Data Description Sheet for the paper "**Detecting Accounting Fraud in Publicly Traded U.S. Firms Using a Machine Learning Approach**" by Yang Bao, Bin Ke, Bin Li, Y. Julia Yu, and Jie Zhang (November 2019)

1. A description of which author(s) handled the data and conducted the analyses.

Yang Bao and Julia Yu handled data. Yang Bao, Bin Ke, Bin Li, and Julia Yu jointly conducted the analyses.

2. A detailed description of how the raw data were obtained or generated, including data sources, the specific date(s) on which data were downloaded or obtained, and the instrument used to generate the data (e.g., for surveys or experiments). We recommend that more than one author is able to vouch for the stated source of the raw data.

Archival data are used in this paper. The data are mainly obtained from commercially available sources. The details are as follows.

1) AAER Data: our initial accounting fraud sample comes from the SEC's Accounting and Auditing Enforcement Releases (AAERs), as compiled by the University of California-Berkeley's Center for Financial Reporting and Management (CFRM).
2) The AAER database used in the current version of our paper includes all the AAERs announced over the period between May 17th, 1982 and September 30th, 2016. Since the CFRM has not updated the AAER database since 2017, we also hand collected additional fraud observations from the SEC website: https://www.sec.gov/divisions/enforce/friactions/friactions2018.shtml that are dated up to December 31, 2018 (AAER #4012). We make the fraud observations available (see item 5)
3) Financial Accounting Data: The publicly traded U.S. firms' accounting data are from COMPUSTAT fundamental annual database fiscal year 1991 to 2014. The data from COMPUSTATA used in the current version of the paper were downloaded in April 2017.

All the authors have access to the raw data mentioned above.

3. If the data are obtained from an organization on a proprietary basis, the authors should privately provide the editors with contact information for a representative of the organization who can confirm data were obtained by the authors. The editors would not make this information publicly available. The authors should also provide information to the editors about the data sharing agreement with the organization (e.g., non-disclosure agreements, any restrictions imposed by the organization on the authors, such as restrictions to publish certain results).

Not Applicable.

4. A complete description of the steps necessary to collect and process the data used in the final analyses reported in the paper. For experimental and survey papers, we require information about the instructions and instruments used to generate the data, subject eligibility and/or selection, as well as any exclusion criteria. The full set of instructions and instruments can be provided in the online appendix.

We provide a complete description of the steps necessary to collect and process the data in Section 3 "The Sample and Data" of the paper, and make our final dataset publicly available as described below.

5. The computer programs or code used to convert the raw data into the final dataset used in the analysis plus a brief description that enables other researchers to use this program. The purpose of this requirement is to facilitate replication and to help other researchers understand in detail how the raw data were processed, the final sample was formed, variables were defined, outliers were treated, etc. This code or programming is in most circumstances not proprietary. However, we recognize that some parts of the code or data generation process may be proprietary, including from the authors' perspective. Therefore, instead of the code or program, researchers can provide a detailed step-by-step description of the code or the relevant parts of the code such that it enables other researchers to arrive at the same final dataset used in the analysis. In such cases, the authors should inform the editors upon initial submission, so that the editors can consider an exemption from the code sharing requirement. Whenever feasible, authors should also provide the identifiers (e.g., CIK, CUSIP) for their final sample. Authors should consult our FAQ Sheet on the JAR website for further details.

The fraud firm years observations, SAS and Matlab programs used in this paper are publicly available on *JAR* online supplements and datasheet webpage (https://research.chicagobooth.edu/arc/journal-of-accounting-research/online-supplements).

The file "AAER_firm_year" contains both the initial fraud firm years from CFRM and additional fraud firm years of AAERs announced after September 30, 2016 up to December 31, 2018 by hand-collection.

The file "SAS coding" shows the process of merging fraud firm years with COMPUSTATA database and prepare necessary accounting features for our prediction models.

The file "run_RUSBoost28.m" is a Matlab program to replicate the results of our best fraud detection model RUSBoost in the paper. To run this program, two additional Matlab files are required: (1) the file "data_reader.m" for reading the data, and (2) the file "evaluate.m" for evaluating model performance.

We also made the final dataset publicly available in our Github repository (https://github.com/JarFraud/FraudDetection).

The file "uscecchini28.csv" is our final dataset which contains the fraud labels and feature variables. The variable name of our fraud label is "misstate" (1 denotes fraud, and 0 denotes non-fraud). The variable names of the 28 raw financial data items are: *act, ap, at, ceq, che, cogs, csho, dlc, dltis, dltt, dp, ib, invt, ivao, ivst, lct, lt, ni, ppegt, pstk, re, rect, sale, sstk, txp, txt, xint, prcc_f*. The variable name of 14 financial ratios are: *dch_wc, ch_rsst, dch_rec, dch_inv, soft_assets, ch_cs, ch_cm, ch_roa, issue, bm, dpi, reoa, EBIT, ch_fcf*. The variable *new_p_aaer* is used for identifying serial frauds as described in Section 3.3 (see the code in "RUSBoost28.m" for more details).

6. An assurance that the data and programs will be maintained by at least one author (usually the corresponding author) for at least six years, consistent with National Science Foundation guidelines.

The authors will retain the data and programs for the required six years.