International Journal of Industrial Organization 89 (2023) 102946

Contents lists available at ScienceDirect

International Journal of Industrial Organization

journal homepage: www.elsevier.com/locate/ijio

Flagging cartel participants with deep learning based on convolutional neural networks

Martin Huber[a], David Imhof[b,*]

[a] University of Fribourg, Dept. of Economics, Switzerland
[b] Swiss Competition Commission, University of Fribourg, Dept. of Economics, and Unidistance, David Imhof, Hallwylstrasse 4, 3003 Bern, Switzerland

ARTICLE INFO

Article history:
Received 6 May 2021
Revised 21 November 2022
Accepted 14 March 2023
Available online 5 April 2023

JEL classification:
C21
C45
C52
D22
D40
K40
L40
L41

Keywords:
Bid rigging
Deep learning
Convolutional neural networks
Bid rotation test

ABSTRACT

Adding to the literature on the data-driven detection of bid-rigging cartels, we propose a novel approach based on deep learning (a subfield of artificial intelligence) that flags cartel participants based on their pairwise bidding interactions with other firms. More concisely, we combine a so-called convolutional neural network for image recognition with graphs that in a pairwise manner plot the normalized bids of some reference firm against the normalized bids of any other firms participating in the same tenders as the reference firm. Based on Japanese and Swiss procurement data, we construct such graphs for both collusive and competitive episodes (i.e when a bid-rigging cartel is or is not active) and we use a subset of graphs to train the neural network such that it learns distinguishing collusive from competitive bidding patterns. With the remaining graphs, we test the neural network's out-of-sample performance in correctly classifying collusive and competitive bidding interactions. We obtain a very decent average accuracy of around 95% or slightly higher when either applying the method within Japanese, Swiss, or mixed data (in which Swiss and Japanese graphs are pooled). When using data from one country for training to test the trained model's performance in the other country (i.e. transnationally), predictive performance decreases (likely due to institutional differences in procurement procedures across countries), but often remains satisfactorily high. All in all, the generally quite high accuracy of the convolutional neural network despite being trained in a rather small sample of a few 100 graphs points to a large potential of deep learning approaches for flagging and fighting bid-rigging cartels.

1. Introduction

Bid rigging or collusive tendering is a pervasive problem in many markets and countries (see for instance Feinstein et al., 1985; Porter and Zona, 1993; Baldwin et al., 1997; Porter and Zona, 1999; Pesendorfer, 2000; Banerji and Meenakshi, 2004; Abrantes-Metz et al., 2006; Lee and Hahn, 2002; Bajari and Ye, 2003; Asker, 2010; Ishii, 2009; Abrantes-Metz et al., 2012; Ishii, 2014; Hueschelrath and Veith, 2014; Conley and Decarolis, 2016; Bergman et al., 2019) and damages the taxpayer through substantially increasing the cost of procurement in public tenders. The OECD estimates that

* Corresponding author.
E-mail address: david.imhof@comco.admin.ch (D. Imhof).

https://doi.org/10.1016/j.ijindorg.2023.102946

the elimination of bid rigging might reduce procurement prices in tenders by 20% or more[1] and recommends developing pro-active tools for detecting bid-rigging cartels to supplement the traditional means as leniency programs (see OECD, 2014). Indeed, a growing literature proposes statistical tools for unveiling cartels (see Abrantes-Metz et al., 2006; Bolotova et al., 2008; Harrington, 2007; Abrantes-Metz et al., 2012; Imhof et al., 2018; Imhof, 2019; Chassang et al., 2022, among others papers). Among them is an increasing number of studies applying machine learning for flagging suspicious tenders (see Anysz et al., 2018; Huber and Imhof, 2019; Rabuzin and Modrusan, 2019; Wallimann et al., 2022; Rodriguez et al., 2020; Silveira et al., 2021), which algorithmically learns to optimally predict cartels from historical data. In this paper, we add to this growing arsenal of artificial intelligence-based methods by proposing a novel approach for detecting bid-rigging cartels based on deep learning, which is an extension to machine learning that has revolutionized the quality of image recognition, natural language processing, and many other tasks.

We apply so-called convolutional neural networks (CNNs) - a deep learning method never implemented before in the screening literature - to graphical plots of bid interactions of one firm with all other firms across various tenders in order to learn whether such interactions occur most likely under collusion or competition. Similar to conventional machine learning approaches, CNNs aim at developing optimal predictive models by learning from systematic patterns in data. In contrast to machine learning, however, they do not require predefining specific variables to be used as predictors in the model, but learn autonomously from graphs which features (e.g. specific shapes formed by pixels in images) are the most relevant for recognizing an image. Due to their high performance in such context, CNNs appear to be a natural choice for developing methods for cartel prediction based on graphs of bidding interactions, permitting competition agencies to screen suspicious bids for deeper investigation and to disclose illegal collusive activities, with potentially large gains for tax payers.

We consider procurement data from Switzerland (see Wallimann et al., 2022) and Japan (see Huber et al., 2022), in which we can distinguish collusive from competitive tenders, for training (i.e. developing) the CNN models for predicting bid-rigging cartels as well as for testing, i.e. assessing model performance in terms of prediction accuracy. As model inputs, we use graphs from the bid rotation test suggested by Imhof et al. (2018) and Imhof (2019), which consists of depicting the strategic interaction of firms in the bidding process. To this end, the bids (i.e. prices) of one specific firm are plotted against the bids of several other firms participating in the same tenders within a quadrant that is normalized such that all bids contain values between zero and one. The prevalence of systematic patterns in such plots of pairwise bids permits detecting strategic interactions between bidding firms across tenders. Similar as in the image recognition of photos (e.g. classifying animals as cats or dogs), CNNs can learn from such bid plots to classify specific interactions as likely collusive or competitive even without prior knowledge about the nature of patterns underlying bid rigging.

When assessing the model performance in the Swiss and Japanese data, we obtain an impressive overall accuracy (or correct classification rate) in classifying collusion and competition of roughly 96% and 95%, respectively. The results therefore suggest that our detection method based on CNN models correctly classifies 19 firms out of 20 as cartel participants or competitors. When pooling the graphs from both Switzerland and Japan to generate a transnational dataset of tenders, we obtain a similar accuracy of 95%. We also consider training the CNN in data from one country and using the trained model for predicting collusion in the other country in order to test performance in an institutional context that is different to the training environment. When training the CNN based on Japanese graphs, we obtain an accuracy of 85–86% when predicting collusion in the Swiss data. Training with Swiss bidding plots data and testing in the Japanese data yields to a similar accuracy amounting to 84%.

We also observe that when training in one country to test in the other one, there are noticeable imbalances across the true positive and true negative rates, i.e. the correct classification rates among truly collusive or truly competitive firms. While Japanese competitive and Swiss collusive bid interactions are mostly correctly classified with an accuracy above 99%, the performance is lower for Japanese collusive and Swiss competitive interactions. This result suggests that the bidding patterns in some of the Japanese collusive and Swiss competitive graphs are insufficiently dissimilar for obtaining a high out-of-sample accuracy. Such difficulties in directly transferring trained models from one country to another might be related to institutional differences in the procurement process (e.g. the number of bids typically submitted), as also acknowledged by Huber et al. (2022) when analyzing the Swiss and Japanese data based on a machine learning approach. However, despite such imbalance, the overall correct classification rates of 84–86% appear quite decent given that the training and testing data come from distinct countries. Moreover, the high accuracy obtained when training and testing within a single country (or even a mixed data set of several countries) suggests that the application of CNNs to graphs depicting the bidding interaction of firms provides national competition agencies with a very powerful tool for screening procurement markets.

Moreover, we investigate whether the high accuracy of the CNN is rather driven by the representation of normalized bids based on two-dimensional plots of bidding interactions or by the fact that CNNs can autonomously learn predictive features from such plots in a data-driven way. We find that depicting bidding information in two-dimensional plots considerably increases the predictive performance also when applying machine learning or (logit) regression techniques that (in contrast to CNNs) rely on predefined predictors, such as the share of bid interactions in different quadrants of the plots. In contrast, the additional gain in accuracy of learning predictors autonomously by a CNN rather than predefining predictors (based on quadrants) and applying machine learning is somewhat limited both in the Swiss and Japanese data. However, this finding does not necessarily generalize to other data, in particular if collusive bids show less distinct patterns in terms of

[1] See http://www.oecd.org/competition/cartels/fightingbidriggininpublicprocurement.htm, accessed in March 2021.

their location in the plots. Indeed, we find for bidding data from Italy that the CNN strongly outperforms machine learners relying on the very same bidding plots.

Our paper is related to an expanding literature using screening methods for detecting cartels (see Abrantes-Metz et al., 2006; Harrington, 2007; Bolotova et al., 2008; Abrantes-Metz et al., 2012; Jimenez and Perdiguero, 2012; OECD, 2014; Froeb et al., 2014). More specifically, it considers statistical screens for flagging bid-rigging cartels as Imhof et al. (2018) and Imhof (2019). Huber and Imhof (2019) use such screens (like the coefficient of variation or the kurtosis of the bids in a tender) as inputs to machine learning algorithms for building predictive models for collusion based on Swiss data. Wallimann et al. (2022) focus on the machine learning-based prediction of incomplete (or partial) bid-rigging cartels by calculating a large number of screens based on sub-groups of bids in tenders. Huber et al. (2022) apply machine learning techniques to a Japanese bid-rigging cartel and also investigate the transnational predictive performance across Swiss and Japanese data (as we also do for deep learning in this paper). All these studies focus on a tender-based approach for flagging bid-rigging cartels, i.e. they analyze the bid distributions within tenders. In this paper, we, however, focus on a firm-based approach that analyzes the bid interactions of firms in a specific region and period across tenders to test if their interactions are of collusive nature.

Furthermore, Chassang et al. (2022) investigate the occurrence of "gaps" in bid values (i.e. bid values with zero probabilities) in the distribution of observed bids. Finding such gaps, especially between the first and the second lowest bid in a tender, might point to collusive behavior as they should not systematically occur in competitive markets. Related to our paper, this phenomenon of missing density implies that some areas of our plots remain empty when firms collude, which may be recognized as predictive feature when training a CNN for detecting collusion. However, when calculating the normalized margin suggested by Chassang et al. (2022), we find no particular gap for the Okinawa bid-rigging cartel, whereas the normalized distance, a statistical screen used in Huber et al. (2022) clearly shows manipulation in the bidding pattern. These results suggest that exclusively relying on a "gap" between the first and the second lowest bids might not be sufficient for detecting bid rigging. For this reason, one should consider both the differences between the first and second lowest bids as well as between the losing bids. In fact, the differences between the two lowest bids relative to the differences among losing bids importantly predict bid-rigging cartels in our Japanese data.

Our paper is also related to further studies on detecting bid-rigging cartels which apply econometric tests, as suggested in the seminal paper of Bajari and Ye (2003). A first test consists of verifying if bids are independent conditional on the costs of each firm by estimating a bidding function and testing if the residuals are correlated between firms, indicating potential collusive issues. A second test examines if firms react similarly as a function of their respective costs, based on the estimated cost-related coefficients in a regression model. Divergent coefficients across firms could indicate potential collusive issues. Subsequent papers replicate and refine the two econometric tests suggested by Bajari and Ye (2003) (see Jakobsson, 2007; Aryal and Gabrielli, 2013; Chotibhongs and Arditi, 2012a; 2012b; Conley and Decarolis, 2016; Imhof, 2017; Bergman et al., 2019). The application of the econometric tests in the Okinawa dataset is difficult, at least for two reasons. We first do not observe an essential part of the market, which impedes the construction of a meaningful proxy for the capacity of each firm. Second, many firms bid few times in the data, making it hard if not impossible to estimate panel models with variable coefficients in order to implement the tests, even for the most frequent bidders. In contrast, the Ticino dataset is characterized by few firms that bid many times, which facilitates the application of those econometric tests. However, Imhof (2017) finds that the tests fail to detect pairs of firms that actually colluded. Moreover, the tests require firm-specific covariates which are good proxies for the respective firm's costs and the measurement or construction of such covariates might create additional administrative burden. Finally, our paper is also broadly associated with further studies investigating bid-rigging cartels or rings (see Baldwin et al., 1997; Porter and Zona, 1999; Banerji and Meenakshi, 2004; Lee and Hahn, 2002; Asker, 2010; Ishii, 2009; 2014).

The remainder of this study is organized as follows. Section 2 introduces the bid rotation screen that underlies our analysis and illustrates how we compute graphs for the bidding interactions of firms. Section 3 introduces the Japanese and Swiss data considered in the empirical analysis. Section 4 provides a brief introduction to deep learning based on CNNs and outlines our CNN architecture for detecting cartels. Section 5 presents the empirical results and Section 6 discusses the application of other detection methods. Section 7 concludes.

2. The bid rotation screen

The inputs for predicting bid-rigging cartels in our empirical analysis consist of graphs depicting the bidding interaction between one firm and a group of other firms in a series of tenders during a specific period. We follow the ideas in Imhof et al. (2018) and Imhof (2019) and use the so-called bid rotation screen to construct the graphs entering our algorithm. In order to plot the interaction of firms for several tenders of different contract values, we apply the so-called min-max transformation to some bid i within some tender t as follows:

$$\hat{b}_{it} = \frac{b_{it} - b_{min,t}}{b_{max,t} - b_{min,t}}, \tag{1}$$

where $b_{min,t}$ and $b_{max,t}$ are the minimum and maximum bids in tender t, respectively. Therefore, each transformed bid takes values between zero and one across all tenders, $\hat{b}_{it} \in [0, 1]$. Such a normalization based on the min-max transformation in the space $[0, 1]$ allows us to compare tenders of different contract values, in order to obtain a scale that remains constant

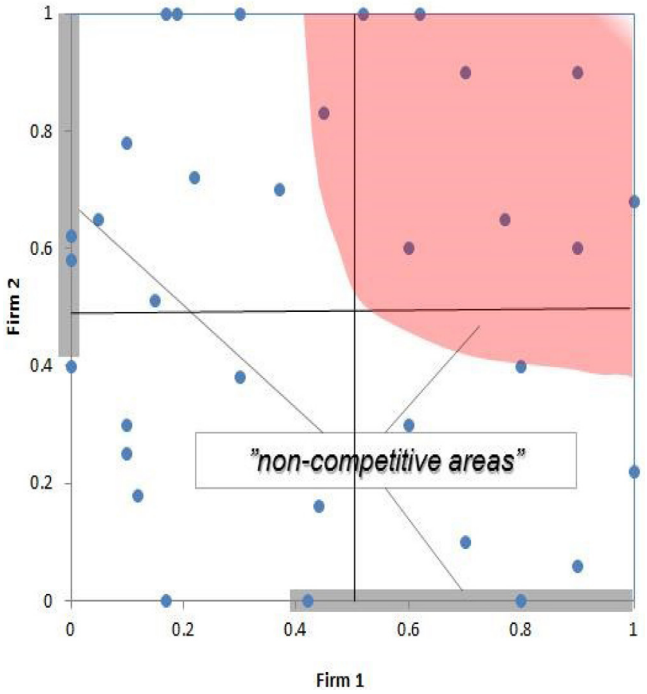

Fig. 1. Example of non-competitive interactions between two firms.

(i.e. between zero and one) across tenders and bids when examining the distribution of bids. In a next step, we calculate the Cartesian coordinates of transformed bids for each possible pair of firms, i.e. of some reference firm and each of the other firms participating in the same tenders as the reference firm in a specific period, on the space $[0, 1] \times [0, 1]$. We then plot the Cartesian coordinates of these pairs of firms in a graph, thus depicting the interactions of the reference firm (on the x-axis) with a group of firms (on the y-axis) bidding in the same tenders as the reference firm.

We construct the graphs separately within collusive periods (with a cartel in place) and competitive periods (after a cartel has been dissolved) for each reference firm, in order to depict the firm's bidding interactions both within cartels and under (post-cartel) competition. Our suggested method exploits systematic differences in the bidding interaction under collusion and competition, but remains agnostic about how these differences look like. It simply relies on the plausible hypothesis that the bidding behavior of firms differs between the cartel and post-cartel periods in some way due to strategic behavior and aims at detecting such differences algorithmically by autonomously learning from data.

Even if the existence of some (and possibly unknown) kind of difference in distributional patterns of bids across collusive and competitive periods is sufficient for applying our method, several studies provide explanations for how bid rigging might change interactions, e.g. by inducing gaps or concentrations in the distribution of bids (see Huber and Imhof, 2019; Imhof, 2019, for examples in Switzerland) and (see Chassang et al., 2022; Huber et al., 2022, for examples in Japan). In Swiss bid-rigging cases, for instance, differences between the two lowest bids in a tender increase whereas the differences between losing bids simultaneously decrease. The two effects combined distort the distribution of bids, which might be captured by tender-specific statistical screens. Also in the Japanese data, bid rigging creates gaps between the first and second lowest bids in the distribution of the bids, whereas losing bids are close to each other, as revealed by screens considered in Huber et al. (2022).

Furthermore, Chassang et al. (2022) formalized a model implying that increasing differences between the first and second lowest bids are not in line with competition. Such a zero density in bid values between the first and second lowest bids creates a gap in the distribution of bids. In the context of our min-max transformation, a gap in the distribution of bids (or a zero density) implies that collusive bids are absent in specific regions of the two dimensional space $[0, 1] \times [0, 1]$, whereas competitive bids are typically scattered all over the space. Considering the interaction of just two firms across different tenders, Fig. 1, which is taken from Imhof et al. (2018) and Imhof (2019), highlights regions (non-competitive areas) in the space in which we would suspect a more concentrated density of bids in the case of bid rigging. This is either due to one firm submitting a comparably high bid to guarantee that the other, lower-bidding firm wins the contract, or due to both firms submitting high bids to allocate the contract to another cartel member. Such distributional patterns would therefore indicate strategic interactions between firms and therefore point to collusive behavior. Therefore, a concentration of bids in specific regions and density gaps in other regions can be learnt by a CNN as a feature for predicting collusion and competition.

3. Data of bid-rigging cartels

3.1. Japanese data from Okinawa

Our Japanese data stem from a bid-rigging cartel in Okinawa. We refer to Ishii (2014) and Huber et al. (2022) for further details concerning the institutional context and data construction and note that we consider all firms that submitted at least ten bids in the data window for our deep learning approach. We focus on contracts of the so-called ranks A+ and A in the Japanese data, with ranks depending on the reserve prices of the contracts as discussed in Huber et al. (2022). The latter study found that the impact of the bid-rigging investigation launched by the Japanese Federal Trade Commission (JFTC) in June 2005 and the sanctions imposed on cartel participants in March 2006 were strongest for contracts of those ranks, in which condemned bidders had participated more frequently. This implies that for such contract types, it was particularly unlikely that former cartel participants immediately colluded again in post-cartel periods after sanctioning, which thus serve as competitive observations in our data.

As in Huber et al. (2022), the pre-investigation period from April 2003 to June 2005 presumably comprises the collusive tenders and the post-amendment period from January 2006 to May 2007 (after sanctioning) the competitive cases.[2] Concerning rank A+ contracts, we consider 48 tenders with 670 bids in the pre-investigation period and 57 tenders with 1049 bids in the post-amendment period. For contracts of rank A, we consider 108 contracts with 1325 bids in the pre-investigation period and 86 contracts with 1439 bids in the post-amendment period. All in all, 161 firms bid for A+ contracts, 491 firms for A contracts, see Table 1.

Table 1
Japanese sample.

Type of contract	Period	Number of bids	Number of firms	Number of contracts
Rank A+	Collusive	1049	161	57
Rank A+	Competitive	670	159	48
Rank A	Collusive	1339	491	108
Rank A	Competitive	1439	481	86

We only calculate interaction graphs for those reference firms submitting at least 10 bids for contracts of rank A+ and A during the data window, but consider all of the submitted bids in the relevant tenders to compute pairwise bidding interactions. This implies that even a firm who only participated in one tender in which the reference firm was present is considered in one of the pairwise interactions in the graph of that tender. As an example, Fig. 2 provides the collusive and competitive graphs for contracts A+ computed for that firm that placed the most bids in our data window. The left graph depicts the strategic interaction of that firm during the pre-investigation period when the cartel was active. The dotted rectangles in Fig. 2 indicate the pairwise interactions located on the axes. On the horizontal (or x-)axis, the pairwise observations in the dotted rectangle correspond to tenders in which the reference firm submits a relatively high bid where another firm is designated to win the contract due to having the lowest bid. On the vertical (or y-)axis, the pairwise observations in the dotted rectangle depict all interactions in which the reference firm submits the lowest bid to win the contract, whereas the other bidders submit higher bids. If the reference firm wins several contracts in the period for which the graph is produced, the vertical axis depicts all pairwise interactions across multiple tenders won by the reference firm. Conversely, aligned vertical observations in the interior of the plot (i.e. mostly away from the axes) as those in the dashed rectangle typically correspond to the pairwise interactions of firms bidding in the same tender. As expected, we find the bottom-left area to be empty, while we observe more pairwise interactions in the top-right region (likely supporting a cartel member designated to win the contract), in the upper area close to the y-axis, or in the right area close to the x-axis (such that one firm likely supports the other in that pair). The right graph shows the interactions of the reference firm during the post-amendment period and in line with competition, pairwise interactions also enter the lower-left area. All in all, we obtain 143 collusive and 144 competitive graphs from the Japanese data.

3.2. Swiss data

Concerning Switzerland, we use data from the canton of Ticino that consist of both collusive and competitive periods (see Imhof, 2019) and observations from two other regions as competitive periods (see Wallimann et al., 2022). Compared to the Japanese data, the data window is much longer for the Ticino data, ranging from January 1999 to March 2005 for the cartel period. As reported in Table 2, we consider 1331 bids in 183 contracts for the collusive period. However, the number of bidding firms is considerably lower than in the Japanese data, as we only observe 21 cartel participants, see Table 2.

[2] We do not consider contracts of so-called ranks B and C in which condemned cartel members participated less frequently such that the number of interactions for constructing the graphs would have been too low. Furthermore, findings in Huber et al. (2022) suggest that there might be collusive interaction for these kind of contracts even after the sanctions by the JFTC such that including these post-amendment observations could contaminate our sample of competitive interactions and thus, reduce the predictive performance of our method.

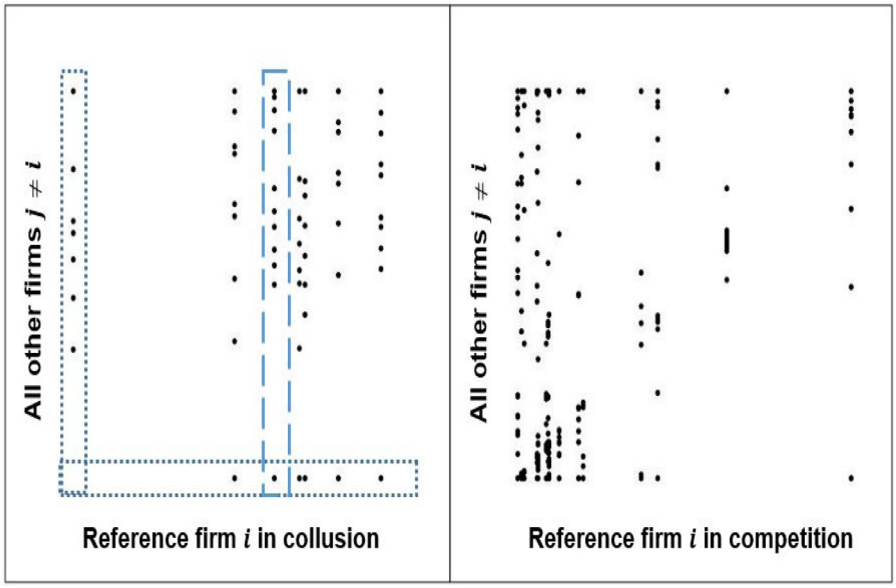

Fig. 2. Interactions of a Japanese firm with other bidders (left: cartel period; right: competitive period).

Table 2
Swiss sample.

Data	Period	Number of bids	Number of firms	Number of contracts
Ticino	Collusive	1331	21	183
Ticino	Competitive	297	21	40
Other Swiss regions	Competitive	3728	188	1018

We compute the graphs on a yearly base for those 16 firms who participate most in the tenders. In the much shorter competitive period from April 2005 to the end of 2006, we can exploit 297 bids in 40 tenders and compute graphs for the same 16 firms over that period (i.e. without dividing the competitive tenders into subperiods).

Fig. 3 reports bidding interactions of that firm that most frequently bid in the Ticino data. The left graph depicts the first year of the collusive period in 1999 and the interpretation of the dash and dot rectangles is identical to Fig. 2. Similar as for

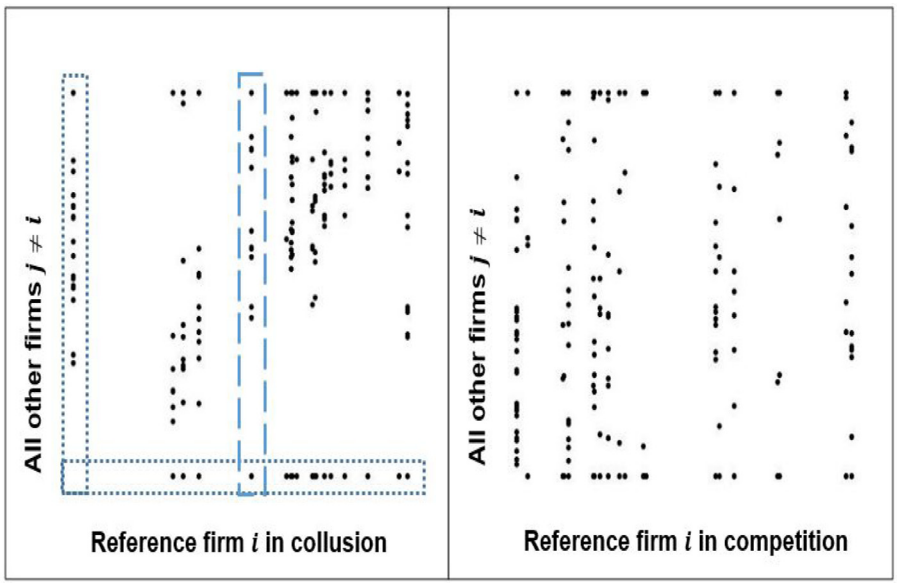

Fig. 3. Interactions of a Swiss firm with other bidders (left: cartel period; right: competitive period).

Table 3
Summary statistics for collusive and competitive normalized bids by country.

normalized bids	mean	lower quartile	median	upper quartile
Swiss collusive	0.669	0.534	0.680	0.821
Swiss competitive	0.453	0.236	0.432	0.661
Okinawa collusive	0.621	0.462	0.667	0.824
Okinawa competitive	0.324	0.058	0.212	0.563

the Japanese data, the bottom-left area is empty in contrast to the top-right region. In the right graph, which provides the interactions in the competitive period, interactions however occur in the bottom-left part, too, such that no apparent gaps in the values of pairwise bids of two firms occur. Fig. 3 therefore suggests that the reference firm behaved differently in the year 1999 when compared to the competitive period.

Since the number of graphs is unsatisfactory and our Ticino sample is imbalanced with many more collusive than competitive graphs, we also consider data from competitive periods in other Swiss regions to obtain a more balanced sample (see Wallimann et al., 2022, for details on the cases). We to this end create graphs on a yearly base that depict the interactions of former cartel participants in competitive periods (after the cartel has been dismantled), based on 3728 bids in 1018 tenders. Combining these data with the Ticino case yields all in all 144 graphs for competitive periods and 96 for collusive periods in various Swiss regions.

3.3. Summary statistics on normalized bids across countries

In Table 3, we provide summary statistics on the distribution of normalized bids of specific firms (rather than pairs of firms as in our previous plots), separately for Japan and Switzerland as well as competitive and collusive periods. We observe that in the presence of a cartel, the distribution of normalized bids is quite similar across countries, as indicated by the comparable values in the mean, median, lower quartile and upper quartile of the respective normalized bids under collusion.[3] This contrasts with the distribution of normalized bids under competition, for which the summary statistics are rather different across countries. Yet, we find that within either country, there are important differences in the distribution of normalized bids across collusion and competition. For both Japan and Switzerland, bid rigging induces a shift in the distribution of normalized bids toward upper segments. For our plots, this implies that collusive observations are predominantly concentrated in non-competitive regions, while the region close to the origin is more likely to be empty.

4. A convolutional neural network for detecting collusive firms

Convolutional neural networks (CNNs), as for instance discussed in LeCun et al. (1998) and Ciresan et al. (2011) are a powerful and increasingly popular deep learning approach for image (e.g. facial) recognition. By their nature, they can learn any kind of patterns in plots of bidding interactions that are predictive for bid rigging from a large enough training data set with collusive and competitive cases, even if the shape of such patterns is a priori unknown to the researcher. This distinguishes such a deep learning approach from previously applied machine learning methods that require so-called feature engineering, namely pre-specifying the features to be used as predictors, such as specific statistical screens like the coefficient of variation or the kurtosis. However, CNNs require choosing a range of model tuning (or hyper-)parameters that determine how the model extracts and learns the most predictive patterns and it needs to be pointed out that the choice of these parameters may crucially affect the performance of the method.

CNNs can be regarded as an extension of standard feed-forward neural networks (FNN), see e.g. McCulloch and Pitts (1943) and Ripley (1996). In the context of bid rigging, such FNNs aim at fitting a system of nonlinear functions that flexibly models the statistical association of a range of predictors (like statistical screens) and an outcome of interest (like a binary indicator for collusion). Specifically, the predictors serve as inputs for specific nonlinear intermediate regression functions (e.g. sigmoid or rectifier functions), also called hidden nodes, which themselves serve as inputs for modeling the outcome of interest. The coefficients in the various nonlinear regression functions are estimated such that they minimize a specific loss function defined in terms of errors (or residuals) when predicting the outcome of interest in the training data, like e.g. the sum of squared residuals. As our collusion outcome is binary, we minimize the so-called binary cross-entropy, which corresponds to a maximum likelihood-based estimation of the coefficients based on the log-likelihood function for binomial outcomes, as e.g. used in logit or probit regression.

It is worth noting that the hidden nodes in an FNN bear some similarity with the baseline functions in spline or series regression, with the difference that they are learned from the data rather than predetermined. The hidden nodes are also related to principal component analysis (PCA), which aims at generating dimension-reducing linear (rather than nonlinear)

[3] We note that for computing the summary statistics, we delete any zeros and ones and thus focus on the internal distribution of normalized bids between the lowest and highest bids submitted in some tender.

Table 4
CNN architecture for predicting collusion.

1. Convolutional layer (two-dimensional for gray-shaded images) with 8 filters, each with a sliding window size 3 × 3 pixels, using the rectifier (or ReLU) function, which is zero for negative values and linear for positive values, as nonlinear function for mapping pixels into numeric values.
2. Max pooling with size 2 × 2: aggregates the previously generated numeric features within sliding matrix windows of size 2 × 2 by taking the maximum value within each window.
3. Convolutional layer with 16 filters, each with kernel size 3 × 3, using the rectifier as nonlinear function for mapping the pooled features.
4. Max pooling with size 2 × 2.
5. Convolutional layer with 32 filters, each with kernel size 3 × 3, using the rectifier as nonlinear function for mapping the pooled features.
6. Flattening: transforming each of the matrix values in the previous convolutional layer into separate input (or predictor) variables for the FNN to follow.
7. Hidden layer with 128 hidden nodes (i.e. 128 nonlinear regression functions), using the rectifier as nonlinear function for mapping the inputs.
8. Hidden layer with 64 hidden nodes, using the rectifier as nonlinear function for mapping the nodes of the previous hidden layer.
9. Hidden layer with 32 hidden nodes, using the rectifier as nonlinear function for mapping the nodes of the previous hidden layer.
10. Output layer for classifying the binary outcome (1=cartel, 0=competition), using the sigmoid function as nonlinear function for mapping the nodes of the last hidden layer in order to predict the outcome.

functions of the predictors. Indeed, when replacing the nonlinear functions by linear ones, FNNs collapse to a linear regression model underlying a PCA. Depending on the model complexity of an FNN, hidden nodes may affect the outcome either directly or through other hidden nodes, such that several layers of hidden nodes allow modeling interactions between the functions. The number of hidden nodes and layers gauges the flexibility of the model, with more parameters reducing the bias but increasing the variance.

In contrast to FNNs, CNNs do not rely on providing a list of predictors but may autonomously learn to derive relevant features from objects like images that do a priori not have a clear data structure. CNNs make use of the fact that an image may be considered as a matrix of pixel values which e.g. correspond to specific shades of gray in a grayscale picture. The convolution step in a CNN consists of sliding over submatrices of pixels in the image with a pre-defined dimension (e.g. 3 × 3 pixels) and multiplying each submatrix element-wise with a so-called filter (or kernel), yet another matrix with the same dimension. The aim of such filters, which may e.g. consist of a combination of ones and zeros, is to map the submatrices of pixels into numeric values. This permits representing specific spatial patterns (such as edges in the original picture) by numeric features, depending on the kind of filter that is applied.

A CNN actually learns the values that the filters shall take to optimize performance in a data-driven way during the training process. However, the researcher must pre-specify the number and the size of the filters as tuning parameters. The generation of numeric features through filtering is typically followed by a so-called pooling step, which aggregates the numeric features, e.g. by taking average or maximum values over a pre-specified number of adjacent features. This is typically followed by further filtering (of the pooled features) and pooling steps for transforming and aggregating the features, which constitute further tuning parameters. Finally, the refined features are used as inputs in a conventional FNN for prediction.

While CNNs do not require direct feature engineering as FNNs or classical machine learning, they obviously rely on making a range of design choices concerning how the important features should be learned by (possibly nested) filtering and pooling. In addition, one is confronted with the same design choices in FNNs concerning the number of hidden nodes and layers as well as further tuning parameters. The latter include the share of the training data, i.e. the part of the total sample used to develop a predictive model, and the validation data for assessing its predictive performance based on the accuracy (or correct classification rate) outside the training sample during the process of refining the model. A further tuning parameters is the batch size, the number of observations in the training data to work through before updating the model parameters (in particular the coefficients in the non-linear models) when aiming at minimizing the classification error in the training data based on so-called stochastic gradient descent. Finally, one needs to specify the number of epochs, i.e. of times that the algorithm works through the entire training data for estimating the model parameters. A too small number of epochs may prevent the algorithm to accurately learn the predictive model in the data, thus entailing biases due to misspecification. In contrast, a too large number of epochs may overfit the model parameters to the specificities of the observations in the training data at hand, thus reducing model generality and increasing the variance.

Due to this large amount of choices to make that may apparently come with little guidance, tuning deep learning models is frequently regarded to be more an art than a science. Also in our empirical application, we considered several different specifications in terms of numbers of hidden layers and nodes, filtering and pooling steps, as well as batch size and epochs. We investigated the accuracy of these alternative specifications in the validation data when drawing a random 75% subset of the Japanese sample and using 90% of that subset as training data and 10% as validation data. More concisely, we compared the accuracy of each network specification in the validation data of the last epoch of the training process, as reported in Appendix A.3. In Section 5, we, however, only present the empirical results for the CNN specification outlined in Table 4, which performed better in the validation data of the last epoch than any of the other network specifications considered. In this model, we set the share of training and validation data per epoch to 90% and 10%, respectively, the batch size to 16, and

Table 5
Summary statistics for accuracy when training and testing in the Japanese sample.

	Minimum	1st quartile	Median	Mean	3rd quartile	Maximum	Observations
All graphs	0.901	0.930	0.958	0.950	0.961	1.000	287
Collusion	0.900	0.944	0.973	0.966	1.000	1.000	143
Competition	0.865	0.914	0.932	0.937	1.000	1.000	144

Note: Mean precision is 0.938

the number of epochs to 40.[4] Other than this we use the default options of the "keras" package by Falbel et al. (2021) for the statistical software "R", an interface to the "Keras" application programming interface (API) for neural networks, which itself makes use of the open source software library "TensorFlow" for implementing deep learning architectures.

We point out that the resolution of the images fed into the CNN was found to be a further parameter affecting predictive performance in the validation data. For instance, using an image resolution of just 48 times 48 pixels entails an accuracy that is generally several percentage points lower than under 96 times 96 pixels, which is the resolution chosen in our application. Increasing the resolution even more (e.g. to 144 times 144 pixels) does, however, not appear to further improve accuracy. The resolution affects the complexity of parameters in (and the accuracy of) a CNN because a higher resolution entails a larger matrix of pixels to which the initial filtering steps are to be applied (by sliding over submatrices of pixels). Higher resolutions may thus reduce the bias by displaying bidding patterns more concisely, but on the other hand increase the variance by raising the input space in terms of the number of pixels, which suggests a bias-variance trade-off in a similar manner as for other tuning parameters.

As indicated before, we apply our approach of training and validating a CNN model only to a random subset of the data, while we hold out the remaining observations for testing our model in unseen data (not used for the iterative training and validation steps) to check out-of-sample predictive performance, using accuracy (or the correct classification rate) as performance measure. For instance, when only considering Japanese data, we randomly choose 75% of the images from the Okinawa region for the training and validation process and 25% for testing performance. We follow the same approach when only investigating Swiss data or when analyzing a mixed (i.e. transnational) sample consisting of both Japanese and Swiss images. For each of the approaches considered, we repeat the procedure of randomly splitting the sample into training/validation and test data to develop and assess the predictive model 20 times. We report the mean, median, worst, and best performance across these 20 simulations, as the variance in performance is non-negligible across simulations due to the limited number of images available. Finally, we also use images from one country for training and validation to test performance based on images of the respective other country in order to check the transnational transferability of trained CNNs. We consider 20 simulations for this approach, too, even though the training/validation and test data are deterministic (in countries), while there is still a stochastic component within training and validation of the model.

5. Empirical results

We first apply our CNN architecture outlined in Section 4 to our 287 Japanese interaction graphs constructed from the Okinawa data (see Huber et al., 2022), running 20 simulations in which we randomly select 75% of images for training and validation and the remaining 25% for testing performance based on classification accuracy. Our outcome variable is a binary indicator taking the value 1 for a cartel period and 0 for competition. Table 5 provides summary statistics for the accuracy in the total test data and separately for truly collusive or competitive test periods (i.e. the true positive and true negative rates) across simulations, namely the respective minimum, first quartile, median, mean, third quartile, and maximum. Despite the rather limited number of images available for training, the median and mean of accuracy amount to 95.8% and 95%, respectively, in the test data. This suggests that the CNN classifies on average roughly 19 out of 20 firms correctly as cartel members or competitive bidders.

When investigating the accuracy within actually collusive and competitive cases, we find that the CNN performs somewhat better for detecting actual cartels, with an average and median of the true positive rate (or recall) of 96.6% and 97.3%, respectively, than for detecting truly competitive firms, with a median and mean of the true negative rate of 93.2% and 93.7%, respectively. Put differently, we find median and mean classification error rates of 6.8% and 6.3%, respectively, under competition and of 2.7% and 3.4%, respectively, under collusion. Despite these slight imbalances across outcome classes, the CNN shows a quite decent overall predictive performance for both collusive and competitive cases. We, however, point out that the variation is non-negligible across our 20 simulations, as can be seen by the differences in the respective minimum and maximum accuracy. This suggests that one should ideally use a substantially larger pool of images for training and testing the model than was available to us. In a note underneath Table 5, we also provide the mean precision across simulations, i.e. the average share of truly collusive cases among all cases classified as collusive, which amounts to 93.8%.

In a second step, apply the same procedure to our 240 Swiss interaction graphs and report the results in Table 6. Overall accuracy is again quite high, with its median and mean amounting to 95% and 95.7%, respectively. This is quite similar

[4] In Appendix A.3, we report the accuracy of our model in the validation data of the last epoch when varying the number of epochs between 20, 40, and 60 while setting the batch size to 16 and when varying the batch size between 4, 16, and 64 while setting the number of epochs to 40. Model performance turns out to be very stable across these choices of tuning parameters.

Table 6
Summary statistics for accuracy when training and testing in the Swiss sample.

	Minimum	1st quartile	Median	Mean	3rd quartile	Maximum	Observations
All graphs	0.933	0.933	0.950	0.957	0.971	1.000	240
Collusion	0.778	0.918	0.957	0.937	0.962	1.000	96
Competition	0.880	0.946	0.972	0.966	1.000	1.000	144

Note: Mean precision is 0.948

Table 7
Summary statistics for accuracy when training and testing in both countries.

	Minimum	1st quartile	Median	Mean	3rd quartile	Maximum	Observations
All graphs	0.863	0.939	0.950	0.949	0.962	0.977	527
Collusion	0.710	0.916	0.937	0.934	0.972	1.000	239
Competition	0.912	0.944	0.969	0.962	0.973	1.000	288

Note: Mean precision is 0.953

Table 8
Summary statistics for accuracy when training in Japan and testing in Switzerland.

	Minimum	1st quartile	Median	Mean	3rd quartile	Maximum	Observations
All graphs	0.729	0.832	0.863	0.849	0.879	0.921	527
Collusion	0.958	0.979	0.990	0.986	0.990	1.000	239
Competition	0.563	0.731	0.781	0.758	0.806	0.882	288

Note: Mean precision is 0.731

as for the Japanese data and suggests that our deep learning approach might work well in different countries. Contrary to the Japanese cases, however, the CNN now entails a larger true negative rate (of correctly classifying truly competitive firms) with a mean and median of 96.6% and 97.2%, respectively, while the respective numbers are 93.7% and 95.7% for the true positive rate (of correctly classifying truly competitive firms). If the overall performance appears very satisfactory (with roughly 19 out of 20 correct classifications), there is again some variation in the predictive performance across the 20 simulations.

Next, we pool Japanese and Swiss graphs to obtain a mixed sample of 527 images in total and conduct the same simulation, training, validation, and testing steps as before. As reported in Table 7, the median and mean accuracy amount to 95% and 94.9%, respectively, which is rather similar as for the analyses within countries. This finding can be explained by the similarity of the distribution of the normalized bids across both countries, especially for the collusive bids. Our results therefore suggest that well-performing CNNs might under certain conditions even be trained in mixed data composed of several countries or regions, which can be helpful for attaining larger samples. Furthermore, the use of combined data might be advantageous for reducing the risk of overfitting, i.e. fitting the model too much to the particularities of one specific country, in particular when predicting collusion for a region where no training data are available.

We now turn to the more challenging task of training the predictive model in one country and testing it in the other country. As discussed in more detail in Huber et al. (2022) for the case of Japan and Switzerland, the transfer of a trained method across borders may not yield the desired performance due to institutional differences in procurement, which may jeopardize the method's ability to distinguish collusion and competition in a country-specific context it has not been trained on. Table 8 reports the summary statistics for accuracy when training the Japanese graphs for testing on the Swiss graphs. While the Swiss test data are now identical across our 20 simulations, the Japanese images randomly chosen for training and validation generally differ. The median and mean accuracy attains 86.3% and 84.9%, respectively, is somewhat lower than before but still appears quite satisfactory. However, the imbalances across outcome classes are now more substantial. While the median and the mean of the true positive rate (among cartels in the Swiss test data) amount to 99% and 98.6%, respectively, the corresponding figures for the true negative rate (among competing firms in the Swiss test data) are 78.1% and 75.8%, respectively. Apparently, bidding interactions of Swiss firms under competition do not always yield a pattern that is close to that of Japanese firms under competition that were used for training the algorithm, while interactions look substantially more similar across Swiss and Japanese cartels. Yet, for neither outcome class, the predictive performance seems unacceptably low and overall accuracy is in fact quite good.

Finally, we swap the roles of the countries and train the CNN based on the Swiss graphs to test it on the Japanese graphs. As shown in Table 9, accuracy in all of the test sample is now lower with a median and mean of 84.2% and 84.3%, respectively, but nevertheless appears acceptable. However (and even more so than in the previous analysis), there is a sizeable imbalance in the performance across classes. While the correct negative rate is very high with its median and its mean amounting to 99% and 98.5%, respectively, the correct positive rate is substantially lower, with a median and mean of just 69.2% and 69.9%, respectively. This suggests that several Japanese collusive graphs are hard to distinguish from Swiss competitive graphs used for training, while Japanese competitive graphs can be very clearly distinguished from

Table 9
Summary statistics for accuracy when training in Switzerland and testing in Japan.

	Minimum	1st quartile	Median	Mean	3rd quartile	Maximum	Observations
All graphs	0.791	0.836	0.842	0.843	0.851	0.902	527
Collusion	0.587	0.678	0.692	0.699	0.722	0.846	239
Competition	0.958	0.979	0.990	0.985	0.993	1.000	288

Note: Mean precision is 0.979

Swiss collusive graphs. Overall, we conclude that the CNN architecture yields a very promising predictive performance for classifying collusive and competitive bidding interactions of firms. Aside from a lower accuracy for specific outcome classes when only training in one country to test in the other one, the deep learning approach performs generally very well (with accuracies typically around 95% or slightly higher) when training and testing within country-specific data or a mixed sample consisting of pooled graphs from both Japan and Switzerland.

6. Discussion of alternative methods for detecting bid rigging

Our suggested approach based on CNNs reaches a high accuracy of 95% on average for both our Swiss and Japanese bid-rigging cases. In other words, we are able to correctly classify 19 firms out of 20 as either "cartel participants" or "competitors" based on deep learning. Nevertheless, it appears interesting to discuss alternative screening methods for better understanding the differences to our deep learning architecture. In the following, we describe three different approaches, which appear popular for detecting bid-rigging cartels. The first one is the game-theoretical approach of Bajari and Ye (2003), the second the method of Chassang et al. (2022) based on missing densities (or density gaps) in bids. The third approach combines machine learning with statistical screens for detecting collusion (see Huber and Imhof, 2019; Wallimann et al., 2022; Imhof and Wallimann, 2021). Furthermore, we investigate if the higher accuracy reached by CNNs is due to plotting information on two dimensions or to the autonomous learning process inherent to deep learning. Finally, we discuss policy recommendations and the looked-for features of a detection method.

6.1. Game-theoretical approach

In their seminal paper, Bajari and Ye (2003) formalize a procurement auction model with asymmetrical bidders. They derive five conditions to be satisfied by the distribution of bids $G_i(b; z)$ in a Bayes-Nash equilibrium, where z is a set of observable covariates and i the index of a firm. One condition, called the conditional independence, implies that firm i's and firm j's bids are independently distributed, conditional on a set of covariates z. Another condition postulates that the distribution of bids is exchangeable across bidders in equilibrium, again conditional on a set of covariates z. It implies that if we exchange the costs of firm i by those of firm j, then the bids of firm i should match the bids of firm j in the case of competition.

The covariates z_i are supposedly unique to each firm i. Bajari and Ye (2003) consider the distance of firm i's location to the project site and the firm's capacity utilization level as such covariates. However, this requires some additional efforts to calculate the respective covariates. Even if it is feasible to directly calculate the capacity used by each firm based on the bid summaries, one must be convinced that these bid summaries reflect an important share of the market. If the bid summaries used to construct proxies for the firms' capacities fail to cover an essential part of the market, one faces the risk of measurement error affecting the estimation (and therefore, the tests for collusion). From the bid summaries, we typically see the project site and if we also know the location of each bidding firm, we can compute the geographic distance. However, in some cases, the location of the contract and/or the firms' addresses might not be fully clear from the bid summaries and would need to be carefully checked. In this case, the requirement of obtaining well measured firm-specific covariates increases the cost of screening, whereas poorly measured covariates might jeopardize the accuracy of the econometric tests for collusion.

If the distribution of bids satisfies the conditions stated by Bajari and Ye (2003), then it is possible to construct firm-specific distributions of costs which uniquely rationalize the observed bids in equilibrium, i.e. under competition. When only some bidders compete while others collude, the structural model must be adapted to recover the costs of competitors and those of the cartel participants. However, the researcher needs to observe some competitive bids in order to reconstruct the distribution of the cartel costs. If all bidders in some tender are cartel participants such that no bidder competes, it is impossible to reconstruct the distribution of costs for such tenders with perfect collusion by this method (which is the case for most of the Swiss and Japanese collusive tenders).

As discussed in Bajari and Ye (2003), we may non-parametrically estimate the distributions of bids in order to test the conditional independence and the exchangeability of bids, given that we have large amounts of data. With limited data, we can estimate the so-called reduced form of the bidding function following Porter and Zona (1993) and Porter and Zona (1999), where the bid is the dependent variable and the covariates are the exogenous cost variables. All subsequent works follow this regression approach (see Jakobsson, 2007; Aryal and Gabrielli, 2013; Chotibhongs and Arditi, 2012a; 2012b; Conley and Decarolis, 2016; Imhof, 2017; Bergman et al., 2019). In a regression framework, the conditional independence

implies that the residuals from the regression are not correlated across firms, since bids should be independent in a competitive situation, conditional on the observable cost proxies, i.e. the covariates. The exchangeability of the bids implies that firms react in the same way after controlling for their costs. We verify this by testing whether the estimated coefficients are comparable or different across firms, with statistically significantly distinct coefficients pointing to potential bid rigging. For running such tests, it is necessary to estimate a panel model with heterogeneous coefficients for each firm. This, however, requires a sufficient number of firm-specific observations, otherwise the method may suffer from a too high variance, impeding the practical implementation of testing.

In our Japanese dataset from Okinawa, a total of 656 bidders are present in the rank A and A+ contracts during the entire data window period including the pre-inspection, post-inspection and the post-amendment periods. However, only 407 bidders won at least one contract. Among the 249 bidders who never won a contract, 108 firms submitted 10 or more bids in the whole data window. This means that 38% of bidders labeled as rank A or A+ never won a contract of any type (including lower rank contracts) in the entire period, but kept submitting bids. This implies that the data from Okinawa do likely not cover the entire market or a substantial share thereof, since it is hard to see how firms can keep submitting bids without ever winning contracts. Not observing a large share of the market makes it difficult to build meaningful proxies for the firms' capacities. A further issue arises concerning the estimation of a panel model with heterogeneous coefficients for each firm as required for implementing the econometric tests. If we only consider the cartel period (the pre-inspection period), we find among the 612 firms labeled as rank A or A+ bidders only 6 firms that submitted at least 20 bids, whereas 490 firms bade less than 10 times. For the post-cartel period (the post-amendment period), only 12 firms submitted at least 20 bids among 580 firms of rank A and A+, while 455 firms bade less than 10 times. For this reason, estimating a panel model with heterogeneous coefficients is practically infeasible for the vast majority of firms in the Okinawa dataset due to too few firm-specific bids being submitted, such that we cannot apply the econometric tests of Bajari and Ye (2003).

Contrary to the Okinawa case, our Swiss dataset from Ticino contains firms which regularly submit bids such that estimating a panel model with firm-specific coefficients is feasible. Only one firm submits less than 20 bids in the data window and most of the firms submit between 50 to 206 bids (see Table 9 in Imhof (2017)). The number of bids per firm and (cartel and post-cartel) period is therefore appropriate to apply the econometric tests of Bajari and Ye (2003). However, Imhof (2017) shows that the reliability of such tests is questionable for the Ticino case. In the cartel period when collusion is present, only 40% of pairs of firms fail at least one test and only 4% fail the two econometric tests considered in Table 6 of Imhof (2017). In other words, at least 60% of the firm pairs are not classified as "cartel participants" during the cartel period. The error rate of the econometric tests is therefore quite substantial for the Ticino cartel and far larger than that obtained in our paper.

6.2. The density gap method

Chassang et al. (2022) investigate the phenomenon of a missing density or "density gap" in the distribution of bids, due to a gap in the values of the first and the second lowest bids in Japanese tenders. To analyze the incidence of a missing density, they use the normalized margin calculated by means of the following formula:

$$\Delta_{i,a} = \frac{b_{i,a} - \wedge b_{-i,a}}{r}, \tag{2}$$

where $b_{i,a}$ is the bid of some firm i in auction a, r is the reserve price and $\wedge b_{-i,a}$ is the minimum bid among bidders other than i. If firm i is the winning firm with the lowest bid, then $\Delta_{i,a}$ corresponds to the normalized distance between the winning bid and the minimum bid among losers, i.e. the second lowest bid in the tender, where the normalization is with respect to the reserve price. But $\Delta_{i,a}$ is computed for any firm in a tender, in order to investigate the density of this statistic. In their Japanese data, Chassang et al. (2022) find a small gap in the density of the normalized margin $\Delta_{i,a}$ around zero, which is incompatible with competition, and also formalize a model to show that bid rigging explains such a missing density around zero.

However, it is worth noting that the formula proposed by Chassang et al. (2022) can also mechanically create a gap around zero, which can be illustrated by a simple numeric example with bids normalized by the reserve price. Consider the following vector of normalized bids $\mathbf{b} = (0.95, 0.98, 0.99, 1)$. The strict application of the formula yields the following normalized margins $\Delta_{i,a}$ for that tender: $\{-0.03, 0.03, 0.04, 0.05\}$. The difference between the first and second lowest bid, which amounts to an absolute value of 0.3, entails a gap in the density of the normalized margin which is twice as large in absolute terms, namely 0.6 (from -0.3 to 0.3), with the center of the gap being at zero. We therefore suspect that this formula might detect too small modifications in the bidding patterns between the first and the second lowest bids.

Bearing this potential caveat in mind, we apply the Chassang et al. (2022) formula to the Okinawa bid-rigging cartels. As displayed in Fig. 4, we do not observe any density gap for the normalized margin in the cartel period around zero when simultaneously considering contracts of ranks A+ and A. Figs. 5 and 6 in Section A.1 of the Appendix depict the density of the normalized margin separately for contracts A+ and A, and also in these cases, no specific gap can be spotted. Therefore, it appears that the normalized margin of Chassang et al. (2022) is not capable to detect bid-rigging cartels in the Okinawa data and we subsequently propose one possible explanation for this failure.

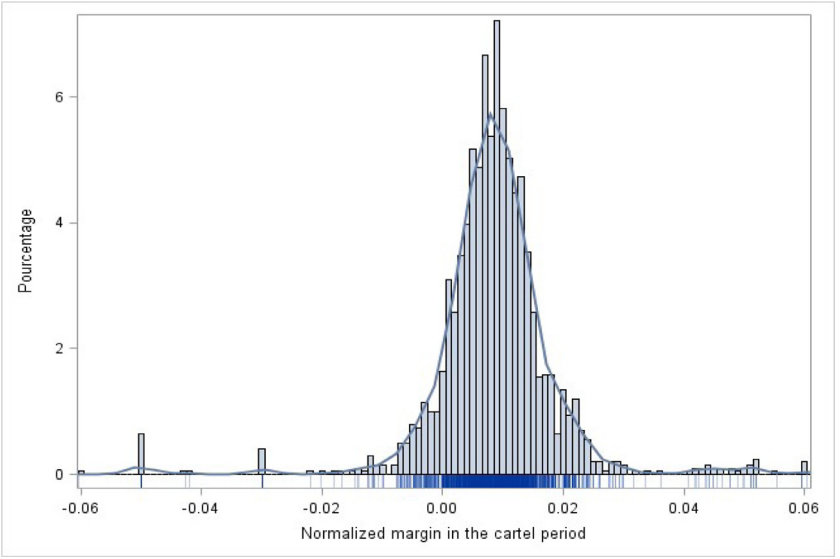

Fig. 4. Distribution of normalized margins of collusive bids for contracts A and A+.

When looking at the descriptive statistics for the percentage difference between the first and second lowest bids in tenders in Okinawa (see the DIFFP in Tables A1 and A2 in Huber et al., 2022), we find small discrepancies between the cartel and the competitive periods. For the cartel period, the lower quartile and the median amount to 0.26 and 0.41, respectively. For the competitive period, the lower quartile and the median amount to 0.17 and 0.52, respectively. About half of the tenders exhibit similar and very small values for the percentage difference between the first and second lowest bids. However, when investigating the descriptive statistics for yet another statistical screen referred to as relative distance (RD, RDNOR or ALTRD in Tables A1 and A2 in Huber et al., 2022), there are strong differences between the cartel and competitive periods, even though differences in the first and second lowest bids are small. This finding suggests that the cartel strongly reduced the differences between losing bids, which is captured by the relative distance measures, while the differences between the first and second lowest bids were hardly affected by bid rigging. This result illustrates the risk of predominantly focusing on the distance between the first and second lowest bids, as it occurs to be the case for the method based on density gaps, and neglecting changes in the remaining bids, which might be captured by appropriate screens.

Finally, we note that the approach of Chassang et al. (2022) cannot be applied to our Swiss bid-rigging cases for which no reserve prices are available. At first glance, an apparent alternative could consist of using the minimal bid in a tender instead of the reserve price for calculating normalized margins. However, this would generally not entail margins which are comparable across collusive and competitive tenders, since minimum bids tend to be higher among collusive tenders. For this reason, the minimum bid is generally not a "neutral" reference price and this issue might distort the computation of the normalized margin.

6.3. Machine learning with screens

A further alternative to our deep learning architecture are machine learning methods using pre-defined statistical screens (rather than unstructured data like images), which are computed from the distributions of bids in the respective tenders, as features for predicting bid rigging. Among such machine learners, we can further distinguish between the so-called tender- and coalition-based approach. The former consist of calculating screens in each tender in order to capture the distributional disrupt when bid rigging occurs. Such tender-based approaches yielded correct classification rates of about 83% to 84% in different data sets from Switzerland (see Huber and Imhof, 2019; Wallimann et al., 2022). For the Japanese data from Okinawa, Huber et al. (2022) obtained correct classification rates ranging from 88.4% to 89.1%.

We apply a so-called ensemble method, a weighted average of five machine learning algorithms, which uses the tender-based screens considered in Huber et al. (2022) as predictors for classifying collusion in the Swiss and Japanese data and compare the performance to our CNN-based approach. We provide further details on the ensemble method along with references to other studies applying the same algorithm in Section A.5 of the Appendix. To facilitate comparability, we repeat the procedure of randomly splitting the sample into training/validation and test data to develop and assess the ensemble-based model 20 times, just as we did it for our CNN. Table 10 provides summary statistics of the accuracy in the test samples, which shows that the ensemble method using tender-based screens performs very well, but yet considerably worse than the deep learning approach. Using our CNN architecture increases average accuracy by at least ten percentage points in the Swiss data and five to six percentage points in the Japanese data when compared to machine learning with

Table 10
Summary statistics for the ensemble method using tender-based screens.

Country	Tenders	Minimum	1st quartile	Median	Mean	3rd quartile	Maximum
Japan	All tenders	0.836	0.864	0.882	0.888	0.918	0.927
	Cartel	0.810	0.861	0.893	0.886	0.909	0.953
	Competition	0.804	0.853	0.909	0.892	0.918	0.943
Switzerland	All tenders	0.760	0.827	0.846	0.841	0.856	0.884
	Cartel	0.711	0.841	0.859	0.853	0.876	0.923
	Competition	0.773	0.813	0.827	0.828	0.841	0.873

Table 11
Summary statistics for logit models using tender-based screens.

Country	Tenders	Minimum	1st quartile	Median	Mean	3rd quartile	Maximum
Japan	All tenders	0.809	0.836	0.855	0.861	0.893	0.927
	Cartel	0.717	0.781	0.810	0.816	0.839	0.933
	Comp.	0.857	0.895	0.920	0.920	0.941	1.000
Switzerland	All tenders	0.781	0.800	0.829	0.825	0.844	0.877
	Cartel	0.708	0.761	0.786	0.784	0.806	0.859
	Comp.	0.797	0.840	0.872	0.869	0.900	0.928

tender-based screens. Such a five percentage points increase implies a reduction of the misclassification rate by half for the Japanese sample, while for the Swiss sample, the relative reduction of incorrect classifications is even higher, amounting to roughly 63%.

In a next step, we also estimate logit models for classifying collusion in order to compare this more classical statistical tool to our machine and deep learning approaches. As conventional logit regression does not permit selecting predictors in a data-driven way, we base it on the three most predictive screens in Huber et al. (2022), namely, the coefficient of variation, the Kolmogorov-Smirnov statistic, and the normalized distance. Like for the other methods, we repeat the procedure of splitting the sample to estimate the logit coefficients (using the three screens) in the training data and assess the model's accuracy in the test data 20 times. The results reported in Table 11 suggest that the logit approach performs somewhat worse then the ensemble method, as the accuracy of the former is on average roughly 2 percentage points higher than that of the former.

Contrary to the tender-based methods discussed so far, the coalition-based approach of Imhof and Wallimann (2021) consists of focusing on fixed sets or "coalitions" of firms and calculating the screens within the respective coalition for each tender in which these firms participate. Once the tender- and coalition-specific screens are calculated for each possible coalition in the data, the screens are aggregated to summary statistics within each coalition, e.g. by computing the median, minimum or maximum of the respective screen. Finally, the coalition-based summary statistics serve as predictors for training the machine learners. Imhof and Wallimann (2021) find the accuracy ranging from 91.9% to 94.9% for the Japanese data from Okinawa and from 86.9% to 90.5% for various Swiss data sets (including the Ticino case), and from 84.8% to 90.1% for an Italian data set, which is introduced in Section A.4 in the Appendix. This implies that the CNN also outperforms the coalition-based approach. On average, the accuracy of the CNN is five percentage points higher in the Swiss data and even seven percentage points higher in the Italian data, see the results in Section A.4 in the Appendix. The CNN's accuracy is also higher for the Okinawa case, even if less substantially so.

6.4. Performance gains related to bidding interactions and autonomous feature learning

The CNN-based approach seems to exhibit a higher predictive performance than the tender- or the coalition-based approach with machine learning and pre-defined screens. This might be explained by the autonomous learning process of CNNs, which aim at constructing and applying the most appropriate features for distinguishing collusion and competition in a data-driven way from the plots. On the other hand, the information contained in the plots of bidding interactions might also play a crucial role for the performance of a CNN. The fact that plots are two-dimensional could be an advantage over the machine learning use of screens, which are one-dimensional summary statistics and might for this reason not capture important distributional changes across collusive and competitive bids that well. We investigate now which role the autonomous learning process of CNNs and the two-dimensional plots play in explaining the higher predictive performance of the CNNs.

In order to assess the merits in terms of predictive performance of depicting bid interactions of a reference firm against all other firms by means of plots, we first use a base model that exclusively exploits distributional statistics coming from the normalized bids of the reference firm, but not the interactions. More concisely, we calculate the 10th percentile, the first quartile, the median, the third quartile and the 90th percentile of the normalized bid for each of the reference firms considered in the plots, in order to use these statistics as predictors for collusion. This approach is one-dimensional in the sense that these statistics come from the reference firm alone, but the statistics are yet different from the screens considered

Table 12
Ensemble method using distributional statistics of normalized bids.

Country	Firms	Minimum	1st quartile	Median	Mean	3rd quartile	Maximum
Japan	All firms	0.792	0.833	0.854	0.867	0.917	0.944
	Cartel participants	0.816	0.86	0.895	0.891	0.915	0.972
	Competitors	0.757	0.805	0.831	0.844	0.882	0.944
Switzerland	All firms	0.661	0.692	0.732	0.728	0.750	0.875
	Cartel participants	0.543	0.612	0.689	0.683	0.760	0.842
	Competitors	0.629	0.736	0.761	0.764	0.791	0.892
Italy	All firms	0.642	0.709	0.746	0.731	0.765	0.806
	Cartel participants	0.000	0.151	0.205	0.188	0.236	0.357
	Competitors	0.868	0.918	0.960	0.949	0.978	1.000

Table 13
Ensemble method using quadrant-based statistics of the plots.

Country	Firms	Minimum	1st quartile	Median	Mean	3rd quartile	Maximum
Japan	All graphs	0.889	0.917	0.944	0.938	0.958	0.972
	Cartel	0.875	0.902	0.930	0.932	0.952	1.000
	Competition	0.872	0.922	0.944	0.944	0.974	1.000
Switzerland	All graphs	0.883	0.933	0.950	0.949	0.971	0.983
	Cartel	0.857	0.882	0.939	0.928	0.962	1.000
	Competition	0.892	0.949	0.971	0.966	0.981	1.000
Italy	All graphs	0.776	0.821	0.836	0.838	0.854	0.881
	Cartel	0.500	0.586	0.659	0.660	0.753	0.800
	Competition	0.833	0.886	0.908	0.905	0.920	0.958

in Section 6.3. Table 12 provides the predictive performance when applying an ensemble method as machine learner.[5] We see that the accuracy of this approach is considerably lower than that of machine learning with screens or our CNN. Table 14 in the Appendix also reports the results when using logit regression rather than the ensemble method, and performs even worse (by on average five to ten percentage points when compared to the ensemble learner).

In a second step, we investigate the value added of two-dimensional statistics coming from the plots. To this end, we divide each plot into quadrants by cutting both the x- and y-axes at a threshold value of 0.5. This provides us with four quadrants of the same dimension, since the normalized bids of reference and co-bidding firms are between 0 and 1. We then compute the percentage of dots (or bid interactions) in each quadrant relative to the total number of dots in a plot. Finally, we use the percentages in the various quadrants as predictors (the logit estimation uses only three out of four variables as including all four would entail perfect multicollinearity). Table 13 reports the predictive performance based on the ensemble method when considering the bid-rigging data from Japan, Switzerland, and Italy, respectively, while Table 15 in the Appendix provides the results for logit regression.

Exploiting two-dimensional information on both reference and co-bidding firms increases the average accuracy by 10 to 20 percentage points for predicting if a firm is a cartel participant or a competitor relative to the one-dimensional approach. This points to a substantial value added of considering the normalized bids in a two-dimensional plot, not only for our CNN approach, which learns predictive features from the plots autonomously, but also for the ensemble and logit models, that rely on the quadrant-specific percentages as pre-defined predictors. The relative advantage of the CNN over the ensemble method appears to differ from case to case. When considering the data from Switzerland and Japan, for which the use of two-dimensional information entails an important boost in predictive performance, the accuracy of the CNN is on average just about one percentage point higher than that of the ensemble method. Both approaches excel with an average accuracy well beyond 90%. In the Italian data, however, the CNN clearly outperforms the ensemble method by 13 percentage points in terms of average accuracy, such that autonomously learning predictive features importantly improves upon the quadrant-based approach.

In conclusion, our results indicate that dividing the bidding plots into four quadrants can be very effective for capturing fundamental interaction patterns among collusive bidders. As discussed in Section 2, most collusive bidders in Japan and in Switzerland submitted high cover bids to make sure that another firm wins the contract (namely, the designated winner by the cartel). For this reason, many collusive dots (or interactions) are located in the top right quadrant (see Fig. 1), but only few (if any) in the bottom left quadrant. In other words, the collusive interactions, once plotted on two dimensions, are so different to competitive interactions in Switzerland and in Japan that any of the investigated methods has a high accuracy in classifying collusion and competition. This might, however, not be a general pattern in all data sets. Indeed, in the Italian data, the autonomous learning approach of CNNs clearly outperforms methods relying on the ad hoc approach of defining predictors based on the quadrants of the plot.

[5] See Section A.5 in the Appendix for further details on the ensemble method along with references to other studies applying this algorithm.

6.5. A practical assessment of alternative detection methods

Even though accuracy is certainly one of the most relevant criteria for assessing a collusion detection method, a law enforcer also considers the data requirements and the general applicability of a method as important factors. The method proposed by Chassang et al. (2022) has comparably modest data requirements, since it only relies on the bids in each tender and the reserve price. However, even with such apparently low data requirements, we cannot apply the methods to the Swiss data since the reserve price – if it exists at all – is not recorded in the bid summaries. Furthermore, the application of their method does not seem to work satisfactorily in the Japanese data from Okinawa, mostly because it focusses on the differences between the lowest bid and the other bids and not on the reduced variance between losing bids, which seems to play a crucial role in those data (as in the Swiss data, too).

Bajari and Ye (2003) also suggest a formal model to test collusion, however, with higher data requirements than any other method considered in this paper. More specifically the method relies on proxies for the cost variables of each firm of sufficiently high quality to be used for regression and statistical inference. In the case of Okinawa, collecting such data for each firm would have been infeasible, as well as estimating a panel model with heterogeneous coefficients across firms (which is difficult if the number of observations per firm is low) as required for applying the econometric tests of Bajari and Ye (2003). However, in the Swiss data from the Ticino cartel, in which the data requirements are met, the majority of the econometric tests fail to detect collusive episodes and are strongly outperformed by a simple screen-based benchmarking approach without machine learning (see Imhof, 2019). For this reason, this method does not seem to be an appropriate general screening tool. However, it might be interesting to apply such econometric tests to cases already flagged as potential cartels by other methods, in order to support the first diagnostic (given that the availability of cost proxies and other data requirements are satisfied).

Both the tender- and the coalition-based approach use screens for detecting cartels based on machine learning and only rely on bid summaries that permit identifying the bidders and the bids. Data requirements are, thus, low, while the average accuracy is very decent by varying between 85% and 95%. Moreover, the coalition-based approach might also be applied to other auction formats as the mean-price sealed-bid auction in Italy. All in all, these methods appear adequate as general tools for screening procurement markets since their cost of implementation is low and the results are sufficiently reliable for flagging markets in which competition might be endangered.

The CNN-based detection method differs from the previously mentioned machine learning methods as it focuses on pairwise firm interactions rather than tenders or coalitions. The data requirements are equally low as for detection methods based on machine learning. Its high accuracy seems to be due to two characteristics: (a) the representation of the interactions between a reference firm and co-bidders in a two-dimensional plot and (b) the autonomous learning process of CNNs. In cases where the interaction pattern of collusive firms is very distinct from that of competitive bidders and, thus, concentrated in specific parts of the plot, using predefined and rather basic statistics of the plots (like the share of dots in some quadrant) as predictors in machine learners might perform almost as well as a CNN. This is what we observe in the Swiss and Japanese data. However, predefining plot-related statistics to be used as predictors (a process known as feature-engineering) generally requires extensive domain knowledge to create predictors that are (close to) optimal in terms of performance. Deep learning has the advantage of learning predictive features autonomously from the plots and might therefore outperform machine learning in terms of accuracy when the complexity of interaction patterns (in terms of locations in the plots) increases, as we observe it in the Italian cartel data. This, however, comes with the caveat that CNNs require us to set up a network architecture, for which we face an infinite number of potential choices concerning properties and tuning parameters like the number of filers, nodes, layers, epochs, observations in a batch, etc. On the positive side, we may use the previously described validation approach to optimize such choices with regard to predictive performance.

From the perspective of a competition agency engaged in screening procurement markets, the high accuracy of the CNN approach (as to a lesser extent the machine learning approach) is an attractive and practically relevant characteristic. In order to open an investigation, a competition agency must have reasonable grounds for a possible law infringement, which means that it must have formed an initial suspicion against specific firms for having fomented a bid-rigging conspiracy. Such a suspicion implies that the likelihood of the presence of a cartel must be significantly higher than that of the absence of a cartel. Being able to correctly classify a firm as a cartel participant or competitor in the vast majority of cases (in our data 19 out of 20 times) might well satisfy the required standards of empirical evidence for deciding whether an investigation should or should not be launched.

Moreover, the CNN-based approach has an advantage over the tender-based and the coalition-based approach because it can deliver a direct probability for each reference firm to be a cartel participant. This allows a competition agency or procurement agency to focus only on the most likely cartel candidates for further observations or tests. With this respect, both techniques could be combined in order to further reduce the risk of false positives (i.e classifying competitive firms as cartel members). A public enforcer might for example use the tender-based screening approach with machine learning to detect a set of potentially collusive tenders and simultaneously apply the CNN approach in order to detect potentially collusive firms. Comparing the findings of both methods might e.g. reveal that the potentially collusive firms are mostly active in potentially collusive tenders. Thus, a large intersection of the results coming from both methods can further increase the confidence in flagging suspicious cases. This can then serve as a base for an investigator's decision to perform additional tests, analyses or research to validate the initial suspicions. After all, it is up to a human (e.g. a judge) to decide whether

a legal investigation should be opened based on the data-driven indications provided by algorithms or other empirical methods.

7. Conclusion

This paper contributes to the literature on the data-driven detection of bid-rigging cartels by proposing a novel method based on a deep learning architecture involving a convolutional neural network (CNN) for image recognition. We to this end construct graphs on pairwise bidding interactions of firms across various tenders, as also considered in the bid rotation test suggested in Imhof et al. (2018) and Imhof (2019), and use them as inputs for the CNN. We obtain an average out-of-sample accuracy of roughly 95% or higher in classifying graphs with collusive and competitive interactions when applying the method separately to Japanese and Swiss data on bid-rigging cartels. Pooling the graphs from both countries also yields a similarly high accuracy. Only when training (or developing) CNN models based on the graphs of one country in order to test the predictive performance in the respective other country, accuracy generally decreases. Yet, we still obtain an average accuracy of 85% when training in the Japanese and testing in the Swiss data and of 84% when training in the Swiss and testing in the Japanese data. However, imbalances in accuracy across collusive and competitive cases increase importantly when transferring trained models to another country, implying that the method no longer works similarly well for correctly predicting either outcome class. Nevertheless, the fact that we attain a high predictive performance in most scenarios considered despite our limited sample of just a few 100 graphs suggests that our CNN approach for bidding interaction classification is a very promising tool for screening procurement markets by inferring if some firm is likely a cartel member or not.

In line with the impressive successes of deep learning in many domains like natural language processing and pattern recognition, our approach has the potential to outperform other quantitative approaches for flagging bid rigging and is one of the rare applications of CNNs in economics to date. Since our method focuses on the bidding interactions of a reference firm with other firms participating in the same tender, it (in contrast to most other approaches) permits identifying potential cartel members directly for further scrutiny. Indeed, a similarly high accuracy as in our analysis, implying that the CNN correctly classifies 19 out of 20 firms as cartel members or competing bidders, provides a solid base for competition agencies concerning the decision to open an investigation against a suspicious firm. Moreover, our method is also well suited for combining different screens and approaches. The accuracy of our CNN-based approach relies on the two-dimensional representation of interactions between a reference firm and co-bidders as well on the autonomous learning process of CNNs. When collusive interactions are very distinct from competitive interactions, using descriptive statistics of the plots as predictors in machine learning might reach the performance of CNNs, as in the Swiss and Japanese cases. However, as soon as interactions become less obvious and more complex, CNNs might outperform predefined predictors with machine learning, as in the Italian cartel data.

Even though the scope for applications of CNNs appears ample for fighting collusion and fraud more generally, a practical issue is that there is an infinite number of ways a CNN could be constructed in terms of tuning parameters like the number/size of filters (for extracting patterns), hidden layers (determining model complexity), nodes (nonlinear functions per layer), and many others. The best performing CNN architecture is a priori not obvious and needs to be iteratively explored, making deep learning seemingly more an art than a science. While the model analyzed in this paper yielded a very decent accuracy considering our comparably small data, it is possible that different specifications may result in higher performance in other contexts and with a larger amount of graphs on bidding interactions.

Disclaimer: All views contained in this paper are solely those of the authors and cannot be attributed to the Swiss Competition Commission, its Secretariat, the University of Fribourg, Unidistance (Switzerland).

Data availability

Data will be made available on request.

CRediT authorship contribution statement

Martin Huber: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration. **David Imhof:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration.

Acknowledgement

The authors would like to thank Emanuel Viklund, an anonymous referee and the Editor for helpful comments.

Appendix A

A1. Figures

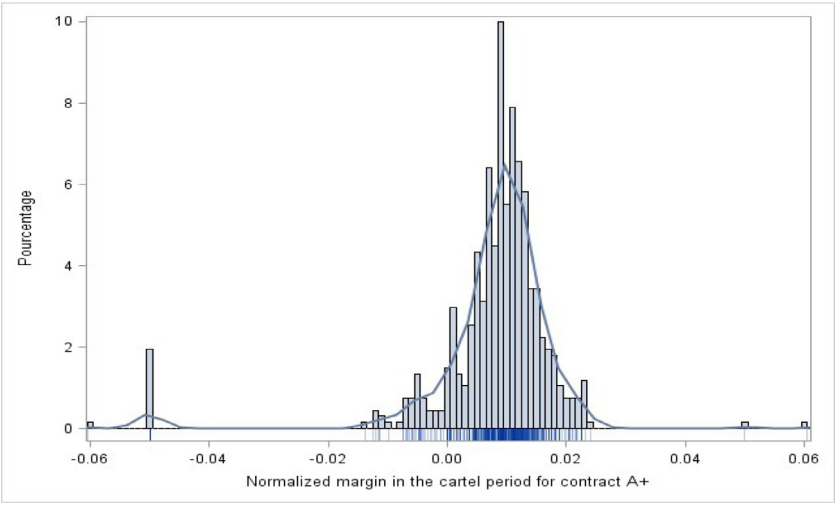

Fig. 5. Distribution of normalized margins of collusive bids for contracts A+.

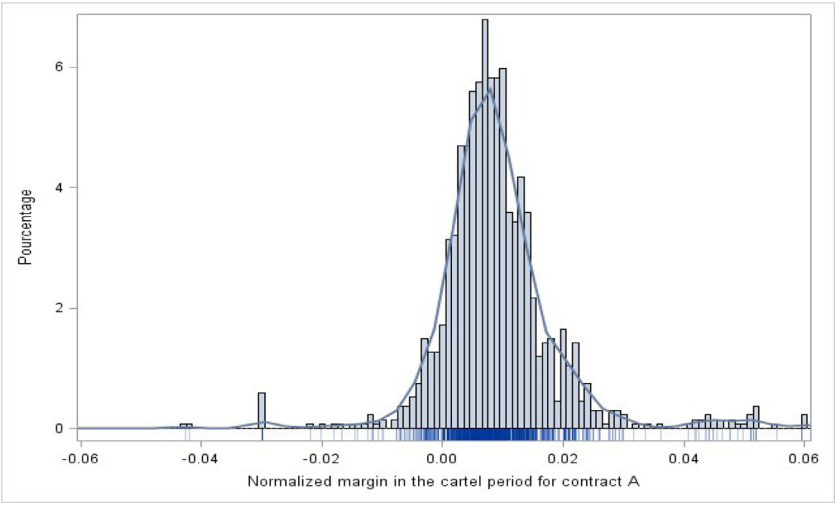

Fig. 6. Distribution of normalized margins of collusive bids for contracts A.

A2. Tables

Table 14
Logit model using distributional statistics of normalized bids.

Country	Firms	Minimum	1st quartile	Median	Mean	3rd quartile	Maximum
Japan	All firms	0.736	0.764	0.785	0.793	0.819	0.861
	Cartel participants	0.579	0.676	0.719	0.714	0.750	0.865
	Competitors	0.757	0.845	0.864	0.867	0.890	0.949
Swiss	All firms	0.583	0.617	0.633	0.648	0.667	0.750
	Cartel participants	0.143	0.181	0.236	0.251	0.333	0.368
	Competitors	0.833	0.886	0.939	0.925	0.946	1.000
Italy	All firms	0.597	0.672	0.694	0.702	0.724	0.821
	Cartel participants	0.000	0.000	0.000	0.000	0.000	0.000
	Competitors	0.978	0.995	1.000	0.995	1.000	1.000

Table 15
Logit model using quadrant-based statistics of the plots.

Country	Firms	Minimum	1st quartile	Median	Mean	3rd quartile	Maximum
Japan	All graphs	0.875	0.913	0.938	0.931	0.948	0.972
	Cartel	0.825	0.899	0.920	0.922	0.957	1.000
	Competition	0.857	0.915	0.939	0.939	0.97	1.000
Swiss	All graphs	0.900	0.933	0.950	0.953	0.967	1.000
	Cartel	0.700	0.908	0.932	0.928	0.959	1.000
	Competition	0.905	0.947	0.971	0.969	1.000	1.000
Italy	All graphs	0.687	0.784	0.806	0.802	0.851	0.881
	Cartel	0.292	0.373	0.477	0.494	0.634	0.800
	Competition	0.852	0.902	0.918	0.92	0.946	0.961

A3. Model validation

The empirical results reported in Section 5 are based on a network architecture with 3 convolutional and 3 hidden layers as outlined in Table 4, which is our benchmark model. We also considered alternative networks in terms of the complexity of convolutional and hidden layers and investigated their predictive accuracy. To this end, we drew a 75% random subset of the Japanese sample (while 25% were held out as test data), used 90% as training data and 10% as validation data per epoch when estimating the model parameters of the various networks, and estimated the accuracy of each network in the validation data of the last epoch of the training process. More specifically, we analysed a more parsimonious network with just 1 convolutional and 1 hidden layer as reported in Table 16, as well as a more complex network with 4 convolutional and 5 hidden layers as reported in Table 17. Furthermore, we considered all possible combinations of the numbers of convolutional and hidden layers as outlined in Tables 4, 16, and 17. That is, we combined 1, 3, or 4 convolutional layers with 1, 3, or 5 hidden layers, yielding 9 alternative network architectures. We repeated this validation procedure (of randomly splitting observations into training and validation data in a sequence of epochs and assessing the accuracy of the neural networks in the last epoch) 20 times. Table 18 reports the mean, median, worst, and best performance across these 20 simulations for each network when setting the batch size to 16 and the number of epochs to 40. Our benchmark model with 3 convolutional and 3 hidden layers (at least slightly) dominates any other neural network in terms of the median and mean accuracy or other summary statistics.

Table 16
Architecture of model 2 (parsimonious model).

1. Convolutional layer (two-dimensional for gray-shaded images) with 8 filters, each with a sliding window size 3 × 3 pixels, using the rectifier (or ReLU) function, which is zero for negative values and linear for positive values, as nonlinear function for mapping pixels into numeric values.
2. Max pooling with size 2 × 2: aggregates the previously generated numeric features within sliding matrix windows of size 2 × 2 by taking the maximum value within each window.
3. Flattening: transforming each of the matrix values in the previous convolutional layer into separate input (or predictor) variables for the FNN to follow.
4. Hidden layer with 32 hidden nodes, using the rectifier as nonlinear function for mapping the nodes of the previous hidden layer.
5. Output layer for classifying the binary outcome (1=cartel, 0=competition), using the sigmoid function as nonlinear function for mapping the nodes of the last hidden layer in order to predict the outcome.

Table 17
CNN architecture for predicting collusion.

1. Convolutional layer (two-dimensional for gray-shaded images) with 64 filters, each with a sliding window size 5 × 5 pixels, using the rectifier (or ReLU) function, which is zero for negative values and linear for positive values, as nonlinear function for mapping pixels into numeric values.
2. Max pooling with size 3 × 3: aggregates the previously generated numeric features within sliding matrix windows of size 2 × 2 by taking the maximum value within each window.
3. Convolutional layer with 32 filters, each with kernel size 3 × 3, using the rectifier as nonlinear function for mapping the pooled features.
4. Max pooling with size 2 × 2.
5. Convolutional layer with 16 filters, each with kernel size 3 × 3, using the rectifier as nonlinear function for mapping the pooled features.
6. Max pooling with size 2 × 2.
7. Convolutional layer with 8 filters, each with kernel size 3 × 3, using the rectifier as nonlinear function for mapping the pooled features.
8. Flattening: transforming each of the matrix values in the previous convolutional layer into separate input (or predictor) variables for the FNN to follow.
9. Hidden layer with 128 hidden nodes (i.e. 128 nonlinear regression functions), using the rectifier as nonlinear function for mapping the inputs.
10. Hidden layer with 64 hidden nodes, using the rectifier as nonlinear function for mapping the nodes of the previous hidden layer.
11. Hidden layer with 32 hidden nodes, using the rectifier as nonlinear function for mapping the nodes of the previous hidden layer.
12. Output layer for classifying the binary outcome (1=cartel, 0=competition), using the sigmoid function as nonlinear function for mapping the nodes of the last hidden layer in order to predict the outcome.

Table 18
Accuracy in validation data (last epoch) across networks, 75% of Japanese sample.

Number of layers	Minimum	1st quartile	Median	Mean	3rd quartile	Maximum
3 convolutional, 3 hidden	0.864	0.955	1.000	0.971	1.000	1.000
1 convolutional, 1 hidden	0.182	0.625	0.886	0.782	0.955	1.000
4 convolutional, 5 hidden	0.864	0.955	0.955	0.952	0.966	1.000
3 convolutional, 5 hidden	0.864	0.955	0.955	0.964	1.000	1.000
3 convolutional, 1 hidden	0.545	0.955	0.977	0.950	1.000	1.000
1 convolutional, 3 hidden	0.864	0.943	0.955	0.950	0.966	1.000
1 convolutional, 5 hidden	0.864	0.909	0.955	0.941	0.955	1.000
4 convolutional, 1 hidden	0.864	0.955	0.955	0.952	0.966	1.000
4 convolutional, 3 hidden	0.455	0.943	0.955	0.932	1.000	1.000

Note: 40 epochs, batch size is 16. The network with 3 convolutional and 3 hidden is the benchmark network.

For our benchmark model only with the batch size set to 16, the upper panel of Table 19 reports the accuracy in the validation data of the last epoch when setting the number of epochs to 20, 40, or 60. The results suggest that varying the number of epochs within barely affects accuracy. The empirical results presented in Section 5 are based on 40 epochs, as this choice seems to perform at least as good as any other numbers of epochs investigated. The lower panel of Table 19 provides the results when setting the number of epochs to 40, but varying the batch size between 4, 16, and 64. The predictive performance is very stable across these values of the batch size. Section 5 reports the empirical results for a batch size of 16, which is not dominated by the other choices of 4 or 64 in terms of accuracy.

Table 19
Summary statistics for accuracy across number of epochs and batch size.

3 convolutional, 3 hidden	Minimum	1st quartile	Median	Mean	3rd quartile	Maximum
varying the number of epochs (batch size is 16)						
40 epochs	0.864	0.955	1.000	0.971	1.000	1.000
20 epochs	0.864	0.955	1.000	0.968	1.000	1.000
60 epochs	0.864	0.955	0.977	0.968	1.000	1.000
varying the batch size (40 epochs)						
batch size is 16	0.864	0.955	1.000	0.971	1.000	1.000
batch size is 4	0.864	0.955	0.977	0.966	1.000	1.000
batch size is 64	0.864	0.955	0.977	0.964	1.000	1.000

Note: The benchmark model uses 40 epochs and batch size of 16.

A4. Applying deep learning to Italian cartel data

In this section we apply our CNN approach to a further data set, which comes from the municipality of Turino in Italy and contains information on roadwork contracts from 2000 to 2003 (Conley and Decarolis, 2016; Imhof and Wallimann, 2021, see for details). Investigations revealed the existence of eight cartels consisting of 95 firms suspected to participate in bid-rigging conspiracies. In 2008, a court sentenced 27 firms for bid rigging. The Italian data differ from the Swiss and Japanese data in two dimensions. First, Swiss and Japanese procurement agencies use first-price sealed-bid auctions to procure a contract, whereas the Italian procurement agencies use mean-price sealed-bid auctions. In the former case, the bidder who submits the lowest bid wins the contract, while the latter mechanism awards the contract to the bidder, the closest to the mean. This difference in the awarding rules implies that interactions and agreements among collusive firms to generate collusive bids differ between first-price and mean-price sealed-bid auctions.

Second, the Italian data consist of incomplete cartel cases, i.e. not all the firms present in the data participated in the cartel, which entails a different data structure than for the Swiss and Japanese samples. As reported in Table 20, out of the 276 contracts in the Italian data, all in all 106 cartel participants (who have submitted at least 20 bids in our data window) were active in 273 actions, submitting a total of 8013 bids. Furthermore, all in all 522 competitive bidders (submitting at least 20 bids) participated in the 276 contracts, submitting a total of 12,202 bids. Given this overlapping data structure of collusive and competitive bids, we cannot construct plots based on distinct collusive and competitive periods as described in Section 2 for the Swiss and Japanese cartels. However, we may adapt the method and generate the plots by type of bidders, i.e. separately for collusive and competitive bidders (who submitted at least 20 bids) in the 276 contracts considered. So for collusive reference bidders, only collusive bids are considered in the plots, and for competitive reference bidders only competitive bids. This allows us to investigate if the behavior of collusive and competitive bidders is different even when they

participate in the same tenders. Indeed, suspected cartel participants won 80% of the contracts while they only represent 10% of the bidders. This suggests that they were successful in implementing their collusive strategies.

Table 20
Italian sample.

Firm	Number of bids	Number of firms	Number of contracts
Collusive	8013	106	273
Competitive	12202	522	276

Fig. 7 depicts the interaction of a typical cartel participant (left plot) and of a competitive firm (right plot). For the competitive bidder, we see that the pairwise observations are scattered across the entire plot whereas for the collusive bidder, there is concentration of pairwise observations in the upper right part of the plot.

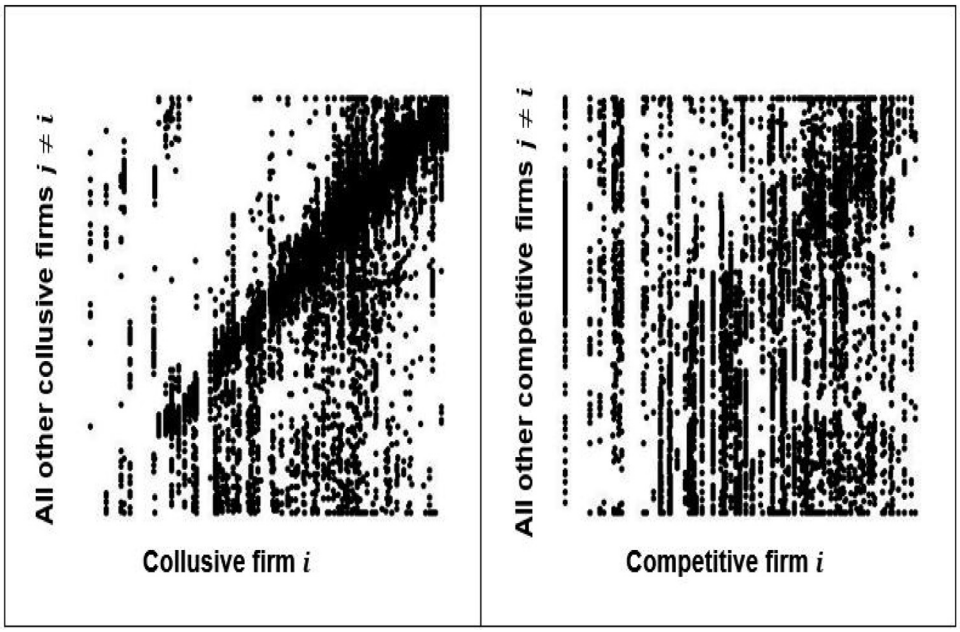

Fig. 7. Interactions of Italian firms with other bidders (left: cartel firm; right: competitive firm).

Table 21 reports the predictive performance of our CNN approach when applied to 266 plots constructed from the Italian data. The median and mean accuracy amount to 97% and 96.7%, respectively, which suggests that our deep learning approach performs decently across different types of auctions (and different countries). However, we observe an imbalance in the accuracy across collusive and competitive classes, which might be due to the limited sample and to the fact that we have more competitive plots than collusive plots. This puts a larger weight on and, thus, favors the correct prediction of competition in the training process, such that the true negative rate (of correctly classifying truly competitive firms) is very high with a median and mean of 100% and 99.9%, respectively. In contrast, the respective numbers amount to 89% for the true positive rate (of correctly classifying truly collusive firms). Overall, the performance appears very satisfactory also in the Italian data (with on average more than 19 correct classifications out of 20).

Table 21
Summary statistics for accuracy when training and testing in Italy.

	Minimum	1st quartile	Median	Mean	3rd quartile	Maximum	Observations
All graphs	0.909	0.951	0.970	0.967	0.989	1.000	266
Collusion	0.667	0.846	0.890	0.890	0.969	1.000	77
Competition	0.979	1.000	1.000	0.999	1.000	1.000	189

Note: Mean precision is 0.997

A5. Ensemble method

As in Huber and Imhof (2019) and in Huber et al. (2022), we apply an ensemble method as machine learner for several approaches presented in Sections 6.3 and 6.4. This ensemble method is a weighted average of five algorithms: bagged decision trees, random forest, lasso regression, support vector machines and shallow neural nets. For either method, we randomly split the data into training and test samples, which contain 75% and 25% of the observations, respectively. Cross-validation in the training sample determines the optimal weight each of the five machine-learning algorithms obtains in the ensemble method. To this end we apply the "SuperLearner" package for "R" by van der Laan et al. (2008) with default values for bagged tree, random forest, Bayesian additive regression tree, lasso regression, support vector machine and neural net algorithms in the "ipred", "partykit", "glmnet", "kernlab" and "net" packages, respectively, see Peters et al. (2002), Hothorn and Zeileis (2015), Friedman et al. (2010), Karatzoglou et al. (2004) and Venables and Ripley (2002). For further details about the various machine learning algorithms, we refer to Huber and Imhof (2019) and Huber et al. (2022).

References

Abrantes-Metz, R.M., Froeb, L.M., Geweke, J.F., Taylor, C.T., 2006. A variance screen for collusion. Int. J. Ind Organiz 24, 467–486.
Abrantes-Metz, R.M., Kraten, M., Metz, A.D., Seow, G., 2012. Libor manipulation. Journal of Banking and Finance 36, 136–150.
Anysz, H., Foremny, A., Kulejewski, J., Nical, A., 2018. Collusion and bid rigging in the construction industry: case studies from poland. Proceedings of the Creative Construction Conference.
Aryal, G., Gabrielli, M.F., 2013. Testing for collusion in asymmetric first-price auctions. Int. J. Ind Organiz vol 31, 26–35.
Asker, J., 2010. A study of the internal organization of a bidding cartel. American Economic Review 100, 724–762.
Bajari, P., Ye, L., 2003. Deciding between competition and collusion. Rev Econ Stat 85, 971–989.
Baldwin, L.H., Marshall, R.C., Richard, J.-F., 1997. Bidder collusion at forest service timber sales. Journal of Political Economy 105, 657–699.
Banerji, A., Meenakshi, J.V., 2004. Buyer collusion and efficiency of governement intervention in wheat markets in northern India: an asymmetric structural auctions analysis. Am J Agric Econ 86, 236–253.
Bergman, M.A., Lundberg, J., Lundberg, S., Stake, J.Y., 2019. Interactions across firms and bid rigging. Review of Industrial Organization 56 (1), 107–130.
Bolotova, Y., Connor, J.M., Miller, D.J., 2008. The impact of collusion on price behavior: empirical results from two recent cases. Int. J. Ind Organiz 26, 1290–1307.
Chassang, S., Kawai, K., Nakabayashi, J., Ortner, J., 2022. Robust screens for noncompetitive bidding in procurement auctions. Econometrica 90 (1), 315–346.
Chotibhongs, R., Arditi, D., 2012. Analysis of collusive bidding behavior. Construction Management and Economics 30, 221–231.
Chotibhongs, R., Arditi, D., 2012. Detection of collusive behavior. J Constr Eng Manag 138, 1251–1258.
Ciresan, D.C., Meier, U., Masci, J., Maria Gambardella, L., Schmidhuber, J., 2011. Flexible, high performance convolutional neural networks for image classification. IJCAI Proceedings-International Joint Conference on Artificial Intelligence 22, 1237–1242.
Conley, T.G., Decarolis, F., 2016. Detecting bidders groups in collusive auctions. American Economic Journal: Microeconomics 8 (2), 1–38.
Falbel, D., Allaire, J.J., Chollet, F.c., Tang, Y., Van Der Bijl, W., Studer, M., Keydana, S., 2021. R interface to keras version 2.4.0. R manual.
Feinstein, J.S., Block, M.K., Nold, F.C., 1985. Asymmetric information and collusive behavior in auction markets. Am Econ Rev 75, 441–460.
Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. J Stat Softw 33, 1–22.
Froeb, L.M., Sibley, D.S., Doane, M.J., Pinto, B.P., 2014. Screening for collusion as a problem of inference. The Oxford Handbook of International Antitrust Economics 2.
Harrington, J.E., 2007. Behavioral screening and the detection of cartels. In: Ehlermann, C., Atanasiu, I. (Eds.), European Competition Law Annual 2006: Enforcement of Prohibition of Cartels. Oxford: Hart..
Hothorn, T., Zeileis, A., 2015. Partykit: a modular toolkit for recursive partytioning in R. Journal of Machine Learning Research 16, 3905–3909.
Huber, M., Imhof, D., 2019. Machine learning with screens for detecting bid-rigging cartels. Int. J. Ind Organiz 65, 277–301.
Huber, M., Imhof, D., Ishii, R., 2022. Transnational machine learning with screens for flagging bid-rigging cartels. Journal of the Royal Statistical Society, Series A (Statistics in Society) 1–41.
Hueschelrath, K., Veith, T., 2014. Cartel detection in procurement markets. Managerial and Decision Economics 35 (6), 404–422.
Imhof, D., 2017. Econometric tests to detect bid-rigging cartels: does it work? Working Papers SES 483, Faculty of Economics and Social Sciences, University of Fribourg (Switzerland).
Imhof, D., 2019. Detecting bid-rigging cartels with descriptive statistics. Journal of Competition Law and Economics 15 (4), 427–467.
Imhof, D., Karagoek, Y., Rutz, S., 2018. Screening for bid rigging, does it work? Journal of Competition Law and Economics 14 (2), 235–261.
Imhof, D., Wallimann, H., 2021. Detecting bid-rigging coalitions in different countries and auction formats. Int Rev Law Econ 68, 1–12.
Ishii, R., 2009. Collusion in repeated procurement auction: a study of paving market in japan. Int. J. Ind Organiz 27, 137–144.
Ishii, R., 2014. Bid roundness under collusion in japanese procurement auctions. Review of Industrial Organization 44, 241–254.
Jakobsson, M., 2007. Bid rigging in swedish procurement auctions. Working Paper.
Jimenez, J.L., Perdiguero, J., 2012. Does rigidity of PPrice hide collusion? Review of Industrial Organization 41, 223–248.
Karatzoglou, A., Smola, A., Hornik, K., Zeileis, A., 2004. Kernlab - an S4 package for kernel methods in r. J Stat Softw 11, 1–20.
van der Laan, M., Polley, E.C., Hubbard, A.E., 2008. Super learner. Statistical Applications of Genetics and Molecular Biology 6.
LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proc. IEEE 86, 2278–2324.
Lee, I.K., Hahn, K., 2002. Bid-rigging in auctions for korean public-works contracts and potential damage. Review of Industrial Organization 21, 73–88.
McCulloch, W., Pitts, W., 1943. A logical calculus of ideas immanent in nervous activity. Bulletin of Mathematical Biophysics 5 (4), 115–133.
McCulloch, W., 2014. Roundtable on ex officio cartel investigations and the use of screens to detect cartels.
Pesendorfer, M., 2000. A study of collusion in first-price auction. Rev Econ Stud 67, 381–411.
Peters, A., Hothorn, T., Lausen, B., 2002. Ipred: improved predictors. R News 2, 33–36.
Porter, R.H., Zona, J.D., 1993. Detection of bid rigging in procurement auctions. J Polit Econ 101, 518–538.
Porter, R.H., Zona, J.D., 1999. Ohio school milk markets: an analysis of bidding. RAND Journal of Economics 30, 263–288.
Rabuzin, K., Modrusan, N., 2019. Prediction of public procurement corruption indices using machine learning methods. KMIS.
Ripley, B.D., 1996. Pattern Recognition and Neural networks. Cambridge.
Rodriguez, M.J.G., Montequin, V.R., Fernandez, F.O., Balsera, J.M.V., 2020. Bidders recommender for public procurement auctions using machine learning: data analysis, algorithm, and case study with tenders from spain. Complexity.
Silveira, D., Vasconcelos, S., Resende, M., Cajueiro, D.O., 2021. Won't get fooled again: a supervised machine learning approach for screening gasoline cartels. CESifo Working papers.
Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with S, 4 ed. Springer.
Wallimann, H., Imhof, D., Huber, M., 2022. A machine learning approach for flagging incomplete bid-rigging cartels. Computational Economics.