the investigated phenomenon. Przekop also claimed that fraud detection methods tended to be very general and in order to provide the best possible performance of the applied algorithm, the methods required an approach tailored to a specific problem.

Ekin et al. (2021) applied the principal component analysis (PCA) to address multicollinearity between variables. PCA discloses hidden variables represented by principal components that are linear combinations of the original variables, thus reducing the dimensionality of an issue. The choice of the number of the principal components total variance can be controlled. In conclusion, PCA improves the model performance in terms of the computational time and data storage parameters.

3.2. Supervised methods

As already mentioned, Ekin et al. (2021) addressed the problem of financial fraud in healthcare. They also investigated a selection of supervised techniques in terms of their application in fraud detection. Linear discriminant analysis (LDA1) is a classifier based on a linear combination of independent variables. It splits the whole dataset into two classes. LDA1 is dedicated to problems involving dichotomous variables. According to the authors, LDA1 underperforms as a fraud detection method. It copes better with undersampled data, although it is not a distinguishing feature of this method. LDA1 works better with smaller and more homogenous datasets, which causes the method to be resistant to imbalanced data. The authors also proposed the quadratic discriminant analysis (QDA), which is a variant of LDA1. The two techniques therefore share certain characteristics, although in contrast to LDA1, QDA is based on a non-linear combination of independent variables. Moreover, QDA produces better results when dealing with pre-processed data based on sampling and collinearity reduction.

Decision trees (DTs) are one of the most popular methods used for fraud classification in healthcare (Ekin et al., 2021). DTs aim to assign the objects from datasets to their relevant classes. This is a tree-structured classifier where the internal nodes represent features, branches designate the decision rules, and leaves class labels. Each node acts like a test which is a premise to binary classification. The popularity of this technique results from the fact it is relatively easy to interpret. On the other hand, DTs become less readable when too many decision rules occur. The algorithm is a greedy search technique. It does not cope with imbalanced data, nor does it distinguish small classes from the large ones. What is more, branch pruning leads to misclassification due to the use of imbalanced data. In general, DTs achieve poorer results when working on small datasets (Ekin et al., 2021), i.e. datasets

comprehensible for humans. As regards collinearity, it does not affect DTs, therefore they do not need any correlation pre-processing. When collinearity does not occur, it makes the code only run faster.

DTs can also be used in an ensembled form as a random forest (RF). This algorithm is based on the average outcome of many DTs obtained through bagging (or bootstrap aggregating), i.e. an algorithm which improves the stability and accuracy of a classifier. RF reduces variance and overfitting. It is resistant to imbalanced data, but not to a high correlation of features. A significant collinearity results in correlation bias, therefore the orthogonalisation of features is necessary (e.g. by using PCA). In contrast to DTs, RFs are intended for processing big data. This method can also be effective with small datasets, however, it would result in a lower variety of patterns. More trees generated within RF leads to a better performance of the algorithm, but on the other hand, it involves a longer processing time. However, if a hyperparameter optimisation is achieved, it makes the RF's performance independent from an increasing number of the generated trees (Wang & Xu, 2018). Ekin et al. (2021) compared the above-mentioned techniques and concluded that tree-based algorithms outperform LDA1 and QDA in terms of AUC and accuracy, but are the most time-consuming.

Wang and Xu (2018) tested the support-vector machine (SVM) as an algorithm supporting text mining within the analysis of vehicle insurance claims. This method is another dichotomous classifier applied in the space of a decision problem. It determines a hyperplane which separates the examples maximising the margin between the two classes. SVM is considered to be one of the best classifiers in terms fraud detection (if relying on references provided by Wang and Xu); therefore, it was included in this study. Although SVMs' performance is poorer than that of deep neural networks, Zhang et al. (2022) concluded that SVM is a better option in terms of classification performance. The algorithm can be combined with others to improve the final recall rate that makes fraud detection more efficient.

The *K*-nearest neighbours (*k*-NN) technique was investigated by Ekin et al. (2021). It classifies objects of an undefined membership to the existing classes. The algorithm uses distance metrics, like the Euclidean distance, to assign the unlabelled objects in the radius of *k* neighbours into the existing classes. The majority of the same class objects in the radius is the criterion of membership. Ekin et al.(2021) demonstrated that this algorithm provided the best accuracy, comparable to that of SVM. The performance of *k*-NN deteriorates as the size of the dataset becomes larger. The algorithm is sensitive to imbalanced data.

Zhang et al. (2022) used the Naïve Bayes classifier (NB) in text classification as a technique characterised by high classification performance. NB is a probabilistic method based on the Bayes theorem. The technique involves the calculation of the

conditional probability of an example belonging to a certain class, assuming the independence of the considered variables. NB is sensitive to imbalanced data and there is no need to remove the collinearity. What is more, the method benefits from correlated attributes (Ekin et al., 2021) and works fast and efficiently only on small samples. It is resistant to the influence of outliers.

Höppner et al. (2022) introduced a novel method – the LASSO-regularised logistic regression (LRLL) – as an extension to the classic logistic regression (LR). The authors tested the algorithm on data produced during card transactions. LR is a classification model with a binary response. Its predictions involve modelling explanatory variables by means of a sigmoid (logistic) function. This method is indifferent to increasing data imbalance (Ekin et al., 2021). The LR outcome is summarised in the confusion matrix. In the context of cost-based modelling, the misclassification of data points can result in a huge financial loss. Therefore, this classification needs improvement. It is possible to achieve this goal and improve the resolution of the LR model through regularisation, i.e. by introducing the penalty function to the basic model. The results of the newly-devised LRLL model allow a more reliable classification which the decision-making process is based on, thus reducing the risk of financial losses.

Höppner et al. (2022) also utilised two methods that can interact with other classifiers to improve their classification performance. The first one is gradient boosting (GB). This technique addresses classification and regression problems by utilising an ensemble of underperforming models to improve their classification capabilities. By combining new models with the existing ones, the algorithm uses the loss function to minimise the overall prediction error. GBs are usually ensembled with DTs by creating gradient tree boosting (GTB). The main problem of DTs is their inaccuracy that impedes predictive learning. GTB improves the precision of trees, in most cases leaving their specific properties that are attributed to data mining intact. These properties include the ability to handle mixed type data, irrelevant outputs and the monotonic transformation of predictors. The improvement, however, at the same time causes the speed, interpretability and resistance to the misclassification of data and overlapping class distributions to decrease. GTB is applicable to cost-related problems where a reliable classification is crucial for the reasonable allocation of financial resources (Höppner et al., 2022).

Extreme Gradient Boosting (XGBoost) is another boosting algorithm based on ensembled simple trees with a weak performance, which enhances the overall performance of this classifier. XGBoost manages to generate more models and process them faster than GTB. Moreover, this model has many other advantages including the ability to handle missing data, scaling according to the dataset size,

greater effectiveness than other GB algorithms, and the ability to rank features according to their importance in the model.

Farrugia et al. (2020) applied the XGBoost classifier to find illicit accounts within the transaction history in the Ethereum network. The authors assert that the method has a potential in many areas involving financial data and it allows the prediction of fraudulent activities based on the detected patterns. Farrugia et al. used balanced data and obtained results with satisfying metrics demonstrating a high performance of the method. The data, however, need to be pre-processed.

Rousseeuw et al. (2019) proposed a method of time series monitoring in terms of unusual patterns. Time series (TS) is a sequence of datapoints as a function of time. Measurements are taken at the same time step. TS might be decomposed to the following elements: trend, seasonal variations and random fluctuations. Monitoring the course of the time series allows the detection of outliers with fraudulent causes. This kind of occurrence might be misleading for conventional TS analysis and produce faulty results. Fraud is observed within time series when a temporary anomaly or level shift occurs. Rousseeuw et al. tested their method on airline and trade data. The algorithm adjusts the curve to the time series, although without taking any irregularities into consideration. This way detecting outliers and level shifts caused by fraudulent activities is possible.

Srinivasan and Kamalakannan (2018) analysed financial data with respect to financial risk. They considered financial risk as a multivariate construct which can be interpreted as a multi-criteria decision-making problem. Therefore, they proposed a multi-objective genetic algorithm (MOGA) as a tool for risk analysis and prediction. The algorithm was tested in terms of predicting decisions on credit card and credit applications. In general, the genetic algorithm is biologically inspired by the evolution process. The process involves the progressive adaptation of biological entities enabling them to survive in a certain environment. To transform this approach into a numerical framework, the algorithm searches the computational space to find the best possible solution for the analysed case. Information included in the dataset is incorporated into a chromosome. The searching rules are formulated on the basis of the chromosome's reproduction mechanisms, such as cross-over, mutation and selection. Only those rules which meet the adopted criteria proceed to the next generation. Srinivasan and Kamalakannan found the best solution by the iteration of the algorithm. The iterations make it possible to find the best solution with an accuracy level exceeding 70%, which renders the performance of MOGA satisfying. What is more, the algorithm outperforms other evolutionary algorithms due to the presence of a memory component which makes the analysis more robust.

3.3. Unsupervised methods

Barabesi et al. (2021) proposed a statistical test on Benford's law (BL), which was based on the sum-invariance property with regard to data entries with the same first significant digits. In general, BL relies on the expected distribution of first digits from the sequence of results. It compares the collected data with the theoretical distribution in order to verify their compliance. Significant differences between the two indicate the occurrence of frauds. The distribution takes into account numbers from 1 to 9 and the fact that their occurrence as the first digit decreases logarithmically. According to this rule, small digits appear more frequently among real-world data than the larger ones. The method is helpful in searching data irregularities in the finance and other sectors. BL might also be applied to further digits – individually or at sequence. The conformity of BL with actual data is verified by using statistical tests that measure goodness-of-fit. The solution proposed by Barabesi et al. was tested in the area of fraud detection in international trade to show its application potential. Firstly, the authors verified the performance of the test on synthetic data. Secondly, the test was applied to real-world data taken from customs declarations of two traders. The final results demonstrated that BL is suitable for labelling fraudulent cases.

Bach, Ćurlin et al. (2020) investigated the relationship between suspicious reports of hours-worked claims and specificity of a project in a project-based company. Therefore, they proposed two data-mining models: one based on chi-square automatic interaction detection (CHAID), aimed at disclosing the relationships between the project attributes and the claims, and the other based on link analysis (LA), whose goal was to detect the potential suspicious claims. CHAID is a decision tree based on chi-square test that allows the split of the dataset into the considered variables. The tree is built progressively from the best to the worst decision rule that differentiates examples. CHAID is usually used in marketing research for customer segmentation, but it generally enables segmentation of any kind. LA, on the other hand, evaluates the relationship between attributes that occur together and are conditioned by each other. This linkage creates a node that implies the association of those two items, and it is called the 'association rule'. Both methods produced similar results that indicated the same areas with the highest probability of fraudulent activity. The research provides a practical tool to detect internal fraud. It allows a better insight into the organisation structure and more efficient control over resource allocation.

Abdul Jabbar and Suharjito (2020) conducted research into fraud in telecommunications company based on call detail records as a dataset. The aim of their study was to propose a method that could detect fraud effectively in order to avoid

financial loss. Two machine-learning algorithms were tested, i.e. *k*-means clustering (*k*-MC), and density-based spatial clustering of applications with noise (DBSCAN), which are similar methods. The difference is that *k*-MC is based on a centroid as a starting point for making clusters, while DBSCAN creates clusters referring to the density of data points. Out of the two methods, only *k*-MC needs a pre-defined number of clusters, which can be optimised. *k*-MC handles large datasets in multidimensional space, but does not work well with outliers. DBSCAN, on the other hand, is less sensitive to outliers, but cannot cope with excessively diverse density, and moreover is applicable only in two-dimensional space. *k*-MC and DBSCAN are both useful in fraud detection by identifying outlying clusters within the analysed dataset. The research outcome showed that both algorithms work well on this kind of data, but *k*-MC performs definitely better when it comes to accuracy, precision and recall. Even though the authors proved that *k*-MC works effectively on data used in the study, this does not necessarily mean that the same method will apply to other cases from the telecommunications industry.

Esen et al. (2019) applied two-step clustering (TSC) to detecting fraud among transactions on a stock market. In this case, fraud was spotted as outlying cases among insider transactions. TSC is a method combining *k*-MC with hierarchical clustering. The algorithm involves two stages: pre-clustering of data into many small sub-clusters, and hierarchical clustering to the expected number of clusters by means of Bayesian Information Criterion (BIC). TSC is more effective in searching outliers than in simple clustering. It seeks to identify examples of unusual behaviour within a peer group, where outliers do not always stand out of the whole population. TSC handles mixed-type data (numerical and categorical). What is more, this method automatically selects the optimal number of clusters, which makes processing extremely large datasets possible. Another advantage of pre-clustering is reducing the size of the distance matrix. The data-mining part of the research was complemented with financial measures that were used to estimate abnormal returns based on abnormal results detected by TSC.

Eshghi and Kargari (2019) proposed a novel technique called the multi-attribute group decision-making method (MCDM), which can be used to detect fraud. The method is a combination of two components: intuitionistic fuzzy sets and evidential reasoning. Fuzzy logic, in contrast to Boolean logic, takes into consideration real numbers in the range of 0-1. Evidential reasoning, on the other hand, refers to inference based on evidence provided by historical data. Eshghi and Kargari's aim was to solve fraud detection problems effectively using real-world data, which tends to be unlabelled. This is the area where other unsupervised methods do not always produce satisfactory outcomes. Another issue is the lack of sufficient information, which hinders the results of fraud detection. This approach reduces the contribution

of expert opinion in favour of information taken from historical data, which eliminates the arbitrariness of decisions. MCDM assigns weights to each attribute based on the provided data only. This method makes it easier to distinguish fraudulent and non-fraudulent cases by taking into account uncertainty as a hesitation margin between these two potential states. Eshghi and Kargari tested their algorithm on bank transaction data. The results showed that MCDM yields satisfactory results with a high level of accuracy and low level of false signals, which makes the fraud detection process based on this method more reliable.

Majadi et al. (2019) investigated an algorithm based on the Markov random field (MRF) called the collusive shill bidding detection (CSBD). They studied its application to identifying fraud based on shill bidding. MRF is a graphical model for inference from noisy data. It is visualised as an undirected graph in which nodes can be in any number of states. There are two types of nodes: the observed nodes and the hidden nodes. The observed nodes are connected with the hidden ones and this relationship is described by the *a priori* belief function. Only hidden nodes are connected with each other, which is described by the compatibility function. The probability of the occurrence of any set of states among the hidden nodes can be calculated using the aforementioned functions. Majadi et al. tested CSBD on synthetically generated auction data charged with shill bidding and on a commercial auction dataset. The algorithm achieved a 99% accuracy in both cases, which renders it an effective tool for colluding shill bidders in online auctions.

Wang and Xu (2018), Zafari and Ekin (2019) as well as Zhang et al. (2022) took up the issue of text mining in the context of fraud detection. Text mining is a part of data mining whose aim is to extract useful information from unstructured data. It is slightly different from traditional data mining due to the nature of the textual data. Semantics presents the main difficulty in text interpretation, which cannot be processed in a computation-based manner. The basis for this approach is the digitalisation of the textual data. The transformed data can then be analysed with quantitative tools.

Wang and Xu (2018) utilised this approach to find fraudulent activity in the automobile insurance industry. Their analysis involved past vehicle insurance claims. The main tool that they used was the latent Dirichlet allocation (LDA2) supported by an AI-driven algorithm. LDA2 was used for feature extraction from the text, which was the initial step of the analysis. The method is based on the decomposition of a text and association of its parts with the main given topics placing them in the context. Wang and Xu combined this method with a deep learning model in order to detect fraudulent behaviours. They used the LDA2 outcome as model input. This approach was compared with two data mining

algorithms: SVM and RF. The comparison demonstrated that LDA2 outperforms SVM and RF while keeping the model quality metrics at the level of 90%.

Zhang et al. (2022) proposed the use of the Bag-of-words (BoW) and Word2Vec (W2V) techniques in financial reports. BoW translates a text into a vector which forms the basis for further analyses. The text is split into single words, which then are digitalised into a vector. Its length is equal to the number of words used, i.e. the size of the dictionary. W2V, on the other hand, is an extension of BoW, which overcomes the acknowledged disadvantages of the initial technique. The length of the W2V vector is limited to a fixed number of words and focuses on the most frequent ones occurring in the text while leaving out the most unusual ones. The analysis was performed on financial reports using both methods. The best results were observed for W2V, whilst BoW demonstrated the highest recall (77%), which is significant in the context of audit work. In conclusion, BoW proves effective in the area of fraud detection.

Zafari and Ekin (2019) investigated a case of prescription fraud by using topic modelling (TM) on the Medicare Part D prescription data (a US prescription drug policy). TM is a statistical model which analyses a text in order to find hidden semantic patterns among words. This approach involves clustering used in text analysis. Related words are assigned to the same topic (cluster) creating a semantic group, which makes the text more interpretable from the computational point of view. TM can be extended to other algorithms, e.g. LDA, which was done by Zafari and Ekin (2019). The outcome of the analysis indicated suspicious prescriptions based on some discrepancies in the distribution of the detected topics. The results can help medical investigators to identify prescription fraud during audits.

Artificial neural networks (ANNs) is one the methods investigated by Ekin et al. (2021). ANNs are biologically inspired computing systems imitating information processing which occurs in real neurones. They are built of at least three layers: the input layer, the hidden layer(s) and the output layer. A single-layer ANN is called a perceptron. This solution can be considered either as supervised, unsupervised or even reinforcement learning depending on the analysed case; this part of the review, however, focuses on unsupervised techniques. The ANN technique is a powerful tool that is commonly used in data analytics. Ekin et al. utilised ANNs for detecting fraud in healthcare. ANNs handle noisy data with great efficiency. They work best on smaller datasets with a low variety of hidden patterns. ANNs are vulnerable to overfitting and imbalanced data, which necessitates data pre-processing.

The issue of imbalanced data, i.e. the low number of cases labelled as fraudulent within the whole dataset, and ways of handling such data was raised by Fiore et al. (2019). They proposed generative artificial networks (GANs) to improve the classification effectiveness in the context of credit card fraud detection. The analysis

was performed on a publicly available dataset concerning credit card fraud. Technically, GANs are multi-layer ANNs composed of two models – generative and discriminative – which compete against each other. The whole model is used to mimic the minority class generating synthetic datapoints. Learning data patterns and generating synthetic data take place all at once with the aim of obtaining examples indistinguishable from the original class as a form of training for the classifier. The generator attempts to cheat the discriminator using its feedback to create new instances similar to the original one. Subsequently, the new examples can be merged with the original ones into an augmented training dataset. The classifier trained on this set outperforms the classifier trained on the original data, thus increasing the efficiency of fraud detection and improving the sensitivity of the classifier.

Bach, Vlahović et al. (2020) investigated the occurrence of fraud in the leasing industry. They performed clustering by using the self-organising map (SOP), also called the Kohonen map or the Kohonen neural network. The research was based on a client database containing leasing contract details. The method involves searching for similarities among datapoints and organising the neurons in the hidden layer into clusters associated with the pattern hidden among the data. The SOP is also an example of competitive learning. Nodes in the neurones compete to represent the pattern and their performance is described with a neighbourhood function. The function decreases according to the distance to the winning node. Nodes with weight which are the closest to the input vector win. Then, their neighbours have their weights updated when moving towards the input pattern. Finally, the displacement of the nodes and their neighbours leads to obtaining clusters whose number corresponds to the grid size. A higher level of accuracy is observed when the number of clusters is below eight. The results were analysed according to the involved categories and their contribution to each cluster. The fraud gradation among clusters proves informative in the context of implementing preventive actions. Cluster characteristics help define customer profiles and predict the potential behaviours and risks among the groups. Experts from the leasing industry confirmed that these results add value to day-to-day business operations and improve planning both at the tactic and strategic levels.

Table. Summary of fraud detection methods

Method	Reference	Source of data	Data category	Purpose
Pre-processing				
SMOTE	Ekin et al. (2021)	Healthcare payment claims	Real-world data	Balancing data
MMWOTE				
RWO				
Random undersampling				
PCA				Dimensionality reduction
Kernel density estimation	Kamalov (2020)	NA	Simulated data	Balancing data
ADASYN				
NearMiss				
Random oversampling				
Feature engineering	Przekop (2020)	New customers' bank application	Real-world data	Dimensionality reduction
Supervised learning				
Linear discriminant analysis	Ekin et al. (2021)	Healthcare payment claims	Real-world data	Classification
Quadratic discriminant analysis				
Decision trees				
Random forest				
	Wang and Xu (2018)	Vehicle insurance claims		
SVM	Zhang et al. (2022)	Financial reports		Classification into two classes and regression
	Wang and Xu (2018)	Vehicle insurance claims		
k-nearest neighbours	Ekin et al. (2021)	Healthcare payment claims		Classification
Naïve Bayes	Zhang et al. (2022)	Financial reports		
	Ekin et al. (2021)	Healthcare payment claims		
Logistic regression + LASSO regularisation	Höppner et al. (2022)	Card transactions		Classification with better performance than w/o regularisation
	Przekop (2020)	New customers' bank applications		
Gradient tree boosting	Höppner et al. (2022)	Credit card transactions		Classification and regression
XGBoost	Farrugia et al. (2020)	Transaction history (Ethereum network)		
Time series	Rousseeuw et al. (2019)	Trade data / airline data		Forecasting
Multi-objective genetic algorithm	Srinivasan and Kamalakannan (2018)	Credit card and credit applications		Classification

Table. Summary of fraud detection methods (cont.)

Method	Reference	Source of data	Data category	Purpose
Unsupervised learning				
Benford's law	Barabesi et al. (2021)	NA	Simulated data	Distribution analysis
		Customs declaration	Real-world data	
CHAID	Bach, Ćurlin et al. (2020)	Working-hours claims and project characteristics		Classification
Link analysis				Association rules
k-means clustering	Abdul Jabbar and Suharjito (2020)	Call detail records (telecommunication)		Classification
DBSCAN				
Two-step clustering	Esen et al. (2019)	Stock market transactions		
Multi-attribute group decision-making method (fuzzy-logic-based)	Eshghi and Kargari (2019)	Bank transactions		
Markov random field (CSBD)	Majadi et al. (2019)	Auctions (synthetically generated)	Simulated data	Clustering
Laten Dirichlet allocation	Wang and Xu (2018)	Vehicle insurance claims	Real-world data	Classification
Word2Vec	Zhang et al. (2022)	Financial reports		Association rules
Bag-of-Words				
Topic modelling	Zafari and Ekin (2019)	Medicare Part D prescription data		Classification and association rules
ANN	Ekin et al. (2021)	Healthcare payment claims		Clustering
GAN	Fiore et al. (2019)	Credit card transactions		
Kohonen neural networks	Bach, Vlahović et al. (2020)	Leasing contracts		

Note. NA – not applicable.
Source: authors' work.

4. Conclusions

The objective of the systematic literature review presented in this paper was to identify and discuss fraud detection methods that are used in finance and related areas. We were looking for methods based on algorithms utilising statistical knowledge in order to delve into and analyse various kinds of data containing examples of fraudulent activity. Finance and related fields indicated the direction of the search, as they handle big data sets and often use analytical solutions. Generalising data to pure numbers without a context, we can see that the tools already adopted in some areas might also be applied to others. The risk of fraud exists not only in the financial sector, but in CTs as well. Large volumes of data generated by CTs would benefit from being subjected to identified algorithms in

order to release the substantial information currently hidden inside them. Our review yielded a set of quantitative methods that were divided into three categories, according to the characteristics of data threatened with fraudulent activity.

The first category comprises pre-processing techniques which are used to prepare imbalanced data for further analysis improving the performance of the applied model. Of course, not all of these methods are sensitive to data imbalance, but the issue must not be neglected. The incidence of fraudulent activities is believed to be usually very low, but its oversight may cause appreciably negative effects. If we ignore the pre-processing, the accuracy of the performed analysis might decrease. Misleading results support wrong decisions that implicate further negative consequences for decision-makers. In the context of CTs, the issue of using unsupervised learning may be a challenge because of a shortage of labelled data. However, it is possible to build a classifier on historical data which could be applied to future studies using data of a similar specificity. Moreover, CTs are embedded in a multivariate space, which allows the reduction of their dimensionality. It can be done not only by means of PCA, but also FE, which needs deeper knowledge about the study to select sensible predictors.

The second category consists of supervised techniques. The key characteristics of these methods is that they need data labelled either as fraudulent or non-fraudulent. However, meeting this condition might be problematic, considering the usual circumstances of committing fraud and the lack of historical data-driven methods for its detection. Therefore, the availability of *a priori* knowledge about which cases are fraudulent and which are not is limited. For this reason, supervised techniques are unlikely to be applicable in CTs, where the knowledge about fraud is usually hidden or unproven. It should be remembered that only supervised algorithms are compatible with the pre-processing ones, as both kinds need knowledge about which data points are fraudulent and which are not. It is then pre-labelled data which makes the application of these kinds of algorithms possible. However, in order to point out fraudulent cases among clinical data, expert knowledge or deep insight into the study specifics and/or a therapeutic area is necessary as well.

The third category, unsupervised techniques, is the largest. These techniques seem to be most useful when only unlabelled data – the most common in CTs – are available. Unsupervised algorithms usually rely on clustering, which leads to obtaining subsets with more homogeneous profiles. The outcome of the application of unsupervised methods enables the indication of cases which might be fraudulent. Basic tools like *k*-MC or Benford's law are a must-have in CTs, but there are also other methods that look promising, such as Markov-random-field-based algorithms or text-mining techniques. The former might be treated as a more sophisticated

improvement to clustering algorithms, while the latter seem to be useful in making clinical reports more time-efficient.

The identified methods should be verified in the context of their utility in CTs. There are other methods already used in this area, but mostly they are simpler algorithms relying on a traditional multivariate analysis (Kirkwood et al., 2013; Venet et al., 2012). What has to be remembered is the fact that the impact of fraud erodes confidence in clinical research, and better-performing, more robust analyses are required to ensure that high-quality medicines are offered on the market. In addition, regulators' requirements are becoming increasingly strict, which obliges drug manufacturers to develop more sophisticated solutions. Therefore, research into fraud detection methodologies for clinical data must continue, and the applicability of the identified methods must be further investigated.

Acknowledgements

A brief summary of a systematic review of the methods from literature was presented as a poster during a conference organised by the Statisticians in the Pharmaceutical Industry (PSI) on 12th–15th June 2022 in Gothenburg. We would like to thank GlaxoSmithKline plc (GSK) for the help extended to Maciej Fronc, especially with regard to his research used in this paper, as well as for the support to his doctorate research at SGH Warsaw School of Economics and everyday work tasks. We would like to specially acknowledge Tim Rolfe of GSK for his ongoing support, encouragement and guidance, as well as his invaluable help in text editing.

Appendix

Keywords used in the query:
(("fraud" OR "manipulation") AND "detection" AND ("machine learning" OR "data mining" OR "artificial intelligence" OR statistic* OR econometric*))

References

Abdul Jabbar, M., & Suharjito. (2020). Fraud detection call detail record using machine learning in telecommunications company. *Advances in Science, Technology and Engineering Systems*, *5*(4), 63–69. https://doi.org/10.25046/aj050409.

Al-Hashedi, K. G., & Magalingam, P. (2021). Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019. *Computer Science Review*, *40*, 1–23. https://doi.org/10.1016/j.cosrev.2021.100402.

Bach, M. P., Ćurlin, T., Dumičić, K., Zoroja, J., & Žmuk, B. (2020). Data mining approach to internal fraud in a project-based organization. *International Journal of Information Systems and Project Management*, *8*(2), 81–101. https://doi.org/10.12821/ijispm080204.

Bach, M. P., Vlahović, N., & Pivar, J. (2020). Fraud Prevention in the Leasing Industry Using the Kohonen Self-Organising Maps. *Organizacija. Journal of Management, Informatics and Human Resources*, *53*(2), 128–145. https://doi.org/10.2478/orga-2020-0009.

Barabesi, L., Cerasa, A., Cerioli, A., & Perrotta, D. (2021). On Characterizations and Tests of Benford's Law. *Journal of the American Statistical Association*. Advance online publication. https://doi.org/10.1080/01621459.2021.1891927.

Botev, Z. I., Grotowski, J. F., & Kroese, D. P. (2010). Kernel density estimation via diffusion. *Annals of Statistics*, *38*(5), 2916–2957. https://doi.org/10.1214/10-AOS799.

Ekin, T., Frigau, L., & Conversano, C. (2021). Healthcare fraud classifiers in practice. *Applied Stochastic Models in Business and Industry*, *37*(6), 1182–1199. https://doi.org/10.1002/asmb.2633.

Esen, M. F., Bilgic, E., & Basdas, U. (2019). How to detect illegal corporate insider trading? A data mining approach for detecting suspicious insider transactions. *Intelligent Systems in Accounting, Finance and Management*, *26*(2), 60–70. https://doi.org/10.1002/isaf.1446.

Eshghi, A., & Kargari, M. (2019). Introducing a new method for the fusion of fraud evidence in banking transactions with regards to uncertainty. *Expert Systems with Applications*, *121*, 382–392. https://doi.org/10.1016/j.eswa.2018.11.039.

Farrugia, S., Ellul, J., & Azzopardi, G. (2020). Detection of illicit accounts over the Ethereum blockchain. *Expert Systems with Applications*, *150*, 1–11. https://doi.org/10.1016/j.eswa.2020.113318.

Federal Trade Commission. (2022, 22 February). *New Data Shows FTC Received 2.8 Million Fraud Reports from Consumers in 2021*. https://www.ftc.gov/news-events/news/press-releases/2022/02/new-data-shows-ftc-received-28-million-fraud-reports-consumers-2021-0.

Fiore, U., De Santis, A., Perla, F., Zanetti, P., & Palmieri, F. (2019). Using generative adversarial networks for improving classification effectiveness in credit card fraud detection. *Information Sciences*, *479*, 448–455. https://doi.org/10.1016/j.ins.2017.12.030.

Gupta, A. (2013). Fraud and misconduct in clinical research: A concern. *Perspectives in Clinical Research*, *4*(2), 144–147. https://doi.org/10.4103/2229-3485.111800.

Höppner, S., Baesens, B., Verbeke, W., & Verdonck, T. (2022). Instance-dependent cost-sensitive learning for detecting transfer fraud. *European Journal of Operational Research*, *297*(1), 291–300. https://doi.org/10.1016/j.ejor.2021.05.028.

International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use. (2016). *Integrated Addendum to ICH E6(R1): Guideline for Good Clinical Practice E6(R2)*. https://database.ich.org/sites/default/files/E6_R2_Addendum.pdf.

Kamalov, F. (2020). Kernel density estimation based sampling for imbalanced class distribution. *Information Sciences*, *512*, 1192–1201. https://doi.org/10.1016/j.ins.2019.10.017.

Kirkwood, A. A., Cox, T., & Hackshaw, A. (2013). Application of methods for central statistical monitoring in clinical trials. *Clinical Trials*, *10*(5), 783–806. https://doi.org/10.1177/1740774513494504.

Majadi, N., Trevathan, J., & Bergmann, N. (2019). Collusive shill bidding detection in online auctions using Markov Random Field. *Electronic Commerce Research and Applications*, *34*, 1–13. https://doi.org/10.1016/j.elerap.2019.100831.

Przekop, D. (2020). Feature Engineering for Anti-Fraud Models Based on Anomaly Detection. *Central European Journal of Economic Modelling and Econometrics*, *12*(3), 301–316. https://doi.org/10.24425/cejeme.2020.134750.

PwC. (2022). *PwC's Global Economic Crime and Fraud Survey 2022*. https://www.pwc.com/gx/en/forensics/gecsm-2022/PwC-Global-Economic-Crime-and-Fraud-Survey-2022.pdf.

Rousseeuw, P., Perrotta, D., Riani, M., & Hubert, M. (2019). Robust Monitoring of Time Series with Application to Fraud Detection. *Econometrics and Statistics*, *9*, 108–121. https://doi.org/10.1016/j.ecosta.2018.05.001.

Sakamoto, J., & Buyse, M. (2016). Fraud in clinical trials: complex problem, simple solutions?. *International Journal of Clinical Oncology*, *21*(1), 13–14. https://doi.org/10.1007/s10147-015-0922-4.

Srinivasan, S., & Kamalakannan, T. (2018). Multi Criteria Decision Making in Financial Risk Management with a Multi-objective Genetic Algorithm. *Computational Economics*, *52*(2), 443–457. https://doi.org/10.1007/s10614-017-9683-7.

Venet, D., Doffagne, E., Burzykowski, T., Beckers, F., Tellier, Y., Genevois-Marlin, E., Becker, U., Bee, V., Wilson, V., Legrand, C., & Buyse, M. (2012). A statistical approach to central monitoring of data quality in clinical trials. *Clinical Trials*, *9*(6), 705–713. https://doi.org/10.1177/1740774512447898.

Wang, Y., & Xu, W. (2018). Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud. *Decision Support Systems*, *105*, 87–95. https://doi.org/10.1016/j.dss.2017.11.001.

West, J., & Bhattacharya, M. (2016). Intelligent financial fraud detection: A comprehensive review. *Computers and Security*, *57*, 47–66. https://doi.org/10.1016/j.cose.2015.09.005.

Zafari, B., & Ekin, T. (2019). Topic modelling for medical prescription fraud and abuse detection. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, *68*(3), 751–769. https://doi.org/10.1111/rssc.12332.

Zhang, Y., Hu, A., Wang, J., & Zhang, Y. (2022). Detection of fraud statement based on word vector: Evidence from financial companies in China. *Finance Research Letters*, *46B*, 1–7. https://doi.org/10.1016/j.frl.2021.102477.

Zhou, W., & Kapoor, G. (2011). Detecting evolutionary financial statement fraud. *Decision Support Systems*, *50*(3), 570–575. https://doi.org/10.1016/j.dss.2010.08.007.