Predicting Public Procurement Irregularity: An Application of Neural Networks

Ting Sun

Rutgers, The State University of New Jersey, Newark

Email: tings@rutgers.edu

Leonardo J. Sales

The Office of the Comptroller General (CGU), Brazil

Email: leonardo.sales@cgu.gov.br

Running Head: Predicting Public Procurement Irregularity: An Application of Neural Networks

Predicting Public Procurement Irregularities: An Application of Neural Networks

ABSTRACT

Using the data describing the characteristics of contractors provided by CGU (the Comptroller General of the Union, Brazil), this paper mainly implements two artificial neural networks, traditional neural network (TNN) and deep neural network (DNN), to develop prediction models of public procurement irregularities designed for initial screening of contractors. This is the first application of DNN in the context of government auditing. To examine the effectiveness of DNN, the authors compare its predictive performance to TNN and other two algorithms (Logistic Regression and Discriminant Function Analysis) and find that DNN significantly outperforms TNN and other algorithms in terms of accuracy, precision, F scores, AUC, and other metrics, as suggested by the high Z scores of Z test. Although TNN has a higher recall than DNN, the difference of recall between TNN and DNN is insignificant. Logistic Regression and Discriminant Function Analysis achieve the highest recall scores but their Z scores are much lower than the those of other metrics. Therefore, DNN generally performs more accurately than other approaches and meets the requirement of CGU for an early alarm system.

Keywords: Public procurement, Bidding, Contractor, Neural networks, Deep learning

1. INTRODUCTION

Government has the fiduciary obligation to provide goods, services, and infrastructures for social development and the welfare of citizens, such as road and rail construction, health care, and education (Odhiambo and Kamau 2003). Public procurement is the process through which government acquires the goods and services from suppliers by signing public procurement contracts. The entire process comprises a variety of activities, including "identification of public needs, framing up contract specifications, selection and solicitation of sources, preparation and award of contract and all phases of contract administration through the end of services contract or the useful life of an asset" (Thai 2008). Public procurements, especially those for developing countries with financial instability and vulnerabilities in the economic system, are thought to have a high risk of fraud and corruption with the outcome of irregularities occurred possibly at any stage of procurement (Lungeanu 2012; Mizoguchi and Quyen 2014).

Recent large-scale corruption scandals in Brazil, like the "Operation Car Wash" which involves 16 companies (including the state-run oil company Petrobras) and approximately 9.5 billion US dollars, show the potential harm to society. The corruption not only results in financial loss but also represents incomplete service or inferior quality. As a central control body of the Brazilian government, the Comptroller General of the Union (CGU) is responsible for auditing all public expenditure transactions made by the ministries and other 2,947 decentralized units. The Brazilian government (including the ministries and the decentralized governmental units) signs approximately 25,000 new contracts each year and, currently, it has 61,000 contracts in progress, with a total value of more than 232 billion dollars[1]. At present, CGU has 2,294 servants[2], among which only 1200 directly perform public procurement auditing. Even if we only consider auditing new contracts for one year, more than 20 contracts would be assigned to each auditor, which is clearly unfeasible as each public procurement is a complicated and burdensome process. On the other hand, the identification of irregularities especially those related to corruptions is challenging due to "the lack of transparency in public procurement process and the involvement of a number of actors working together for a long time" (Mizoguchi and Quyen 2014).

[1] Data source: the Brazilian's system of purchases and contracts (SIASG)
[2] Data source: the Brazilian system of personnel (SIAPE)

As a result of the heavy workload and the difficulty in irregularity detection, CGU requests an initial screening system (before the bidding process) that is able to automatically and efficiently identify highly suspicious contractors at early stage to prevent potential irregularities before the formal bidding process and to facilitate the follow-up monitoring process for identified problematic contractors. Thus, the objective of this study is to develop such a system that only uses the characteristics of the bidding company to predict the risk of public procurement irregularities such as contract default. To minimize the cost of follow-up investigations for the identified companies, the system should ensure that the identified companies are actually with high risk of public procurement irregularity. Meanwhile, since this system is designed for the initial screening process and trained with the data of company characteristics only, it needs to allow a sufficient number of contractors to participate in the bidding process. In other words, we do not expect the system to detect all possible risky contractors but try to ensure that each identified contractor is worthy of the costly follow-up investigation.

There is a favorable data environment for this research. As a control unit, CGU receives structured data of governmental spending and contracts from all transactional information systems. It also has access to governmental databases with aggregated information in other aspects of public contracts, such as the characteristics of private companies. For the purpose of combining all databases for studies in a variety of government spending related topics, CGU launches the Public Spending Observatory (Observatório da Despesa Pública) teamed with specialists in database administration, business intelligence, and data mining. With the multiple data sources provided by Public Spending Observatory, we have a wealth of data attributes (which is as many as 183 attributes) to extract insights for our research.

Innovations in Machine Learning and Artificial Intelligence (AI) empower us with an effective data analysis technique called Deep Neural Network (DNN) (or Deep Learning). Neural Networks experienced a renaissance around a decade ago as enhanced data storage capability and boosted computing power made the traditional Artificial Neural Network (it is called TNN for short in this paper) smarter, faster, more intuitive, and structured more like the real human brain. Without the burdensome data preprocessing (i.e., extracting features from numerous candidate attributes or removing the html tag for text analysis), DNN works efficiently for high-dimensional or unstructured data. Because of its excellent performance, DNN has made great achievements in natural language processing, image recognition, and many other complex data analytics tasks.

Implementing this innovative technology is appropriate for the current study because of the large number of input variables. With Deep Learning, a complex classification model can be built and tested with the large number of labeled data attributes obtained from the multiple governmental databases. The input data originated from 7 CGU datasets contains information describing the bidding company's characteristics from 2011 to 2014. The label variable is "Penalty", which equals 1 if the contractor receives at least one penalty from 2015 to 2016 due to the serious irregularity. This study removes bidding companies with public contracts priced less than R$ 1 million as of 2014 because CGU focuses on companies using relatively large amount of public resources. This leads to 10,186 records in our data, among which 744 are positives (PENALTY=1) and 9,442 are negatives (PENALTY=0). 80% of the data is assigned as training set while the remaining 20% forms the test set. The proposed classification model (DNN) attempts to identify the data pattern underlying the training set with 5-fold validation and the out-of-sample test will be carried out for the model using the test set. Using the classification system, the new contractor can be evaluated to suggest if it is a risky company that could execute the contract improperly. In addition, for comparison purpose, we apply a TNN algorithm, multilayer perceptron (MLP), to predict the irregularity. As an additional analysis, we also employ Logistic Regression and Discriminant Function Analysis to detect irregularity risks. We find that DNN outperforms TNN and other two algorithms in terms of accuracy, precision, AUC, and etc. We also use Z test to examine the statistical significance of the differences between the value of these measures, and the results show that they are all significant at 0.01 level. While TNN is slightly better than DNN in terms of recall, the difference is not significant. Logistic Regression and Discriminant Function Analysis have the highest recall, but the Z scores are much lower than those of other measures when DNN performs better. Furthermore, when considering the joint effect of precision and recall, measured by F scores ($F_1, F_2, and\ F_{0.5}$), DNN significantly performs better than TNN and other algorithms.

This is the first application of Deep Learning to government auditing. It contributes to existing literature in similar topic by establishing a risk monitoring system based on the characteristics of the contractor. This system proactively predicts potential irregularities before the bidding and execution of contracts, which provides recommendations for government auditors to select auditees and for the government decision of contract awarding.

The paper is organized as follows. Section 2 reviews the prior research regarding the critical issue in public procurement, the role of auditing in detecting and monitoring irregularities, and the effort made for the identification of irregularity indicators. Section 3 presents the research methodology including the data analysis method, the data, and the modeling process. It is followed by the prediction result of four classification models. The concluding section summarizes the present work, highlights its contribution to the literature, and provides the direction for future work.

2. PRIOR LITERATURE

- Challenges in public procurement

In a public procurement, the bureaucrat uses public funds to obtain the goods and services provided by the supplier. The supplier must execute the contracted work in accordance with the requirement of the contract. It is the public fund that makes the public procurement different from private acquisition and other related activities. Specifically, to ensure the efficient use of public fund, public procurement should include a bidding process of supplier selection in which only the company that offers the best price and quality wins (Patras and Banacu 2016). Therefore, greater focus should be put on the selection of bidders before or during the bidding process before the contract is awarded to a bidding company that has conflicts of interest or high risk of fraud and other irregularities.

Corruption is perceived as one of the greatest barriers to efficient and sustainable economic and social development, especially for developing countries (Wade 1982; Anderson, Kovacic and Müller 2011). Corruption is the public official's act of abusing the entrusted power in exchange of personal gain. In the context of public procurement, the perpetrator, instead of acting for the public interest, may award the contract to an unqualified supplier who provides reciprocal benefits such as bribes (Anderson, Kovacic and Müller 2011). Corruption is also a major cause to various public irregularities. Nag (2015) investigates corruption in the Indian public procurement sector and suggests that irregularity occurs when the execution of contract departs from the standard operating procedure of procurement and/or when the interest of the public agency or the supplier is compromised. For instance, the contractor provides goods of a lower quality than specified in the contract or does not complete the contracted goods or services by the due date (OECD, 2007). For the same country, Tabish and Jha (2011) provide a comprehensive list of 61 irregularities derived from prior literature, cases in the technical vigilance inspection reports of the Chief Technical Examiner's Organization (CTEO) of India, and interviews with key officials involved in the technical vigilance inspection. Examples of those irregularities include "administrative approval and financial sanction not taken to execute the work", "work is not executed for the same purpose for which the sanction was accorded", "some components are repeated in more than one item", "the reimbursement of service tax, excise duty, etc. is not done after obtaining the actual proof of depositing the same", and "stipulated prequalification criteria for selection of contractor are stringent". A questionnaire based on the 61 irregularities is prepared and completed by 8 chief technical examiners of the CTEO. The participants are requested to rate the likelihood of occurrence on average using a six-point scale for each of the irregularities. The survey reports top 15 most frequent irregularities, among which the top 3 irregularities are "realistic technically sound estimates are not prepared", "the consultant is not appointed after proper publicity and open competition", and "the provisions are not made for payment to consultant for part performance or repetitive work". Furthermore, it classifies the 61 irregularities under five categories: (1) transparency irregularities; (2) professional standards irregularities; (3) fairness irregularities; (4) contract monitoring and regulation irregularities; (5) procedural irregularities. Finally, it suggests the policy-makers to use the five categories of irregularity to develop preventive strategies to combat corruption.

Patras and Banacu (2016) propose a model to describe the public procurement process and indicate that there are four stages throughout the entire public procurement process: need assessment and definition, process design, evaluation, and contract implementation. They highlight a variety of irregularities mainly for the stage of process design and evaluation, based on 54 decisions of National Council of Solving Complains in Romania regarding the complaints made by bidders against the outcome of the procurement procedure in 2015. For instance, one of the irregularities for the process design process is "unclear or unjustified award criteria".

- Role of government auditing

Audit authorities play an important role in combating corruption and detecting irregularities. Lungeanu (2011) asserts that the objective of the audit of public procurement is to ensure integrity in the bidding process and to examine the effectiveness of the internal control of the bidding company for the proper execution of the contracted work. Audit authorities aim to provide sufficient, relevant, and reliable evidences to support their opinion on whether the statement of expenditure is fair and the transaction follows the contract and the regulation. Consequently, auditors usually need to conduct extensive substantive procedures to detect public procurement irregularities. Aquino, Ângelo, and Cardoso (2017) analyze 728 public procurement contracts involving 228 federal public works that are audited by TCU (The Brazilian Court of Audit) to investigate whether TCU's monitoring helps reducing overpricing (that is, the actual cost exceeds the price quoted) and the contract irregularities, including the misevaluation of the cost for contract execution and, the postponing the due date. Their findings show that TCU's efforts are not sufficient to discourage the misevaluation of the execution cost, but they do seem to avoid deadline postponing.

The detection of public procurement irregularities is challenging because "there is often neither a clear perpetrator nor a victim, rather a group of individuals in collusion with common interests in maintaining secrecy around the corrupt acts" (OECD 2005; Nag 2015). Therefore, solely relying on classical audit procedures like inspection, observation, inquiry, and confirmation is not effective and efficient to identify irregularities (Lungeanu 2011).

- Irregularity indicators

A growing number of studies aim to find indicators of irregularities based on historical or survey data and a vast majority of them are in conjunction with corruption. For example, Le, Shan, Chan, and Hu (2014), using principal component analysis (PCA), investigate the relationships between two causes of corruption, including the deregulation in the public construction sector (hereafter, the cause is called "regulation" for short) and the lack of a positive industrial climate (hereafter, the cause is called "climate" for short), and five types of vulnerabilities in construction. They find that regulation is more likely to cause vulnerabilities than climate. Among all factors under the category of regulation, the most influential one is negative leader roles (that is, the leader of an organization engages in or overlooks corruptions), followed by inadequate sanctions (that is, the government authority does not impose significant sanctions on corrupt crimes), lack of rigorous supervision, and multifarious licenses and permits (that is, the contractor holds several compulsory licenses and permits from government agencies). With similar approach, Shan, Chan, Le, and Hu (2015) examine four response strategies for corruption vulnerabilities in the public construction sector that were raised by Tabish and Jha (2012), namely leadership, rules and regulations, training, and sanctions). Leadership involves how the leader of an organization communicates values of integrity to the staff and creates an environment motivating people to behave honestly. Rules and regulations are designed to assist an organization to implement its mission and vision of anti-corruption policies. Training is provided by an organization to equip practitioners with necessary knowledge and skills about corruptions, their damaging effects on the society, and how to manage the risk of corruption. Finally, sanctions are the punishment imposed for the detected corruptions. The authors conduct a questionnaire survey to collect opinion-based data from target respondents in the context of Chinese public construction sector. The respondents are public construction practitioners (including designers, consultants, and governmental officials) and academics from 8 institutions, which are all active players in the Chinese public sector. It is shown that these response strategies, except leadership, are not significantly effective in preventing corruptions. Shan, Chan, Le, Xia, and Hu (2015) focus on the measurement of corruption and develop a fuzzy measurement model based on 24 measurement items and their constructs. They claim that the model facilitates the assessment and monitoring of the potential corruption in public construction projects.

As an extension to prior research, Shan, Le, Yiu, Chan, and Hu (2016) use a hybrid approach to explore the underlying factors of corruption for Chinese public construction market. The research conducts 14 structured interviews to obtain information regarding irregularities due to corruption. Then it develops a questionnaire survey for 188 professionals from a Chinese public construction sector to obtain opinion data on the irregularities. Next, a PCA is used to create corruption factors based on the data collected from last step. Five factors named “immorality”, “unfairness”, “opacity”, “procedural violation”, and “contractual violation” are successfully extracted, with which the research develops two statistical models (stepwise multiple regression and partial least squares structural equation) to do the data analysis and finds that the most influential factor is immorality, which is followed by opacity, unfairness, procedural violation, and contractual violation. Finally, the results are used to develop a case study.

Fazekas, Tóth, and King (2016) develop a composite indicator of corruption. They first extract data from Hungarian public procurement announcements of 2009–2012 and identify corruption red flags based on their literature review and fieldwork with public procurement experts. They use a multiple regression linking those red flags to likely corruption outcomes (including restricted competition and recurrent contract award to the same company). The coefficients of regression represent the strength of association between the input variable and the corruption. Next, they assign a weight between 0 to 1 to each powerful corruption input based on the coefficient and the authors' understanding of how each input leads to outcomes. The corruption risk composite indictor of individual transaction is obtained using the weight associated with the elementary indicator. The result shows that, compared to firms with lower corruption scores, high risk firms have relatively higher profitability, higher ratio of contract value to initial estimated price, greater likelihood of politicians managing or owning them and greater likelihood of registration in tax havens.

To summarize, prior efforts are mainly placed on corruption related irregularities, and researchers usually perform statistical analysis based on specialists' opinion data. Limited research attempts to predict the possibility of public procurement irregularities using data related to the characteristics of the contractor. There is an urgent need to monitor and prevent malpractices in public procurement projects. Auditors should rethink innovative approaches of Machine Learning and AI to predict the risk of irregularity before or during the bidding process of public procurement contract. This paper proposes a proactive measure which could greatly improve the effectiveness and efficiency of public procurement compliance auditing as it, in the first place, directs the attention of the auditor to the contractor who has a high risk of irregularity and helps preparing the appropriate audit plan for the targeted auditee.

3. RESEARCH METHODOLGY

- Method

Artificial Neural Network (ANN)

Artificial neural networks are a set of Machine Learning algorithms modeled loosely after the biological neural networks in the human brain (DeepLearning4J, 2017). The objective of developing a neural network is to identify patterns underlying the data and use them to predict the feature of future unseen data. Figure 1 shows the structure of a simple example of TNN. This neural network consists of one input layer (the leftmost layer), one output layer (the rightmost layer), two hidden layers (the middle layers between the input layer and the output layer). Within each layer, there are a number of neurons where complex computation takes place. In the leftmost input layer, the neuron (also called input neuron) receives the raw input training data, modifies it with linear and/or non-linear transformations, and passes the output data to the next layer (the first hidden layer); the first hidden layer transmits the data received by its neurons on a modified version to the second hidden layer. Similarly, the data is transferred to the last layer which provides the final output (Sun and Vasarhelyi 2017).

INSERT FIGURE 1 HERE

Unlike a TNN that has one or two hidden layer(s), a DNN consists of multiple hidden layers made of more neurons (Deng and Yu, 2014) (see figure 2), allowing the algorithm to execute more complex linear and non-linear transformations. One of the earliest DNN has three densely connected hidden layers (Hinton, Osindero, and Teh 2006), and a neural network with more than 10 layers is considered "very deep" (Schmidhuber 2015). In a model training process, the output value is compared to the actual value of the training data and the error will be calculated. The parameter that used to perform data transformation for model training is adjusted to reduce the error. Such process will be repeated thousands or even billions of times to refine the parameter to minimize the error. It is noteworthy that the training process has no human intervention. In other words, the model “learns” from the data on its own (Sun and Vasarhelyi, 2017).

INSERT FIGURE 2 HERE

Due to the depth of the architecture, DNN is powerful in processing complex datasets, especially semi-structured data (i.e., text) or unstructured data (i.e., sensory data like image or sound) (Najafabadi et al., 2015; Chen and Lin, 2014; Haber and Ruthotto, 2017). In a DNN, as the labeling or clustering raw data is transformed from one layer to the next, each successive layer uses features extracted by the previous layer to form more advanced and more complex features (Sun and Vasarhelyi 2017; Najafabadi et al. 2015). As an illustration, trained with massive volumes of face images, a deep learning system aimed to recognize human faces first extract the most essential element of the raw data, pixels. Then, as the layer goes to the next level, it can recognize more complex elements, edges of pixels. After that, the edges are transmitted to the next layer to form a set of combinations of edges that resemble things like eyes and noses. This is followed by another layer that is able to recognize initial face models, and after many layers of data analysis the system finally recognizes faces (Akagi, 2014; Sun and Vasarhelyi 2017).

The application of DNN has led to many complicated breakthroughs in a variety of areas and the Big Four accounting firms are implementing this technology to perform a number of auditing procedures or tasks, such as reviewing contracts (Deloitte) and assessing the risk of bank loans (KPMG) (Kepes 2016). Deep learning is a stable learning algorithm. In other words, the output of a deep learning model does not change much when the training data is modified slightly, such as leaving out a small number of samples. This leads to better generalization, which is the ability of an algorithm to accurately predict outcome values for future unseen data. Another advantage of deep learning is its scalability. It can efficiently learn patterns from large volumes of data, like hundreds of millions of documents (Saxe 2017; Spring and Shrivastava 2016). Due to the training stability, generalization, and scalability (Candel, LeDell, Parmar, and Arora 2016), we choose deep learning as a classification algorithm for the current study considering the data size and the number of data attributes.

- Data

To form our data, we combine 7 CGU datasets, including RAIS, RFB, SIAFI, SIASG, SIAPE, TSE, and SICONV, reflecting 5 risk dimensions: Operational Capabilities, Profile of Participation in Biddings, History of Punishments and Findings, Conflict of Interests, and Political Bonding. Table 1 summarizes the datasets and relevant risk dimensions.

Operational Capabilities

Operational Capabilities are the ability that a company's employees and partners could use public or nonpublic resources to deliver excellent goods or service to its clients. RAIS (Portuguese: Relação Anual de Informações Sociais; English: Annual Social Information Report) contains information about the employees of the bidding company, i.e., the total number of employees in the year and the average salary of the employees. RFB (Portuguese: Receita Federal do Brasil; English: Federal Revenue Office System) describes the characteristics of the partners of the bidding company, i.e., the total number of partners in the year, the average salary of the partners in the year, and the average age of the partners in the year. Characteristics, like the structure, the compensation, and the qualification, of the employee and the partner of a company affect its operational capabilities, as they reflect the internal control and management from the perspective of the personnel. Another aspect of operational capabilities is the extent to which the company has relied on the government resources. An example of attributes used in this study is the amount of public spending received by the company per employee. Relevant information is provided by SIAFI (Portuguese: Sistema Integrado de Administração Financeira; English: Integrated System of Financial Administration), the transaction processing dataset for the spending of government bodies. The dataset contains information about the purpose of the spending, the party that receives the payment, and related bidding process. Irregularities are more likely to take place in a company with operational deficiencies (i.e., insufficient employees, low salary of the employees, and overreliance on public funds), as they provide incentives and opportunities for irregularities, especially frauds or corruptions (Luo 2002; Luo, 2007).

Profile of Participation in Biddings

SIASG (Portuguese: Sistema Integrado de Administração de Serviços Gerais; English: System of Integrated Administration and General Services) describes historical biddings of the company, such as the total number of biddings disputed in the year, the total value proposed by the company for all disputed biddings in the year, and the percentage of the number of awarded biddings within all disputed biddings. This dataset characterized the company's behavior as a bidder, which is insightful for the evaluation of future biddings. For example, compared to the company that always win in the bidding process, the company with lower percentage of awarded contracts is generally less qualified to conduct a public procurement contract and has more incentives to commit fraud to maximize its benefits once a contract is awarded to the company.

History of Punishments and Findings:

SIASG also contains information about the previous light penalties (i.e., fines and warnings) received by the company due to less severe irregularities, such as delayed delivery of goods. Examples of risk indicators include the quantity of received punishments and the number of alerts generated by CGU during its monitoring process.

Conflict of Interests:

SIAPE (Portuguese: Sistema Integrado de Administração de Recursos Humanos; English: Civil Servants Integrated Administration System) is the system controlling the public servant. This dataset can be used to examine the relationship between the company and the federal agency. A risk indicator is the number of company partners who are also public servants. SICONV (Portuguese: Sistema de Convênios; English: Federal Government Agreements Management System) provides agreements showing the incidence that the bidding company, as a NGO (Non-governmental organization), receives money from the government or the company contracts with a NGO who receives money from the government. This is a conflict of interests because the company acts as a for-profit corporation to participate in the bidding process of a public contract, whereas the same company receives fund from the government as a NGO.

Political Bonding

TSE (Portuguese: Tribunal Superior Eleitoral; English: System of Brazilian Electoral Court) describes the company's involvement in elections. The politically connected bidders have more opportunities to collude with government officials. A risk indicator is the total amount donated by the company to support the candidates (or parties) in political campaigns.

INSERT TABLE 1 HERE

Our dataset contains 183 variables[3]. The target variable is “Penalty”, which equals 1 if the bidding company receives at least one severe penalty[4] from 2015 to 2016 due to the serious irregularity. A serious irregularity occurs when the contracted work is not executed in accordance with the contracted specifications. Examples of those serious irregularities include rendering of fictitious work, inflating the work volume, supplying goods or services of a lower quality than quoted. Our sample contains 10,186 observations, among which 744 are positives (PENALTY=1) and 9442 are negatives (PENALTY=0). Table 2 displays the distribution of the frequency of the penalties received by the bidding companies. The data is randomly separated into training set and test set, with 80% of the population as training set and the remaining 20% as the test set in each modeling process. We use 5-fold cross-validation when training the model, and the trained model is tested with the test data. We have severe data imbalance issue as negatives are much more than positives. We use over-sampling method to randomly replicate positives to balance the data for the training set to avoid biased predictions and misleading accuracies which are caused by the fact that the algorithm cannot obtain the necessary information about the minority class due to the data imbalance. We do not increase the positives for the test dataset as we try to validate the predictive performance of our models for a real life imbalanced data.

INSERT TABLE 2 HERE

In this study, we mainly employ two approaches: TNN vs. DNN. Both approaches use all 183 data fields. We compare the out-of-sample test results for both approaches.

[3] Due to the length limit, the full list of the 183 variables is not reported in this paper

[4] The severe penalty is the termination of the current contract or/and the prohibition of the bidding company for future public biddings. To reduce bias, we only consider those companies received at least 1 penalty for the first time from 2015 to 2016 and discard those companies that received penalties before 2015.

- Model Training

The TNN is trained with Multilayer Perceptron (MLP) algorithm. Table 3 is the coincidence matrix for training dataset. After the over-sampling, the training data has 15,119 observations. 946 out of 7559 negatives are identified as positives while 485 out of 7560 positives are identified as negatives.

INSERT TABLE 3 HERE

In order to train a neural network that best fit our data, we use the "RandomDiscrete" strategy to randomly search for all the combinations of the hyperparameters for model training. We then employ 5-fold cross-validation and train our neural networks with different hyperparameters on the training set and compare validation error on the validation set via a grid search. The grid search is a brute-force exhaustive search paradigm that is commonly used for parameters tuning. It evaluates the model performance for each possible combination of the hyperparameters and search for the one with improved accuracy (Muller and Guido, 2016). Specifically, we choose the model with hyperparameters at the grid point with the lowest validation error (Logarithmic Loss) based on the 5-fold cross-validation. This process trains 47 neural network models and leads to the choice of the deep neural network with five learned fully connected layers (one input layer, three hidden layers, and one output layer) whose Logarithmic Loss is 0.2045. The input layer has 222 neurons. The first of the three hidden layers contains 175 neurons, the second hidden layer contains 350, and the third one has 150 neurons. The final output layer contains 2 output neurons, which is the classification result of this research (whether or not the company has irregularities). Table 4 is the coincidence matrix of DNN for the training data. 1834 out of 7525 negatives are identified as positives, while 1048 out of 7537 positives are identified as negatives.

INSERT TABLE 4 HERE

Furthermore, we employ other two approaches, Logistic Regression and Discriminant Function Analysis, to conduct the same task. Logistic Regression is one of the most frequently used statistical technique for classifying records based on values of input fields. Discriminant Function Analysis is similar to Logistic regression except that it has more stringent assumptions and can be used as an alternative to Logistic Regression. Together with TNN, proposed models with these two approaches are analyzed and the compared to DNN.

4. RESULTS

- Predictive performance

Table 5 reports the predictive performance of the TNN, DNN, Logistic Regression, and Discriminant Function Analysis measured by a set of metrics for the test data. The overall accuracy of the DNN is 0.9157, the highest among all models. The second highest one is produced by TNN, 0.8383, followed by 0.7948 and 0.7789 generated by Discriminant Function Analysis and Logistic Regression, respectively. Similarly, we observe that DNN also has the largest number of True Negatives, which is 1803, while Logistic Regression has the smallest number of True Negatives, which is 1479. Additionally, the number of false positives of DNN is the fewest, 114, while TNN has 267 false positives. Nevertheless, compared to TNN and the other two algorithms, DNN is less powerful for identifying all positives. Its true positives are the fewest (which are 76), and the highest is 100, produced by Discriminant. Similarly, DNN has more false negatives than Logistic Regression and Discriminant, but less than TNN.

Specificity[5] is the true negative rate measuring the portion of identified negatives in all actual negatives. In our case, it is the percentage of the model identified normal contractors in observed normal ones. DNN's specificity is 0.9405, the highest among all models, followed by TNN, Discriminant and Logistic.

Precision and recall are two measures for the ability of the classifier for irregularities detection, where precision[6] measures the percentage of actual irregularities in all perceived irregularities and recall[7] indicates that, for all actual irregularities, how many of them are successfully identified by the classifier. We find that the recall of TNN is slightly higher than that of DNN, and Discriminant has the highest recall followed by Logistic Regression. However, it is observed that DNN has the highest precision, which is 0.4. This number is much higher than the second largest precision made by TNN, which is only 0.2283. To consider the effect of both precision and recall on the predictive accuracy of the model, we employ three F scores, F_1, F_2, and $F_{0.5}$, which have been frequently used by existing data mining research. The F_1 score[8] is the harmonic mean of precision and recall, which treats precision and recall equally. While F_2[9] treats recall with more importance than precision by weighting recall higher than precision, $F_{0.5}$[10] weighs recall lower than precision. We found that DNN has the greatest value for all F scores, suggesting that DNN has the best predictive accuracy, considering both precision and recall. The $F_{0.5}$ for DNN approach is 0.5634, higher than that of TNN approach (0.2795).

[5] Specificity = true negative/ (true negative + false positive)
[6] Precision = true positive/ (true positive + false positive)
[7] Recall = true positive / (true positive +false negative)
[8] F_1 =2 × (precision × recall) / (precision + recall)
[9] F_2 =5 × (precision × recall)/ (4×precision + recall)
[10] $F_{0.5} = \frac{5}{4}$ × (precision × recall) / ($\frac{1}{4}$ ×precision + recall)

Finally, we examine the Area Under the Receiver Operating Characteristic (ROC) curve, called AUC, as it provides a general evaluation of the model. Our classifiers generate estimates of the probability that the input data belongs to the positive (irregularity) class. The ROC curve plots the true positive rate ("recall") versus the false positive rate[11] as the discriminative threshold is varied between 0 and 1. A higher AUC suggests an improved ability of the model to discern between the two classes. The highest AUC is 0.878, produced by DNN. This emphasizes the superiority of DNN as opposed to TNN, Logistic Regression, and Discriminant Analysis.

INSERT TABLE 5 HERE

- Z Test

Based on the predictive results, we note that while DNN performs more powerful than other models in terms of the overall accuracy, precision, F scores, AUC, and etc., it is less effective in identifying all the positives. It seems difficult to determine the performance of DNN as compared to TNN and other two approaches. Therefore, we follow O'Leary (1998) as well as Coats and Fant (1993) and use the normally distributed Z test of equality of proportions to examine the probability that there is a difference between the key predictive measures using DNN and those using TNN as well as Logistic Regression and Discriminant Function Analysis. The key predictive measures include the overall accuracy, AUC, Type I hit, and Type II hit, where Type I hit is the true negative rate which is used for companies without irregularities, and type II hit is the true positive rate which is used for companies with irregularities. Other measures (i.e., precision and recall) have similar result, as they are derived from these key measures.

[11] False positive rate = false positive/ (false positive + true negative)

We first compare the differences of performance metrics between DNN and TNN (see table 6). The null hypothesis is that the proportion of hits, including the overall accuracy, AUC, Type II hit, and Type I hit, for the TNN model is greater than or equal to the proportion of hits for the DNN model. It is shown in table 6 that the overall accuracy, AUC, and type I hit of DNN are significantly greater than those of TNN at the significance level of 0.01. Although the type II hit of DNN is smaller than that of TNN, the difference is insignificant.

INSERT TABLE 6 HERE

Table 7 shows the result of the comparison between DNN and Logistic Regression. We found that except for type II hit, all measures of DNN are significantly better than those of Logistic Regression at the level of 0.01. Despite Logistic Regression's type II hit is significantly higher than that of DNN at the level of 0.01, its Z score (-5.14) is much smaller than the other three Z scores of AUC (11.78), accuracy (16.94), and type I hit (31.81).

INSERT TABLE 7 HERE

Similar pattern is shown for the comparison of DNN and Discriminant (See table 8). The Z score for the case of type II hit is only -6.38 as opposed to 29.33 for type I hit, 20.92 for accuracy, or 16.15 for AUC.

INSERT TABLE 8 HERE

To summarize, DNN performs better than other three algorithms from the aspects of the overall accuracy, F scores, AUC, and etc. Although it is relatively less effective for the task of detecting all possible positives, the difference of such capability between DNN and TNN is insignificant. When compared to Logistic Regression and Discriminant Function Analysis, the difference is significant but with a much smaller magnitude as opposed to its counterparts when DNN outperforms other models.

As an early warning system of irregularity risk for public procurement competitors, the objective of this system is to ensure all identified contractors actually have high risk of irregularity rather than to identify all possible risky contractors. This is because CGU performs costly follow-up investigations and extensively substantial tests for all identified suspicious companies and tries to minimize such cost. In addition, the system is designed for the initial screening of the candidates before they formally start bidding. At this stage, the data about the contract bidding and execution is not yet available. All the data used in this study is about the characteristics of the candidates themselves. It cannot and is unnecessary to provide complete information for the assessment of the irregularity risk. Therefore, it is unrealistic to expect this system to identify all of the actual positives.

5. CONCLUSIONS

This paper mainly employs two neural network algorithms, multilayer perceptron (MLP) of TNN and DNN, to establish prediction models for irregularities in public procurement, using the data of the contractor's characteristics provided by CGU. As a developing country, Brazil has been suffering from frequently occurred frauds, corruptions, and other irregularities in public procurement. The OECD (Organization for Economic Cooperation and Development) Foreign Bribery Report (2014) shows that 57% of international bribery cases occur in public procurement contracts. This is a timely study that meets CGU's requirement for an irregularity risk prediction system designed for the initial screening of bidding companies before they enter the bidding process. Specifically, this system can rule out the inherently risky bidding companies by only analyzing the company's characteristics. This feature is important as the audit authority does not have access to the data of bidding process or contract execution at the initial screening stage. In addition, this paper applies another two frequently used algorithms for classification problems, Logistic Regression and Discriminant Function Analysis, and compare their performance to that of DNN.

The prediction result shows that, compared to other models, DNN has much less false positives but slightly more false negatives. With the DNN model, our goal is to ensure that all identified companies are actually risky and worthy of the costly follow-up investigations, rather than blocking all possible risky companies since only company characteristics are used to train the model. Furthermore, besides the overall accuracy, AUC, and so forth, DNN outperforms other algorithms in terms of all F scores, suggesting that DNN has better predictive accuracy when considering precision and recall together.

Finally, this study examines the statistical significance of the difference of prediction power between DNN and other approaches. The Z test suggests that DNN is significantly more powerful than other algorithms in terms of accuracy, AUC, and type I hit. With regards to the type II hit, TNN is more effective than DNN, but the difference of the type II hit between TNN and DNN is insignificant. For the same metric, Discriminant Function Analysis has the best result and the difference is significant, but the Z score (which is 6.38) is much lower than those of other metrics when DNN performs better.

Future research will be conducted to integrate the result of the current study into a compound risk scoring system not only using company characteristics but also analyzing the contract bidding and execution characteristics.

References

Akagi D. 2014. A primer on deep learning. DataRobot. https://www.datarobot.com/blog/a-primer-on-deep-learning/

Anderson, R.D., Kovacic, W.E. and Müller, A.C., 2011. Ensuring integrity and competition in public procurement markets: a dual challenge for good governance. *Arrowsmith and Anderson (2011)*, 681-718.

Aquino, A.C., Ângelo, D.I. and Cardoso, R.L., 2017. Public works bidding process (procurement): the case of the Brazilian Court of Audit. *Working paper*.

Candel, A., LeDell, E., Parmar, V., and Arora, 2016. A. Deep Learning with H2O. H2O.ai, Inc. Fifth edition. http://www.webpages.uidaho.edu/~stevel/504/Deep%20Learning%20with%20H2O.pdf

Chen, X. W., and Lin, X. 2014. Big data deep learning: challenges and perspectives. *IEEE access*, 2, 514-525.

Coats, P.K. and Fant, L.F., 1993. Recognizing financial distress patterns using a neural network tool. *Financial management*, 22(3), 142-155.

DeepLearning4J. 2017. Introduction to Deep Neural Networks. https://deeplearning4j.org/neuralnet-overview#define

Deng, L.; Yu, D. 2014. Deep Learning: Methods and Applications. *Foundations and Trends in Signal Processing*. 7 (3–4): 1–199. doi:10.1561/2000000039. https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/DeepLearning-NowPublishing-Vol7-SIG-039.pdf

Fazekas, M., Tóth, I.J. and King, L.P., 2016. An Objective Corruption Risk Index Using Public Procurement Data. *European Journal on Criminal Policy and Research*, 22(3), 369-397.

Haber, E., and Ruthotto, L. 2017. Stable Architectures for Deep Neural Networks. arXiv preprint arXiv:1705.03341.

Hinton, G. E., Osindero, S., and Teh, Y. W. 2006. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7), 1527-1554.

Kepes, B., Mar 16, 2016. Big four accounting firms delve into artificial intelligence. *Computerworld.* http://www.computerworld.com/article/3042536/big-data/big-four-accounting-firms-delve-into-artificial-intelligence.html

Le, Y., Shan, M., Chan, A. P. C., and Hu, Y., 2014. Investigating the causal relationships between causes of and vulnerabilities to corruption in the Chinese public construction sector. *Journal of Construction Engineering and Management*, 140(9), 05014007.

Lungeanu, E., 2012. Audit of public procurements under projects financed from structural funds. *Economy Informatics*, 12(1),108-117.

Luo, Y. 2002. Corruption and organization in Asian management systems. *Asia Pacific Journal of Management*, 19(2-3), 405-422.

Luo, Y., 2007. Guanxi and business (Vol. 5). *World Scientific*.

Mizoguchi, T., and Van Quyen, N. 2014. Corruption in Public Procurement Market. Pacific *Economic Review*, 19(5), 577-591.

Muller A.C. and Guido S. 2016.Introduction to Machine Learning with Python: A Guide for Data Scientists O'Reilly Media, Inc.

https://books.google.com/books/about/Introduction_to_Machine_Learning_with_Py.html?id=vbQlDQAAQBAJ&source=kp_cover

Nag, B., 2015. Combating Corruption in Indian Public Procurement--Some Exploratory Case Studies. *Journal of Institute of Public Enterprise, 38(1&2).* https://ssrn.com/abstract=2928667

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., and Muharemagic, E. 2015. Deep learning applications and challenges in big data analytics. *Journal of Big Data*, 2(1), 1.

Nielsen, M. A., 2015. Neural Networks and Deep Learning, *Determination Press*. http://neuralnetworksanddeeplearning.com/chap1.html

OECD. 2007. Bribery in Public Procurement: Method, Actors and Counter-measures. https://www.oecd.org/daf/anti-bribery/anti-briberyconvention/44956834.pdf

OECD, 2005. Fighting Corruption and Promoting Integrity in Public Procurement. *OECD Publishing*, Paris.

OECD, 2014. OECD Foreign Bribery Report: An Analysis of the Crime of Bribery of Foreign Public Officials, *OECD Publishing*, Paris. DOI: http://dx.doi.org/10.1787/9789264226616-en

Odhiambo, W., and Kamau, P. 2003. Public Procurement. *OECD Development Centre Working Papers*, No. 208, OECD Publishing, Paris.

O'leary, D. E. 1998. Using neural networks to predict corporate failure. *International Journal of Intelligent Systems in Accounting, Finance & Management*, 7(3), 187-197.

Patras, M. V., and Banacu, C. S. 2016. Critical Phases in the Process of Awarding Public Procurement Contracts: A Romania Case Study. *Journal of Engineering, Project, and Production Management*, 6(2), 104.

Saxe J. 2017 (June 2). We're taking a quantum leap over traditional machine learning. *Invincea endpoint security blog*. https://www.invincea.com/2017/06/were-taking-a-quantum-leap-over-traditional-machine-learning/

Schmidhuber, J. 2015. Deep Learning in Neural Networks: An Overview. *Neural Networks*. 61: 85–117. arXiv:1404.7828free to read. doi:10.1016/j.neunet.2014.09.003.

http://www2.econ.iastate.edu/tesfatsi/DeepLearningInNeuralNetworksOverview.JSchmidhuber2015.pdf

Shan, M., Chan, A. P., Le, Y., and Hu, Y. 2015. Investigating the effectiveness of response strategies for vulnerabilities to corruption in the Chinese public construction sector. *Science and engineering ethics*, 21(3), 683-705.

Shan, M., Chan, A. P., Le, Y., Xia, B., and Hu, Y. 2015. Measuring corruption in public construction projects in China. *Journal of Professional Issues in Engineering Education and Practice*, 141(4), 05015001.

Shan, M., Le, Y., Yiu, K.T., Chan, A.P. and Hu, Y., 2016. Investigating the Underlying Factors of Corruption in the Public Construction Sector: Evidence from China. *Science and Engineering Ethics*: 1-24.

Spring, R., and Shrivastava, A. 2017. Scalable and sustainable deep learning via randomized hashing. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 445-454. ACM. https://arxiv.org/abs/1602.08194

Sun, T. and Vasarhelyi, M.A., 2017. Deep Learning and the Future of Auditing: How an Evolving Technology Could Transform Analysis and Improve Judgment. *CPA Journal*. June, 24-27

Tabish, S.Z.S. and Jha, K.N., 2011. Analyses and evaluation of irregularities in public procurement in India. *Construction Management and Economics*, *29*(3): 261-274.

Tabish, S. Z. S., and Jha, K. N. 2012. The impact of anti-corruption strategies on corruption free performance in public construction projects. *Construction Management and Economics*, 30(1): 21–35.

Thai, K.V. ed., 2008. International handbook of public procurement. *CRC Press*.

Wade, R., 1982. The system of administrative and political corruption: Canal irrigation in South India. *The Journal of Development Studies*, 18(3), 287-328.

Tables and figures

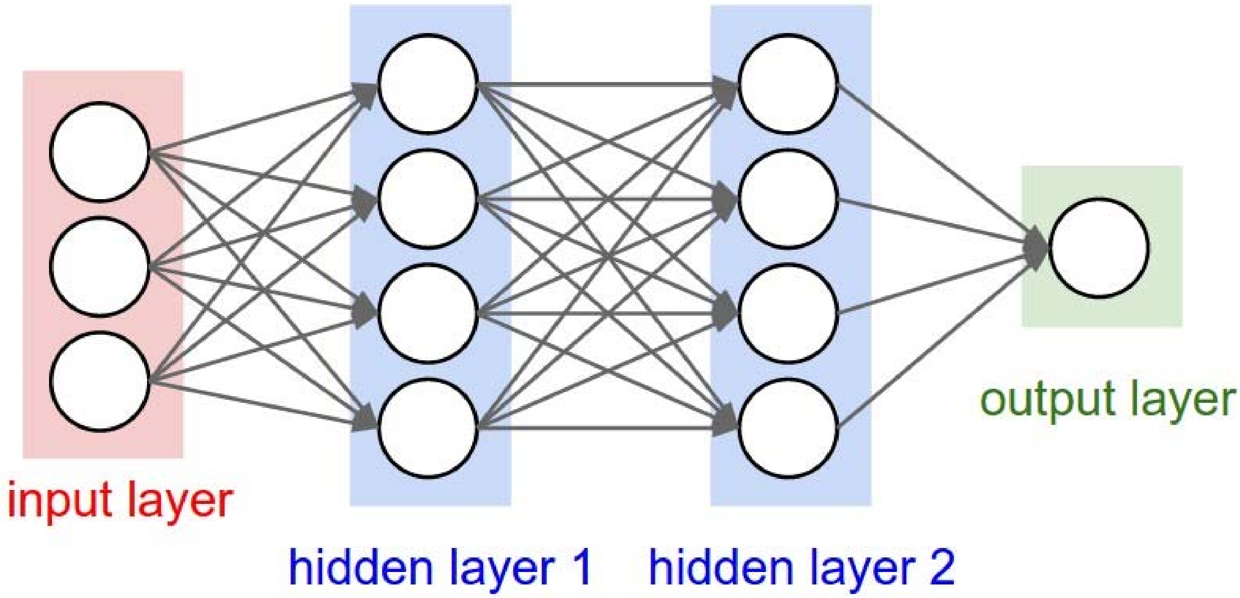

Figure 1: The structure of a traditional neural network (adapted from Nielsen 2015)

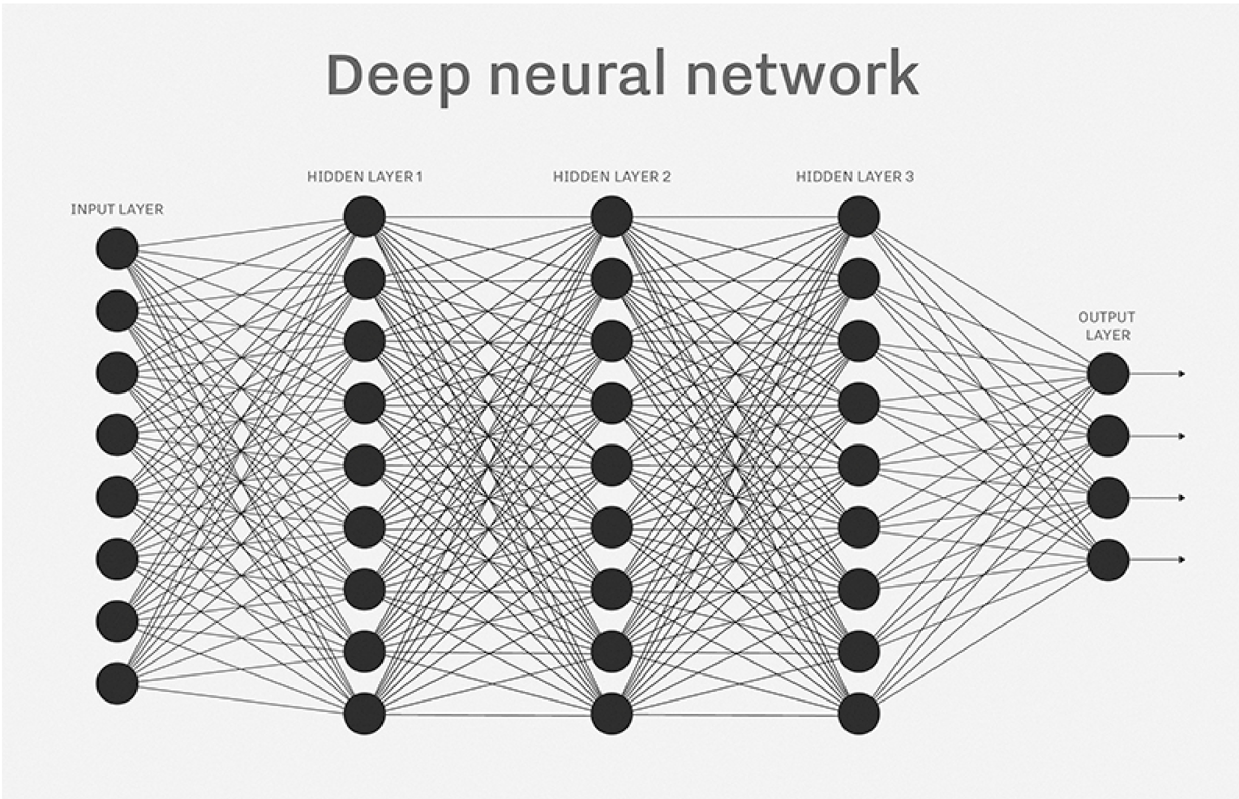

Figure 2: An example of a deep neural network (adapted from Nielsen 2015)

Table 1 The Bidding Company Datasets

Dataset Name	Description	Time period	Risk Dimension
RAIS	Employee information	2011-2014	Operational Capabilities
RFB	Partners information	2011-2014	
SIAFI	Public spending information	2011-2014	
SIASG	Historical Biddings information	2011-2014	Profile of Participation in Biddings /History of Punishments and Findings
SIAPE	Public servant information	2011-2014	Conflict of Interests
SICONV	Amount transferred to the bidding company by a non-governmental organization (NGO)	2011-2014	
TSE	Information of the involvement of the company in elections	2011-2014	Political Bonding

Table 2 The number of penalties received by the bidding companies

The number of penalties received	The number of companies
0	9442
1	574
2	164
3	6
Total	10186

Table 3 The coincidence matrix of TNN for the training data

	Identified negatives	Identified positives	Total
Negatives	6613	945	7559
Positives	485	7075	7560
Total	7095	8020	15119

Table 4 The coincidence matrix of DNN for the training data (overall accuracy=80.87%)

	Identified negatives	Identified positives	Total
Negatives	5691	1834	7525
Positives	1048	6489	7537
Total	6739	8323	15062

Table 5 Predictive performance

	DNN	TNN	Logistic Regression	Discriminant
Overall Accuracy	0.9157	0.8383	0.7789	0.7948
F_1	0.4677	0.3259	0.3005	0.3251
F_2	0.5205	0.4381	0.4545	0.4844
$F_{0.5}$	0.4246	0.2594	0.2244	0.2447
Precision	0.4000	0.2283	0.1920	0.2101
Recall	0.5630	0.5683	0.6906	0.7194
Specificity	0.9405	0.8582	0.7854	0.8003
True Negatives	1803	1616	1479	1507
True Positives	76	79	96	100
False Negatives	59	60	43	39
False Positives	114	267	404	376
False Negative Rate	0.4370	0.4317	0.3094	0.2806
False Positive Rate	0.0595	0.1418	0.2146	0.1997
AUC	0.8780	0.82	0.8190	0.817

Table 6 Z Test for differences between proportions (DNN vs. TNN)

H_0: $P_{TNN} \geq P_{DNN}$ versus H_a: $P_{TNN} < P_{DNN}$				
n	DNN	TNN	z	P value
For all firms: Percentage of hits (AUC)				
10186	87.80	82.00	11.60***	0.001
For all firms: Percentage of hits (accuracy)				
10186	91.57	77.89	27.64***	0.001
For firms with irregularities: Percentage of hits (type two hit[12])				
744	56.29	56.83	-0.21	0.4168
For firms without irregularities: Percentage of hits (type one hit[13])				
9442	94.05	85.82	18.97***	0.001

[12] true positive rate
[13] true negative rate

Table 7 Z Test for differences between proportions (DNN vs. Logistic Regression)

H_0: $P_{Logit} \geq P_{DNN}$ versus H_a: $P_{Logit} < P_{DNN}$				
n	DNN	Logit	z	P value
For all firms: Percentage of hits (AUC)				
10186	87.80	81.90	11.78***	0.001
For all firms: Percentage of hits (accuracy)				
10186	91.57	83.83	16.94***	0.001
For firms with irregularities: Percentage of hits (type two hit)				
744	56.29	69.06	-5.14***	0.001
For firms without irregularities: Percentage of hits (type one hit)				
9442	94.05	78.54	31.81***	0.001

Table 8 Z Test for differences between proportions (DNN vs. Discriminant Function Analysis)

H_0: $P_{Discriminant} \geq P_{DNN}$ versus H_a: $P_{Discriminant} < P_{DNN}$				
n	DNN	Discriminant	z	P value
Percentage of Hits (AUC) for all firms				
10186	87.80	79.48	16.15***	0.001
Percentage of Hits (accuracy) for all firms				
10186	91.57	81.70	20.92***	0.001
For firms with irregularities (type two hit: true positive rate)				
744	56.29	71.94	-6.38***	0.001
For firms without irregularities (type one hit: true negative rate)				
9442	94.05	80.03	29.33***	0.001