Mapping the Technological Landscape of Emerging Industry Value Chain Through a Patent Lens: An Integrated Framework With Deep Learning

Guannan Xu, Fang Dong, and Jiawen Feng

***Abstract*—Recent research applies patent autoclassification using machine learning to map the technological landscape of an industry value chain. However, when these methods are applied to emerging industries, the available patent sample data are small-scale and unevenly distributed, which cause overfitting and reduce the accuracy of patent classification. Therefore, this article proposes a framework to map the technological landscape of an emerging industry value chain through patent analysis with deep learning, which integrates a generative adversarial network as a data-augmentation method to overcome the problem of low-quality emerging-industry patent samples, and a deep neural network as a patent classifier. Based on this framework, this article conducts an application case of the 3-D printing industry. The evaluation results show that the integrated framework can effectively classify the patents with small-scale and unevenly distributed sample data, and depict the technological landscape of an emerging industry value chain. This article develops an efficient, reliable framework for patent autoclassification of emerging industries to overcome the lack of high-quality training samples, and it sheds light on the emerging industry value chain analysis with deep learning.**

***Index Terms*—Deep neural network (DNN), emerging industry, generative adversarial network (GAN), patent auto-lassification, value chain.**

I. Introduction

INDUSTRY value chain approach can help governments and enterprises to find out where the industrial bottlenecks are, and to develop strategies to achieve competitive advantages ([1], [2]; and [3]). Especially for a nascent industry, mapping the technological landscape of the emerging industry value chain can provide a framework for sector-specific action, from which knock-on effects up and down the industrial value chain become more apparent, and complex interdependencies can be visualized and communicated more easily [4]. Existing research has explored various methods to map the technological landscape of an industry value chain (Boehe *et al*., [5]; Noruzi *et al*., [6], [7]; and [8]), and recent attempts find that it is efficient and effective to apply patent autoclassification using machine learning [8]. However, when it is adopted to emerging industries with only a limited number of patent applications, there are problems such as small-scale sample sets and difficulty obtaining sufficient data. Thus, new methods are needed.

Traditional methods for value chain mapping are based on expert domain knowledge, such as interviews, questionnaire surveys, and the Delphi method (Boehe *et al*., [5], [9], [10]). These methods are prone to subjective bias due to the solidification of expert knowledge. In recent years, the use of bibliometric techniques to classify patents to map the technological landscape of an industry value chain has become a new research trend, including international patent classification (IPC) analysis and patent citation analysis (Noruzi *et al*., [6] and [11]). However, as to an emerging industry, the IPC code-based analysis is usually inapplicable. On the one hand, the IPC coding is generally based on the classification of traditional industries, and it is hard to find corresponding codes for emerging industries. On the other hand, patent citation analysis is also inefficient due to the short period of industrial development.

As machine learning can help to reveal hidden information and relationships in patent data to reduce the reliance on expert domain knowledge and manual analysis, it has been used to identify knowledge domains in recent years ([7], [8], and Islam, [12]). Machine learning can be classified into supervised and unsupervised learning models, according to whether external

Manuscript received 22 June 2020; revised 23 November 2020; accepted 29 November 2020. Date of publication 24 December 2020; date of current version 1 November 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 71872019, Grant 71974107, Grant 91646102, Grant L1924062, Grant L1824040, Grant L1924058, Grant L1824039, and Grant L1724034, in part by the Beijing Natural Science Foundation under Grant 9182013, in part by the Beijing Social Science Foundation under Grant 17GLC058, in part by the Fundamental Research Funds for the Central Universities under Grant 2018XKJC04, in part by the Ministry of Education in China Project of Humanities and Social Sciences under Grant 16JDGC011, in part by the CAE Advisory Project "Research on the strategy of Manufacturing Power towards 2035" under Grant 2019-ZD-9, in part by the National Science and Technology Major Project "High-end Numerical Control and Fundamental Manufacturing Equipment" under Grant 2016ZX04005002, in part by the Chinese Academy of Engineering's China Knowledge Centre for Engineering Sciences an Technology Project under Grant CKCEST-2020-2-5, Grant CKCEST-2019-2-13, Grant CKCEST-2018-1-13, Grant CKCEST-2017-1-10, and Grant CKCEST-2015-4-2, in part by the UK-China Industry Academia Partnership Programme under Grant UK-CIAPP\260, and in part by the Volvo-supported Green Economy and Sustainable Development Tsinghua University under Grant 20153000181. Review of this manuscript was arranged by Department Editor F. Tietze. *(Corresponding author: Fang Dong.)*

Guannan Xu and Jiawen Feng are with the School of Economics and Management, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: xuguannan@139.com; carmenfung.kk@foxmail.com).

Fang Dong is with the School of Public Policy and Management, Tsinghua University, Beijing 100084, China (e-mail: dongfang199310@126com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TEM.2020.3041933.

Digital Object Identifier 10.1109/TEM.2020.3041933

labeled data is required [13]. In addition to existing methods mainly based on unsupervised learning models ([14] and Rodriguez *et al*., [15]), patent autoclassification based on supervised learning models can help to analyze the patent distribution for mapping the technological landscape of industry value chains more conveniently and efficiently. Patent autoclassification requires a supervised learning model to discover implicit relationships between patent features and categories. However, supervised learning models have two problems in patent analysis in emerging industries. First, the patent dataset is small in an emerging industry, which lead to overfitting of the model; hence, it cannot obtain an effective training model. Second, the patents in different segments along the emerging industry value chain are unevenly distributed (or called imbalanced), which causes imbalanced labeled samples and reduces the accuracy of the patent classification.

Considering the above issues, with the rapid development of the artificial intelligence (AI), a generative adversarial network (GAN) based on deep learning as a powerful data augmentation method has provided opportunities to solve the problem of low-quality labeled samples. GAN is a generative model [16] introduced in 2014 by Goodfellow *et al*. [17]. A GAN consists of two deep-architecture functions for the generator and discriminator, which can learn simultaneously from trained data in an adversarial fashion, as opposed to oversampling technology [18]. In the learning process, the generator captures the potential distribution of the real data and generates synthetic samples, while the discriminator discriminates between real and synthetic samples as accurately as possible. Recent work has shown that a GAN has some successful applications, and the potential to solve the problem of small-scale and imbalanced sample data in mapping an industry value chain ([19] and [20]). After data augmentation based on a GAN, a deep learning classifier called a deep neural network (DNN) classifier can be employed in data classification. As a typical supervised learning model, deep learning has a relatively complex model structure and good performance [19], and has led to significant breakthroughs in many fields.

Hence, we propose a novel framework to map the technological landscape of an emerging industry value chain from a patent lens, integrating a data augmentation based on GAN and a classification based on DNN classifiers. This framework can effectively deal with small-scale and imbalanced sample data for emerging industry analysis. We employ the framework to depict the value chain of the three-dimensional (3-D) printing industry. First, patent data are collected from the Derwent Innovation (DI) Index patent database. Second, the latent Dirichlet allocation (LDA) algorithm is employed to extract the topics' probability distribution as features, and patents are randomly selected as the sample set and manually labeled by experts. Third, the sample set is divided into training and testing sets, and the GAN is employed to generate a large number of synthetic samples to amplify the scale of the training data. Fourth, the DNN classifiers are trained by a generated augmented training set, and the patent distribution classified by the DNN classifiers is used to understand the current technological innovation status of the 3-D printing industry. Finally, the performance of the proposed framework is evaluated by metric indicators. This article develops an efficient, reliable framework for patent autoclassification of emerging industries based on GAN and DNN to overcome the lack of high-quality training samples, and the proposed framework can provide innovative insight into emerging industries.

The rest of this article is organized as follows. Section II reviews the literature. Section III explains the research process and methodology. Section IV provides guidelines for the implementation and evaluation of our framework. Finally, Section V concludes this article.

II. Literature Review

A. Methods to Map an Industry Value Chain

Previous methods of mapping the technological landscape of an industry value chain consist mainly of expert domain knowledge, such as interviews, questionnaire surveys, and the Delphi method. For instance, Boehe *et al.* [5] identified the basic issues of a sustainable global value chain by interviewing representatives of the main players. Zhang *et al*. [10] constructed questionnaires of 342 manufacturing companies to study the relationship between industrial value chain positioning and innovation intensity. Brink [9] employed the Delphi method to forecast the global trends of the forest engineering value chain.

Thereafter, bibliometric methods are used to map the technological landscape of an industry value chain, with patent data most frequently used. As an essential form of technological intellectual property, patents are a rich source of data to study technological innovation and evolution in industries [53], [49]. In research based on bibliometric methods, IPC analysis and patent citation network analysis are frequently used. Noruzi *et al*. [6], for example, mapped patent activities based on the IPC to discover the track and change of science and technology in Iran by statistical analysis; and Li *et al*. [11] applied patent citation network analysis to discern the evolution of major roles in the industry value chain in different technological eras. However, patent data of an emerging industry are slow to acquire new IPC codes or patent citations. Thus, these bibliometric methods cannot effectively use patent data to map the technological landscape of an emerging industry's value chain.

Recently, machine learning methods have been applied to detect knowledge domains because they can discover hidden information and relationships in patent data to reduce the dependence on expert domain knowledge and manual analysis. Sakata *et al*. [7] developed a methodology to examine the structure and distribution of knowledge by citation network topology clustering and visualization. Kong *et al*. [8] used a support vector machine (SVM) based classifier to single out high-quality patents for each value chain of industrial robotics. However, in an emerging industry with a small number of patents, the effectiveness of this classification method will be significantly reduced, as there will be insufficient sample data for the machine to learn, leading to poor model expression ability. To address this problem, new innovative methods are required.

B. Methods for Data Augmentation and Classification

Previous data-augmentation studies focused on the problem of imbalanced sample data, where the number of samples in some categories is large, and in others small. A supervised machine learning model obtained through imbalanced sample data training has low classification accuracy. The solution is to expand the number of samples in categories with few samples. Chawla *et al.* [21] proposed a classical data-augmentation method, known as the synthetic minority oversampling technique, to generate synthetic samples. Derouin and Brown [22] extended an original sample set by copying samples and adding artificial noise. However, in some cases, not only is the sample data imbalanced, but the sample size in all categories is very small, and traditional data augmentation cannot solve this problem.

A new model—GAN based on AI and deep learning—can potentially solve the problem of small and imbalanced samples. The GAN contains two DNNs, consisting of a generator and discriminator. In the training process of the GAN, the generator and discriminator compete against each other. The generator generates synthetic samples according to the distribution of the real samples, the generated synthetic samples fool the discriminator as much as possible, and the discriminator distinguishes between the synthesized and real samples. When the generator and discriminator reach the Nash equilibrium, the trained GAN can be obtained. Recent studies have demonstrated successful applications of GAN in many fields. Fiore *et al*. [23], for example, used GAN to generate synthetic samples and added them to the original sample set, effectively solving the problem of data imbalance in credit card fraud identification. Hwang *et al.* [24] used a combination of GAN and an auxiliary classifier for disease identification. They compared it to classifiers that did not use GAN, and the results showed that the combination of GAN and auxiliary classifiers could achieve better classification results. Li *et al.* [25] utilized GAN to capture the relevance of dialogue and to generate corresponding synthetic text. Pascual *et al*. [26] proposed a speech-enhancement framework based on GAN. Zhou *et al*. [20] proposes a novel approach that integrates data augmentation and deep learning methods to overcome the problem of lacking training samples for forecasting emerging technologies. These successful applications of GAN provide adequate support to solve the issues of small-scale and imbalanced sample data.

In the previous patent classification studies, many supervised learning models have been employed in patent classification studies, such as K-Nearest neighbor (KNN), SVM, and Naive Bayes (NB). Larkey [27] developed a system based on a KNN model to categorize U.S. patent documents. Wu *et al*. [28] proposed an approach integrated with a hybrid genetic-based SVM model to develop a patent classification system with high classification accuracy and generalization ability. Guo *et al*. [29] proposed a patent classification approach using a NB classifier. These studies use shallow supervised learning models based on statistics; however, they require high-quality samples and features to ensure accurate prediction.

Compared to shallow supervised learning models based on statistics, deep learning has a relatively complex model structure and good performance [19] and has led to breakthroughs in many fields. Recent studies have explored the value of deep learning in bibliometrics. Li *et al*. [30], for example, proposed an effective algorithm based on deep learning to solve the multiclass patent classification problem. Trappey *et al*. [54] applied deep learning analysis methods to automatic and intelligent patent value estimation. Hassan *et al*. [31] compared deep learning and classical statistical supervised learning models in classifying the importance of a citation using the same dataset and found that deep learning has higher accuracy than SVM and RF. Zhang *et al.* [32] utilized a word embedding algorithm based on deep learning to discover the latent semantics in large-scale text. Studies in bibliometrics have shown that deep learning exhibits better performance than classical statistical supervised learning methods; thus, DNN can effectively map the technological landscape of an emerging industry value chain by patent classification.

GAN and DNN related studies provide a new solution to the problem of small and imbalanced sample data. Based on these studies, we propose to combine GAN and DNN, where GAN is used to generate a large number of synthetic samples to increase the quality of the sample, and DNN is used to map the technological landscape of an emerging industry value chain by automatically classifying patents based on these high-quality samples.

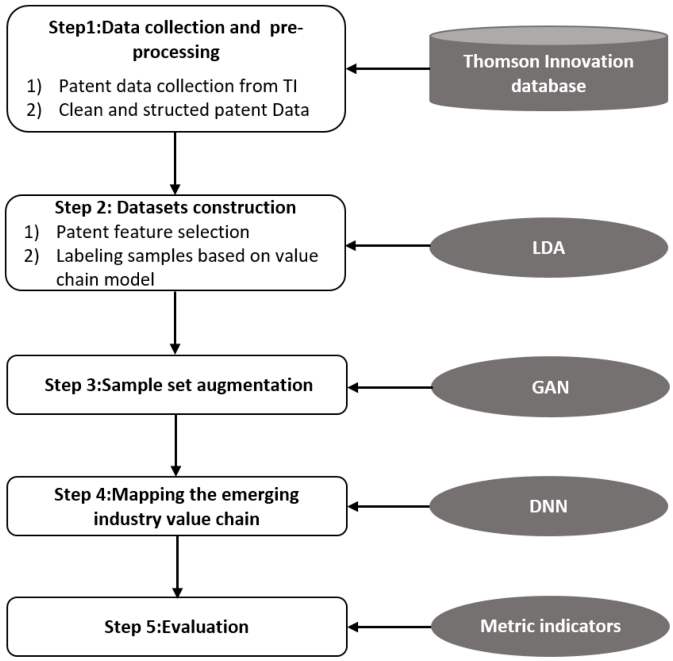

Fig. 1. Overall process of the proposed framework.

III. METHODOLOGY

A. Framework

We propose an integrated framework based on GAN and DNN to map the technological landscape of an emerging industry value chain (see Fig. 1). The framework has five steps: data

collection and pre-processing; dataset construction; data augmentation based on GAN; mapping based on DNN classifiers; and evaluating the framework by accuracy, F-measure, and G-mean.

B. Data Collection and Preprocessing

Patent data can be obtained from the patent database. Once an industry of interest is chosen, the relevant patents are collected based on the search query as an original patent set C_0. As patent autoclassification is based on textual information, in order to determine if the mining field of original patent data will affect its effectiveness. Previous research has indicated that the title and abstract fields of patent data contain the most meaningful textual information ([33] and [34]); thus, they are selected for subsequent analysis.

As to the preprocessing of the original patent data, Python and NLTK packages are used to remove useless information and store the data in a structured local database. The preprocessing of original patent data includes the following steps.

1) Remove non-English characters.
2) Remove the stop-words of the text.
3) Remove the morphological affixes from words.
4) Lower the characters.
5) Segment the sentences into words.

C. Dataset Construction

1) Feature Extraction Based On LDA: The topics' probability distribution discovered by LDA algorithm is used as the feature of each patent (Liefa *et al*., [35], [36]). Following the patent data preprecession, it is necessary to extract features from the text mining and select useful features to construct the text vectors, i.e., feature vectors. In our framework, the advanced topic model of LDA is used to extract the hidden topics as features, develop the corresponding vector space model (VSM) as feature vectors of patent text, and achieve dimension reduction compared to methods such as TF-IDF. LDA is an unsupervised machine learning technique for natural language processing (NLP) processing that is usually used to identify hidden topic information in large-scale document collections or corpuses. LDA can infer topics from a corpus based on abstracts or full text rather than keywords, without prior information on the topics. LDA uses the bag-of-words approach, which treats each document as a word-frequency vector, transforming the unstructured text data into numeric data that is compatible with data mining algorithms. Compared to other methods, such as TF-IDF, LDA has two main advantages. First, LDA artificially designs the number of topics in the document and represents the document through the topic distribution of the document (Wei *et al.*, [37]). Second, LDA can consider the semantic information of the text, instead of expressing the document only by expressing word frequency, it can reflect the semantic characteristics of the document [38].

The parameters of LDA are calculated by Gibbs sampling, and the determination of the number of topics for a corpus has always been challenging. Specifically, if the number of topics is too large, then the similarity among some topics will be relatively high. By contrast, if too small, then the content in one topic will show great differences. Moreover, the number of topics is the dimension of VSM, which impacts the quality of classification. To solve this problem, we use perplexity to determine the optimal number of topics in this article. Perplexity indicates the uncertainty of whether a document belongs to a topic. The higher the certainty of a document belonging to a topic, the lower the perplexity of the model, and the better the performance for the model [39]. LDA topic models are obtained by training them on the patent dataset with different numbers of topics, while the perplexity is calculated separately. Then the number of topics corresponding to the smaller perplexity is selected as the optimal number of topics for this patent dataset. Perplexity is defined as

$$\text{perplexity} = \exp\left\{-\left(\sum_{m-1}^{M}\sum_{n-1}^{N_m}\log\left(\sum_{k-1}^{K}p(w_n|z_k)\, p\left(z_k|d_m\right)\right)\right) \Big/ \left(\sum_{m-1}^{M} Nm\right)\right\} \quad (1)$$

where M is the number of texts in the dataset; is the total number of terms in the M_{th} text; K is the topic number; is the probability of term under topic; and is the probability of testing text under topic.

2) Labeling Samples Based on Value Chain Model: The patent samples of each functional module of the industrial value chain are labeled by experts. Consistent with existing research ([40], [41]; and [42]), first, patents are randomly selected as sample set C_1 from the original patent set C_0. Then, according to the structure of the value chain, each patent in sample set C_1 is manually labeled by two experts separately based on their domain knowledge and the patent content. Finally, it needs to be noted that if a patent is labeled as different functional modules by the two experts, the two experts will discuss to get a consensus.

To precisely evaluate the performance of mapping the technological landscape of the industry value chain, the sample set C_1 is randomly divided into a training set T_0 and a testing set T_1 at a ratio of 7:3. Due to the small scale and imbalanced training samples, we utilize a data-augmentation method based on GAN to generate a large scale and balanced synthetic samples to train the DNN classifiers and improve the accuracy.

D. Data Augmentation Based On GAN

To solve the problem of a small scale and imbalanced training samples, GAN is employed for data augmentation to enlarge the data scale of training samples in the training set T_0. After the training set is built, the training samples for each category in the training set are used to construct the GAN and generate synthetic samples for the corresponding category. The synthetic samples are used as an augmented training set T_2 for subsequent classification tasks.

GAN consists of two deep-architecture functions for the generator (G) and the discriminator (D). The goal of the generator network is to simulate the probability distribution of the training data by mapping the random noise z to generate synthetic

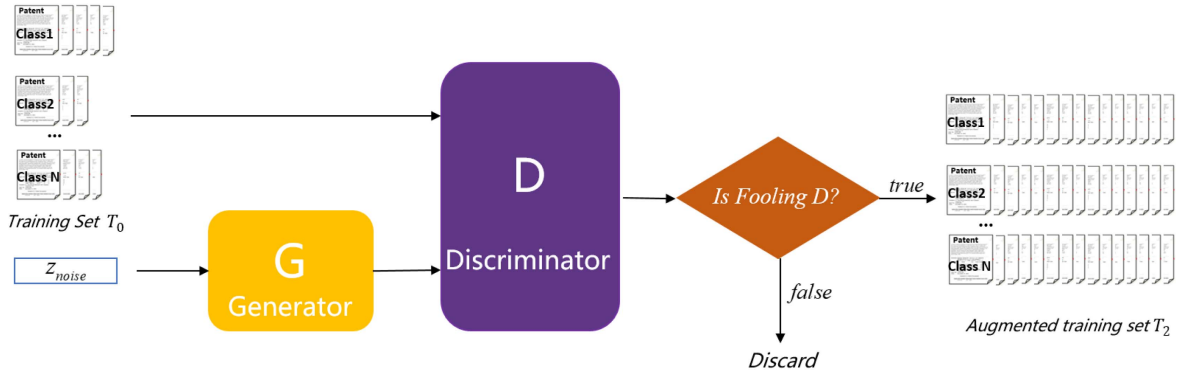

Fig. 2. Workflow of data augmentation using GAN.

data as close as possible to the real data. The discriminator network can learn the effective characteristics of the real data and can discriminate between the synthetic and real data. In other words, the generation process is a cat-and-mouse game, and when it achieves equilibrium, the synthetic and real data are indistinguishable.

Synthetic samples generated by the GAN has two steps, as shown in Fig. 2. First, the trained generator of the GAN is obtained when the loss functions of the generator and discriminator of GAN converge after being trained by training samples for each category in the training set T_0 several thousand times. Second, due to the adversarial idea in GAN (McDaniel *et al*., [43]), the generator attempts to generate synthetic samples that can fool the discriminator while the discriminator tries to distinguish between real and synthetic samples. This means that when synthetic samples are discriminated as real by the discriminator, synthetic samples are more akin to the distribution of the real training samples. The trained generator of the GAN is used to generate the original synthetic samples, and the discriminator is used to filter them. In the synthesized samples for each category created by the GAN generator, samples for each category that fool the discriminator are selected as the final synthesized samples.

The training process of GAN involves finding both the parameters of a discriminator that can maximize its classification accuracy and the parameters of a generator that can maximally confuse the discriminator. The training objective function of GAN is evaluated using a value function, as defined in (1), that depends on both D and G. These are updated, and the iteration stops when a Nash equilibrium is achieved. The input of the generator is white noise, and the output is the synthetic sample. The number of input units equals the dimension of the white noise, and the number of output units equals the number of patent features extracted by LDA. The number of neural network layers and the number of units per layer must be determined experimentally. The input of the discriminator is a real or synthetic sample, and the number of input units equals the number of patent features extracted by LDA. The output unit is one no-activation-function unit. The number of neural network layers and the number of units per layer also must be determined experimentally as follows:

$$\max_D \min_G V(D,G) = E_{P_{\text{data}}(s)}[\log D(s)] + E_{P_z(Z)}[\log(1 - D(G(z)))]. \quad (2)$$

In the above equation, s is a sample from the training set T_0; $D(s)$ is the probability that s comes from the training set T_0; z is white noise; $G(z)$ is the synthetic sample generated by the generator; and $D(G(z))$ is the probability that a synthetic sample is discriminated as real by the discriminator.

E. Mapping the Technological Landscape Based on DNN Classifiers

In this article, the DNN classifier, as a powerful supervised learning model, is used to classify patents into different value chain functional modules to better map the technological landscape of an emerging industry value chain. DNN classifiers based on deep learning are complex and have a large model capacity. After extensive training on large-scale labeled samples, they can exhibit superior performance vis-a-vis classical statistical supervised learning models because the multilayered neural network structure of DNN can learn the multilevel abstract features of samples [17].

The construction of the DNN classifier includes training and testing. First, the DNN classifier is trained by the augmented training set T_2 with large-scale and balanced synthetic samples generated by GAN. Then, the DNN classifier is tested by the original testing set T_1, which can effectively reflect the generalization performance of a classifier. The hyperparameters of the DNN classifier are similar to the discriminator of GAN. The input of the DNN classifier is synthetic samples in the augmented training set T_2, and the output is the corresponding value chain segments or functional modules in the value chain. The number of input units of the DNN classifier equals the number of selected patent features extracted by LDA. The number of output units equals the number of value chain segments or functional modules. The number of neural network layers and the number of units per layer must be determined experimentally.

When the trained DNN classifiers are obtained, it is necessary to classify all original patents that are not in the sample set C_0 into the corresponding value chain functional modules for mapping the technological landscape of the industry value chain. The patent feature vector of each original patent not in the sample set C_0 is input to the DNN classifier, which can directly and automatically classify the original patents into the corresponding

TABLE I
DIAGRAM OF THE CONFUSION MATRIX

		Predictive Class					
		C_1	C_2	•••	C_m	•••	C_L
Real Class	C_1	N_{11}	N_{12}	•••	N_{1m}	•••	N_{1L}
	C_2	N_{21}	N_{22}	•••	N_{2m}	•••	N_{2L}
	•••	•••	•••	•••	•••	•••	•••
	C_m	N_{m1}	N_{m2}	•••	N_{mm}	•••	N_{mL}
	•••	•••	•••	•••	•••	•••	•••
	C_L	N_{L1}	N_{L2}	•••	N_{Lm}	•••	N_{LL}

value chain functional module. As a result, we can obtain the distribution of all original patents in the structure of the industry value chain.

F. Evaluation of the Proposed Framework

The performance of the DNN classifiers directly reflects the quality of mapping the technological landscape of the industry value chain. We use three multiclassification metrics based on a confusion matrix: accuracy, F-measure, and G-mean, in order to test the performance of each DNN classifier. The accuracy is the proportion of predictions that are correct, the F-measure is the harmonic mean of precision and recall [44], and the G-mean is the geometric mean of recall [45]. The confusion matrix is given in Table I. The accuracy, F-measure, and G-mean are respectively defined by (3), (5). Accuracy is widely used as a basic indicator to evaluate the performance of a multiclassification classifier, but it cannot precisely describe the classifier performance on imbalanced data. Thus, the F-measure and G-mean are used to evaluate the performance of the multiclassification classifier on imbalanced datasets

$$\text{Accuracy} = \frac{\sum_{i=1}^{L} N_{ii}}{\sum_{i=1,j=1}^{L} N_{ij}} \tag{3}$$

$$\text{F} - \text{measure} = \frac{2}{L} g \frac{\sum_{i=1}^{L} R_i \sum_{i=1}^{L} P_i}{\sum_{i=1}^{L} R_i + \sum_{i=1}^{L} P_i} \tag{4}$$

$$\text{G} - \text{mean} = \left(\prod_{i=1}^{L} R_i \right)^{\frac{1}{L}}. \tag{5}$$

In (3)–(5), *L* is the number of classes, *Ri* and *Pi* are respectively the recall and precision of class *Li*, which are defined by (5) and (6). *Nii* and *Nij* are the number of class *Li* samples that are correctly predicted as class *Li* and incorrectly predicted as class *Lj*, respectively,

$$R_i = \frac{N_{ii}}{\sum_{j=1}^{L} N_{ij}} \tag{6}$$

$$P_i = \frac{N_{ii}}{\sum_{j=1}^{L} N_{ji}}. \tag{7}$$

To evaluate the effectiveness of the integrated GAN-DNN framework to map the technological landscape of the industry value chain with small-scale and imbalanced training samples, the SVM, as a statistical supervised learning classifier, is employed for comparison experiments. SVM has an excellent generalization performance and has been widely used in text classification (Kreuchauff, 2018), image classification, and character recognition [46].

IV. EMPIRICAL ANALYSIS AND RESULTS

This article was conducted in the 3-D printing industry to map the technological landscape of the industry value chain by describing and analyzing the patent distribution of each segment in the 3-D printing value chain. 3-D printing is an emerging industry that was first commercialized in the late 1980s. It has a wide spectrum of applications in various industries, including aerospace, industrial machinery, motor vehicles, architectural designs, national defense, medicine, consumer products, and academic research.

The patent data were obtained from the DI index patent database. As the world's most comprehensive collection of global patents and patent documents, it is well-structured, thus enabling researchers to conduct extensive real-time analyses. The global 3-D printing patent data from 1990 and 2017 were retrieved and downloaded from the DI patent database on January 17, 2018.[1] A total of 26 080 patent documents containing title, abstract, and assignee information were retrieved from the DI patent database and used as the original patent set C_0, and the year distribution of the collected patents is shown in Fig. 3.

A. Results of Dataset Construction

The VSM of patent set C_0 was constructed based on the topics' probability distribution of patent documents calculated by LDA. The LDA model contains three hyperparameters, *K*, α, and β. *K* is the number of topics [47]. The relationship between the number of topics and the perplexity of LDA is shown in Fig. 4. When the number of topics is less than 200,

[1]Patent search query for the 3-D printing industry: ALLD=(3*D ADJ printing OR three-dimensional ADJ printing OR 3-D ADJ printing OR material ADJ increase ADJ manufact* OR additive* ADJ manufact* OR rapid* ADJ prototyp* OR rapid ADJ manufact* OR rapid* ADJ prototyp* ADJ manufact* OR Layered ADJ Manufact* ADJ Technology OR Solid ADJ free-form ADJ Fabrication OR Stereo ADJ Lithography ADJ Apparatus OR Laminated ADJ Object ADJ Manufact* OR Selective ADJ Laser ADJ Sinter* OR Fused ADJ Deposition ADJ Model* OR Laser ADJ Engineered ADJ Net ADJ Shap* OR Patternless ADJ Casting ADJ Manufact* OR Direct ADJ Metal ADJ Laser-Sinter* OR Direct ADJ Laser ADJ Fabrication OR direct ADJ metal ADJ deposition OR Laser ADJ clad* ADJ forming ADJ technology OR Electron ADJ Beam ADJ Selective ADJ Melt* OR Digital ADJ bricklay* OR 3D ADJ mosaic OR ballistic ADJ particle ADJ manufact*) AND (PRYS>=(1900) AND PRYS<=(2017)).

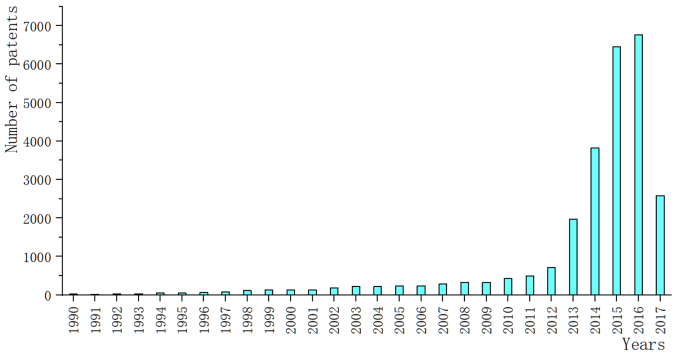

Fig. 3. Year distribution of patents of the patents.

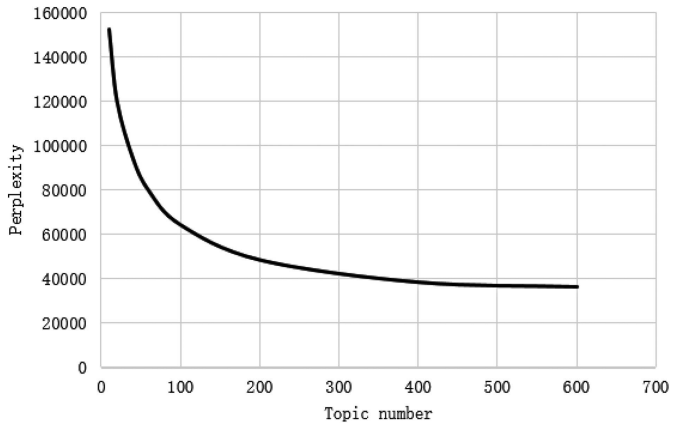

Fig. 4. Perplexity varies with topic number in the LDA model.

the perplexity decreases with increasing numbers of topics; and when the number of topics is larger than 200, the perplexity gradually becomes smooth so that over-fitting occurs. Thus 200 ($K = 200$) was selected as the number of topics of LDA to perform the feature extraction of patent set C_0 in this article. The α adjusts the number of topics contained in a document. The higher the α value, it means that a document should contain more topics. Consistent with the existing literature, we set $\alpha = 50 / K = 0.25$. The β adjusts the ambiguity of each topic. The higher the β value, the higher the ambiguity of each topic. Conversely, the lower the β value, the more specific the topic. Consistent with the existing literature, we set $\beta = 0.01$. [37]. The number of iterations was set to 10 000 to train LDA.

Experts labeled the samples based on their domain knowledge of the 3-D printing industry value chain and the annual report, *Wohlers Report 2017*. Released by the 3-D printing-analysis firm Wohlers Associates Inc., this has been the undisputed industry-leading report on the subject for more than two decades. Based on the domain knowledge of experts and *Wohlers Report 2017*, the 3-D printing value chain has four major segments: material, design, equipment manufacturing, and application and services, which can be subdivided into 11 functional modules, as shown in Fig. 5. The manufacturing process in this value chain is as follows:

1) Select the printing materials needed from composite, polymer, metal, and ceramic materials, as needed.

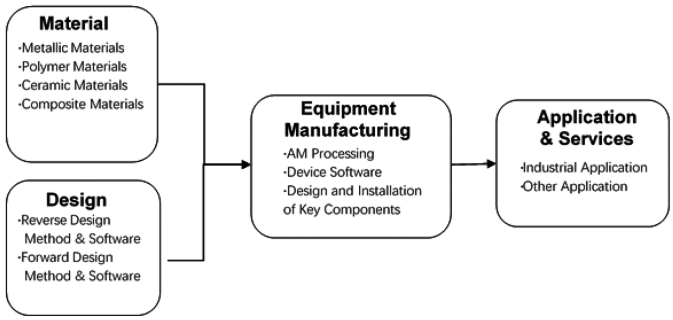

Fig. 5. Industry value chain structure of 3-D printing.

TABLE II
CLASS DISTRIBUTION OF SAMPLE SET

Class Label	Number of Samples	Class Label	Number of Samples
Material		**Equipment Manufacturing**	
Ceramic Materials	34	Design and Installation of Key Components	34
Composite Materials	149	Device Software	57
Polymer Materials	100	AM Processing	129
Metallic Materials	46		
Design		**Application and Services**	
Reverse Design Method & Software	122	Industrial Application	33
Forward Design Method & Software	85	Other Application	57

2) Select the corresponding design methods and software between the reverse and forward to establish the digital model to be printed.
3) Input the selected materials and digital models to the appropriate process and equipment, i.e., design and installation of key components to output the finished product, select the device software and additive manufacturing (AM) process.
4) Use the product in industrial or other services.

According to the definitions of the structure of the 3-D printing value chain, the sample set $\mathbf{C}_1$ was manually labeled through the patent content. Then, to precisely evaluate the performance of mapping the technological landscape of the 3-D printing value chain, the sample set $\mathbf{C}_1$ was randomly divided into a training set T_0 and a testing set T_1 at a ratio of 7:3. From the original patent set C_0, 846 patent documents were randomly selected as sample set C_1 and the category of each patent document in the sample set C_1 was manually identified based on the structure of the 3-D printing value chain, as given in Table II. The functional modules in the structure of the 3-D printing value chain were selected as the criteria for labeling patent documents, and the sample set C_1 was labeled according to the following 11 functional modules: *ceramic material, composite material, polymer material, metallic material; reverse design methods and software, forward design methods and software; design and installation of key components, device software, AM processing; industrial applications,* and *other applications*. For subsequent classification tasks, the sample set C_1 was partitioned into a training set T_0, accounting for 70% of C_1, with the remaining 30% constituting the testing set T_1.

TABLE III
NUMBER OF HIDDEN LAYERS AND HIDDEN UNITS OF EACH GAN

Dataset		Generator			Discriminator		
		First Hidden Layer	Second Hidden Layer	Output Layer	First Hidden Layer	Second Hidden Layer	Output Layer
Material	Ceramic Materials	4	N/A	20	4	N/A	1
	Composite Materials	4	4	20	4	4	1
	Polymer Materials	8	8	20	8	8	1
	Metallic Materials	8	8	20	8	8	1
Design	Reverse Design Method & Software	4	4	20	4	4	1
	Forward Design Method & Software	4	4	20	4	4	1
Equipment Manufacturing	Design and Installation of Key Components	4	4	20	4	4	1
	Device Software	4	4	20	4	4	1
	AM Processing	4	4	20	4	4	1
Application and Services	Industrial Application	4	4	20	4	4	1
	Other Application	4	4	20	4	4	1

TABLE IV
NUMBER OF HIDDEN LAYERS AND HIDDEN UNITS OF EACH DNN

Classification process	Hidden units of each DNN
First Layer	[20, 64, 20]
Second Layer	
Materials	[20,32,32,20]
Design	[5,5,5]
Equipment Manufacturing	[5,5,5]
Services	[16,16,16]

B. Hyperparameters for GAN

Hyperparameters for GANs are empirically determined. The generator and discriminator of GAN are both DNNs, and due to their complex structure, many hyperparameters can influence the performance of GAN. The fundamental hyperparameters of the generator and discriminator are the number of layers and the units of each hidden layer. Too few layers will hinder the ability of a network to build a representation at a level of abstraction appropriate to adequately capture data complexity, and too many layers will complicate training and likely cause overfitting [23]. Since training and tuning a GAN is expensive, we conducted a limited number of experiments to determine the optimal hyperparameters. Thus, 1 to 5 hidden layers were tested in the generator and discriminator. For the number of units per hidden layer, 2, 4, 8, 12, 16, and 32 nodes were tested.

Through a series of experiments, the optimal hyperparameters were determined for all 11 GANs by the loss function's convergence value of the generator and discriminator. The number of hidden layers and hidden units for all generators and discriminators are given in Table III, and all hidden layer nodes of generators and discriminators are rectified linear units (ReLUs). The dimension of the noise vector z for all generators was set to 5; 20 sigmoid units and one no-activation-function unit were used as the output layer of the generator and discriminator, respectively. In each iteration of GAN training, the discriminator first iterated 100 times, whereas the generator iterated once. The development environment of a GAN was TensorFlow 1.1 with Python 3.5.2, and each GAN was trained through a GPU. When all the trained GAN are obtained, the trained GAN are employed to generate 1000 corresponding synthetic samples for each functional module of the 3-D printing value chain.

C. Classification Results of DNN Classifiers

According to the structure of the 3-D printing value chain presented in Fig. 2, the functional modules were selected as the criteria of patent classification, and the sample set was then labeled into 11 functional modules. As these 11 functional modules corresponded to 4 value segments, we performed the classification task on two layers with five times classification to map the technological landscape of the 3-D printing value chain. The first layer was to classify the sample set based on value chain segments, and the second layer was to classify the sample set based on functional modules of each value chain segments.

Hyperparameters for the DNN classifier were also empirically determined. The structure of DNN classifier was a multilayer neural network, and was similar to the discriminator of the GAN [20]. We also conducted a limited number of experiments, in which 2 to 32 nodes with 1 to 5 hidden layers were tested, and the optimal parameter was determined by the accuracy, F1, and G-mean. Through a series of experiments, the best-performing hyperparameters for all five DNN classifiers were determined. The number of hidden layers and hidden units for all DNNs are given in Table IV. All hidden layer nodes of DNN were ReLU, the dimension of input units was 20, two softmax units were used as the output layer for each DNN, and cross-entropy was used as the loss function. The number of iterations was set to 20 000 when classifying the first layer and application and services of the second layer, 6000 when classifying the design, 3000 when classify equipment manufacturing, and 12 000 when classifying materials. The development environment of each DNN was TensorFlow 1.1 with Python 3.5.2, and each DNN was trained through a GPU.

The classification results of the first layer with four segments on the testing set data can be given in Table VI of appendix. A total of 163 patents were correctly classified into corresponding value chain segments with an accuracy of 64%. Specifically, 21 of 27 services patents, 38 of 66 equipment manufacturing patents, 37 of 61 design patents, and 67 of 99 materials parents were identified correctly. In the second layer, the functional module classification resulted in four value chain segments, given in Tables VII–X of appendix. In the services segment, 6 of 10 industrial applications patents and 9 of 17 other applications patents were classified correctly. In the equipment manufacturing segment, 13 of 17 device software patents, 32 of 39 AM processing patents, and 3 of 10 design and installation of key components patents were classified correctly, achieving 72% overall accuracy. In the design segment, 31 of 36 reverse design methods and software patents and 20 of 25 forward design methods and software patents were classified correctly, achieving 83% accuracy. In the materials segment, 3 of 10 ceramic materials patents, 32 of 45 composite materials patents, 17 of 30 polymer materials patents, and 8 of 14 metallic materials patents were classified correctly.

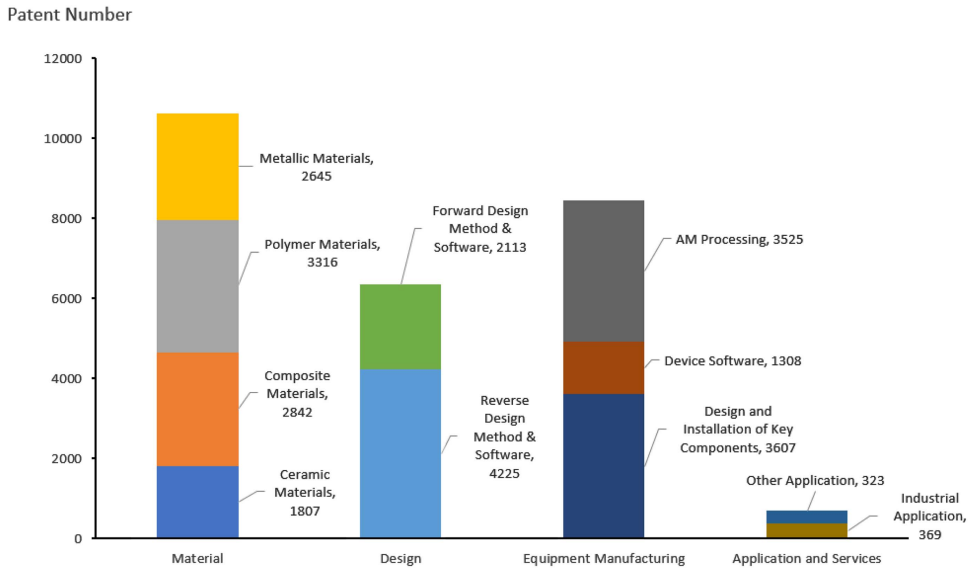

Fig. 6. Patent distribution of the 3-D printing industry.

D. Mapping the 3-D Printing Industry Value Chain

We used the framework of GAN-DNN, as trained and tested above, to classify the original patent set into 11 functional modules based on the 3-D printing industry value chain, and visualized the result, which mapped the current technological landscape of the 3-D printing industry value chain by patent distribution, as shown in Fig. 6. The 10 610 patents in the materials section are more than those in the design, equipment manufacturing, and services sections, respectively. The process and technology innovation of materials has a vital effect on the entire 3-D printing industrial value chain. Moreover, based on the patent distribution, in the materials section, polymer materials, metal materials, and composite materials are the three main types of 3-D printing materials. In the services section, technology development is mainly focused on industrial applications, indicating that general application scenarios in the 3-D printing industry are still on the industrial level.

Fig. 7 shows the technological landscape of the 3-D printing industry value chain of the world, China, the United States, and Germany through a patent lens. The technological focus and trajectory selection are different in these countries. As to the material segment, China has advantages in metallic materials; U.S. concentrates on composite materials; Germany focuses on polymer materials. As to the design segment, China primarily develops reverse design method and software, while the U.S. and Germany select the forward design method and software as their development focus. As to the equipment manufacturing segment, the advantages of China are the design and installation of key components, while the superiorities of the U.S. and Germany are AM processing. As to the application and service segment, it is still weak in all these countries. From the technological landscape mapping, we can identify the strengths, weaknesses, and competitive advantages along the industry value chain.

E. Evaluation of the Proposed Framework

To demonstrate the effectiveness of our process, data augmented based on GAN and classification based on DNN classifiers, comparisons with SVM, and DNN were also conducted. We used the same original training set to train the SVM and DNN classifier. The SVM parameters were applied to the default value: C = 1, gamma = "auto," kernel = "rbf," but in order to avoid the situation of classification result as extremely skewed, the SVM parameters will be moderately adjusted based on the default value. The DNN classifier for comparison used the same parameters as we did. The parameters of each classification process are given in Tables XI and Table XII of appendix. In Table V, we compare our proposed framework with alternative methods in terms of accuracy, F-measure, and G-mean.

As given in Table XI, the accuracy, F-measure, and G-mean of SVM in five classification processes are generally smaller than those of DNN, which shows that DNN can easily distinguish multiple categories in the limited training set, while SVM based on statistics algorithm performs poorly. Furthermore, all the reported metrics improve appreciably when an augmented training set is used by comparing DNN and GAN-DNN, as the GAN algorithm offers massive training data to help DNN learn more useful features that can significantly improve classification performance. The comparison result is consistent with the existing literature, after large-scale data training and sufficient calculation, GAN and DNN will have a strong ability to learning input features, and thus have better performance than the SVM model based classic statistical method (Liu *et al.*, 2020 and [20]). This indicates that our proposed framework, data augmented based on GAN and classification based on DNN, has overall better performance and can effectively determine the classification of small-scale and imbalanced sample data in an emerging industry value chain.

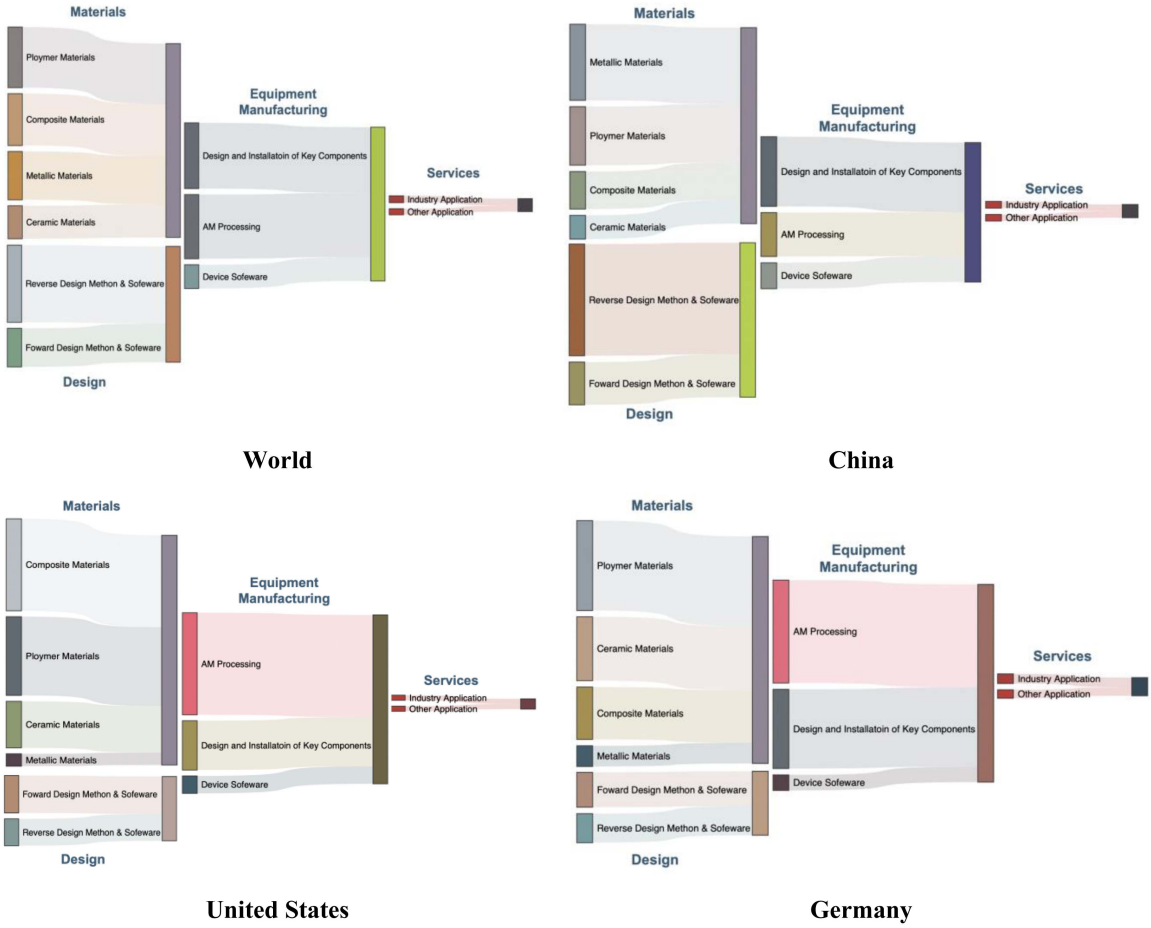

Fig. 7. Technological landscape of 3-D printing industry value chain.

TABLE V
CLASSIFICATION METRICS

Classification Metrics			
Model	**Accuracy**	**F-measure**	**G-mean**
First Layer-Value Chain Segment			
SVM	0.3992094862	0.3883465693	0.3110978077
DNN	0.581027668	0.5787502432	0.5685602719
GAN-DNN	0.6442687747	0.6602764888	0.6547894482
Second Layer-Application and Services			
SVM	0.4444444444	0.437416876	0.433860916
DNN	0.518519	0.494217894	0.48507125
GAN-DNN	0.555556	0.562564633	0.563601862
Second Layer-Equipment Manufacturing			
SVM	0.621212121	0.534969186	0.180516551
DNN	0.681818	0.548814696	0.386906387
GAN-DNN	0.727273	0.637307543	0.573104325
Second Layer-Design			
SVM	0.590163934	0.564076387	0.53748385
DNN	0.655738	0.696719631	0.663324958
GAN-DNN	0.836066	0.830555556	0.829993307
Second Layer-Material			
SVM	0.515151515	0.473501054	0.283216103
DNN	0.535354	0.493265381	0.476506542
GAN-DNN	0.606061	0.537019505	0.512669032

V. CONCLUSION

We propose an integrated framework, i.e., data augmentation based on GAN and classification based on DNN, which effectively classify the patents with small-scale and imbalanced sample data to map the emerging industry value chain. Compared to applying machine learning classifiers alone, the integrated framework increases the accuracy of patent classifications and improves the quality of the mapping. We use the 3-D printing as a case of the emerging industry, and apply the proposed framework to identify the strengths, weaknesses, and competitive advantages along the 3-D printing industry value chain of China, the United States, and Germany. The proposed integrated framework was demonstrated to be an effective tool to map the technological landscape of an emerging industry value chain.

The contributions of this article were twofold.

First, from an academic perspective, this article contributes to research on mapping the emerging industry value chain through a patent lens. By using key technologies in the AI domain, such as GAN and DNN, it effectively reduces the dependence on large-scale and balanced patent sample data in existing research on mapping emerging industry value chain, and promotes the application of AI technology in research related to the emerging industry value chain. The proposed framework can overcome the

problem of weak sample data, and was suitable for emerging industries with small-scale patents and imbalanced patent sample data in different segments along the industry value chain.

Second, from a practical perspective, this article helps policymakers and enterprises assess industrial competitiveness along a value chain and find out where the bottlenecks were and which parts speed up progress. It depict the technological landscape of an emerging industry value chain, provide an overview of the technology layout, and evaluate the strengths and weakness along the industry value chain to identify strategic opportunities and take action.

Nevertheless, further study was required. The semantic patent features extracted by LDA were used to verify the effectiveness of the proposed framework in this article. Besides LDA-based semantic patent features, several other patent feature extraction methods was used for patent autoclassification, such as named entity recognition method and word vectors method. Subsequent studies use the proposed framework to explore more effective patent feature extraction methods.

APPENDIX

TABLE VI
CONFUSION MATRIX OF VALUE CHAIN SEGMENT IN FIRST LAYER

	Classification result				Total
	Material	Design	Equipment Manufacturing	Application and Services	
Material	3	3	0	21	27
Design	18	9	38	1	66
Equipment Manufacturing	8	37	16	0	61
Application and Services	67	8	19	5	99
Total	96	57	73	27	253

TABLE VII
CONFUSION MATRIX OF MATERIALS IN SECOND LAYER

	Classification result				Total
	Ceramic Materials	Composite Materials	Polymer Materials	Metallic Materials	
Ceramic Materials	3	3	2	2	10
Composite Materials	4	32	7	2	45
Polymer Materials	2	9	17	2	30
Metallic Materials	2	2	2	8	14
Total	11	46	28	14	99

TABLE VIII
CONFUSION MATRIX OF DESIGN IN SECOND LAYER

	Classification result		Total
	Reverse design method & software	Forward design method & software	
Reverse design method & software	31	5	36
Forward design method & software	5	20	25
Total	36	25	61

TABLE IX
CONFUSION MATRIX OF EQUIPMENT MANUFACTURING IN SECOND LAYER

	Classification result			Total
	Design and Installation of Key Components	Device Software	AM Processing	
Design and Installation of Key Components	2	13	2	17
Device Software	6	1	32	39
AM Processing	3	1	6	10
Total	11	15	40	66

TABLE X
CONFUSION MATRIX OF APPLICATION AND SERVICES IN SECOND LAYER

	Classification result		Total
	Industrial applications	Other applications	
Industrial applications	6	4	10
Other applications	8	9	17
Total	14	13	27

TABLE XI
CLASSIFICATION PARAMETERS IN FIRST LAYER

Classification process	Classification parameters			
	SVM	DNN		
		Number of hidden layers	Units of each layer	Iterations
First Layer	default value: C = 1, gamma = 'auto', kernel = 'rbf'	3	[20, 64, 20]	20,000

TABLE XII
CLASSIFICATION PARAMETERS IN SECOND LAYER

Classification process	Classification parameters			
	SVM	DNN		
		Number of hidden layers	Units of each layer	Iterations
Materials	default value: C = 1, gamma = 'auto', kernel = 'rbf'	4	[20,32,32,20]	20000
Design	adjusted value: C = 30, gamma = 15, kernel = 'rbf'	3	[5,5,5]	3,000
Equipment Manufacturing	adjusted value: C = 7, gamma = 'auto', kernel = 'rbf'	3	[5,5,5]	6,000
Application and Services	adjusted value: C = 30, gamma = 15, kernel = 'rbf'	3	[16,16,16]	12,000

REFERENCE

[1] T. A. Chiang and A. J. Trappey, "Development of value chain collaborative model for product lifecycle management and its LCD industry adoption," *Int. J. Prod. Econ.*, vol. 109, no. 1/2, pp. 90–104, 2007.

[2] J. Peppard, and A. Rylander, "From value chain to value network: Insights for mobile operators," *Eur. Manage. J.*, vol. 24, no. 2/3, pp. 128–141, 2006.

[3] Y. T. Li, M. H. Huang, and D. Z. Chen, "Semiconductor industry value chain: Characters," *Technol. Evol. Ind. Manage. Data Syst.*, vol. 111, no. 3, pp. 370–390, 2011.

[4] H. Schmitz, *Value Chain Analysis for Policy-Makers and Practitioners*. Geneva, Switzerland: Int. Labour Org., 2005

[5] D. M. Boehe, and L. B. Cruz, "CSR in the global marketplace: towards sustainable global value chains," *Manage. Decis.*, vol. 46. no, 8, pp. 1187–1209, 2008.

[6] A. Noruzi and M. Abdekhoda "Mapping Iranian patents based on international patent classification (IPC), from 1976 to 2011," *Scientometrics*, vol. 93, no. 3, pp. 847–856, 2012.

[7] I. Sakata, H. Sasaki, M. Akiyama, Y. Sawatani, N. Shibata, and Y. Kajikawa, "Bibliometric analysis of service innovation research: identifying knowledge domain and global network of knowledge," *Technolog. Forecasting Social Change*, vol. 80, no. 6, pp. 1085–1093, 2013

[8] D. Kong, Y. Zhou, Y. Liu, and L. Xue, "Using the data mining method to assess the innovation gap: A case of industrial robotics in a catching-up country," *Technol. Forecast. Soc. Change*, vol. 119, pp. 80–97, 2017.

[9] M. Brink, "Development of a method to forecast future systems in the forest engineering value chain," Ph.D. dissertation, *Stellenbosch Univ.*, 2001.

[10] J. Zhang, Z. B. Liu and J. H. Zheng, "Industrial chain positioning, divide, agglomeration and innovation: an empirical study based on questionnaire of manufacturing firms in Jiangsu Province," *China Ind. Economy*, vol. 7, pp. 47–55, 2007.

[11] Y. T. Li, M. H. Huang, and D. Z. Chen, "Is Foundry only a capacity provider still?: Relations of role playing for semiconductor industry value chain by patent analysis," in *Proc. IEEE Int. Conf. Manage. Innov. Technol.*, 2010, pp. 220–225.

[12] N. Islam "Crossing the valley of death—An integrated framework and a value chain for emerging technologies," *IEEE Trans. Eng. Manage.*, vol. 64, no. 3, pp. 389–399, Aug. 2017.

[13] C. M. Bishop, *Pattern Recognition and Machine Learning*. Berlin, Germany: Springer, 2006.

[14] Y. Zhou, F. Dong, D. Kong, and Y. Liu, "Unfolding the convergence process of scientific knowledge for the early identification of emerging technologies," *Technol. Forecast. Soc. Change*, vol. 144, pp. 205–220, 2019.

[15] A. Rodriguez, A. Tosyali, B. Kim, J. Choi, J. M. Lee and B. Y. Coh, "Patent clustering and outlier ranking methodologies for attributed patent citation networks for technology opportunity discovery," *IEEE Trans. Eng. Manage.*, vol. 63, no. 4, pp. 426–437, Nov. 2016

[16] K. Wang, F. Y. Wang, X. Li, and L. Yan, "Generative adversarial networks: Introduction and outlook," *IEEE/CAA J. Autom. Sinica.*, vol. 4, no. 4, pp. 588–598, 2017.

[17] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, and S. Ozair "Generative adversarial nets," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2014, vol. 3, pp. 2672–2680.

[18] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv: 1511.06434*.

[19] Y. Liu, Y. Zhou, X. Liu, F. Dong, C. Wang, and Z. Wang, "Wasserstein GAN-based small-sample augmentation for new-generation artificial intelligence: A case study of cancer-staging data in biology," *Engineering*, vol. 5, no. 1, pp. 156–163, 2019.

[20] Y. Zhou, F. Dong, Y. Liu, Z. Li, J. Du, and L. Zhang, "Forecasting emerging technologies using data augmentation and deep learning," *Scientometrics*, vol. 123, no. 1, pp. 1–29, 2020.

[21] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. 1, pp. 321–357, 2002.

[22] E. Derouin and J. Brown, "Neural network training on unequally represented classes," *Intell. Eng. Syst. Through Artif. Neural Netw.*, pp. 135–145, 2012.

[23] U. Fiore, A. De Santis, F. Perla, P. Zanetti, and F. Palmieri, "Using generative adversarial networks for improving classification effectiveness in credit card fraud detection," *Inf. Sci.*, vol. 479, pp. 448–455, 2017.

[24] U. Hwang, S. Choi, and S. Yoon, "Disease prediction from electronic health records using generative adversarial networks," 2018, *arXiv:1711. 04126*.

[25] J. Li, W. Monroe, T. Shi, S. Jean, A. Ritter, and D. Jurafsky "Adversarial learning for neural dialogue generation," 2017, *arXiv:1701.06547*.

[26] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," 2017, *arXiv:1703.09452*.

[27] L S. Larkey, "A patent search and classification system," in *Proc. 4th ACM Conf. Digit. Libraries*, 1999, 179–187.

[28] C. H. Wu, Y. Ken, and T. Huang, "Patent classification system using a new hybrid genetic algorithm support vector machine," *Appl. Soft Comput.*, vol. 10, no. 4, pp. 1164–1177, 2010.

[29] W. Q. Guo, J. Wen, and G. H. Wen, "Patent categorization based on Bayes model," *Comput. Eng. Des.*, vol. 26, no. 8, pp. 1986–1996, 2005.

[30] S. Li, J. Hu, Y. Cui, and J. Hu, "DeepPatent: Patent classification with convolutional neural networks and word embedding," *Scientometrics*, vol. 117, no. 2, pp. 721–744, 2018.

[31] S. U. Hassan, M. Imran, S. Iqbal, N. R. Aljohani, and R. Nawaz, "Deep context of citations using machine-learning models in scholarly full-text articles," *Scientometrics*, vol. 117, no. 3, pp. 1645–1662, 2018.

[32] Y. Zhang *et al.*, "Does deep learning help topic extraction? A kernel k-means clustering method with word embedding," *J. Informetrics*, vol. 12, no. 4, pp. 1099–1117, 2018.

[33] Y. Qi, N. Zhu, Y. Zhai, and Y. Ding "The mutually beneficial relationship of patents and scientific literature: Topic evolution in nanoscience," *Scientometrics*, vol. 115, no. 2, pp. 893–911, 2018.

[34] W. Q. Li, Y. Li, J. Chen, and C. Y. Hou "Product functional information based automatic patent classification: method and experimental studies," *Inf. Syst.*, vol. 67, pp. 71–82, 2017.

[35] L. Liefa and Z. Y. Le Fugang, "The application of LDA model in patent text classification," *J. Modern Inf.*, vol. 37, no. 3, pp. 35–39, 2017.

[36] S. Venugopalan and V. Rai, "Topic based classification and pattern identification in patents," *Technol. Forecast. Soc. Change*, vol. 94, pp. 236–250, 2015.

[37] X. Wei and W. B. Croft, "LDA-based document models for Ad-hoc retrieval," in *Proc. 29th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2006, pp. 178–185.

[38] S. Zhou, K. Li, and Y. Liu, "Text categorization based on topic model," *Int. J. Comput. Intell. Syst.*, vol. 2, no. 4, pp. 398–409, 2009.

[39] D. M. Blei, A. Y. Ng, M. I. Jordan, and J. Lafferty, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[40] W. Cynthia, "Analyzing interview data: the development and evolution of a coding system," *Qualitative Sociol.*, vol. 24, no. 3, pp. 381–400, 2001.

[41] K. M. Macqueen, E. Mclellan, K. Kay, B. Milstein, and A. Cdc, "Codebook development for team-based qualitative analysis," *Field Methods*, vol. 10, no. 2, pp. 31–36, 1998.

[42] J. C. Chang, S. Amershi, and E. Kamar, "Revolt: Collaborative crowdsourcing for labeling machine learning datasets," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2017, pp. 2334–2346.

[43] P. Mcdaniel, N. Papernot, and Z. B. Celik, "Machine learning in adversarial settings," *IEEE Secur. Privacy*, vol. 14, no. 3, pp. 68–72, May/Jun. 2016.

[44] L. R. Sousa, T. Miranda, R. L. Sousa, and J. Tinoco, "The use of data mining techniques in rockburst risk assessment," *Engineering*, vol. 3, no. 4, pp. 552–558, 2017.

[45] Y. Sun, M. S. Kamel, A. K. Wong, and Y. Wang, "Cost-sensitive boosting for classification of imbalanced data," *Pattern Recognit.*, vol. 40, no. 12, pp. 3358–3378, 2007.

[46] P. Kenekayoro, "Identifying named entities in academic biographies with supervised learning," *Scientometrics*, vol. 116, no. 2, pp. 751–765, 2018.

[47] D. M. Blei, A. Y. Ng, M. I. Jordan, and J. Lafferty, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2012.

[48] F. Kreuchauff and V. Korzinov, "A patent search strategy based on machine learning for the emerging field of service robotics," *Scientometrics*, vol. 111, no. 2, pp. 743–772, 2017.

[49] B. P. Abraham, and S. D. Moitra, "Innovation assessment through patent analysis," *Technovation*, vol. 21, no. 4, pp. 245–252, 2001.

[50] C. J. Liang and M. L. Yao, "The value-relevance of financial and nonfinancial information—evidence from Taiwan's information electronics industry," *Rev. Quant. Finance Account.*, vol. 24, no. 2, pp. 135–157, 2005.

[51] M. Porter, *Competitive Advantage: Creating and Sustaining Superior Performance*, New York, NY, USA: Free Press, 1985.

[52] J. C. Short, Jr., D. J. Ketchen., T. B. Palmer, and G. T. M. Hult4, "Firm, strategic group, and industry influences on performance," *Strategic Manage. J.*, vol. 28, no. 2, 147–167, 2007.

[53] M. Trajtenberg, "Innovation in Israel 1968–1997: A comparative analysis using patent data," *Res. Policy*, vol. 30, no. 3, pp. 363–389, 2001.

[54] A. J. Trappey, C. V. Trappey, U. H. Govindarajan, and J. J. Sun, "Patent value analysis using deep learning models - The case of IoT technology mining for manufacturing industries," *IEEE Trans. Eng. Manage.*, to be published, doi: 10.1109/TEM.2019.2957842.

[55] H. W. Wang and M. C. Wu, "Business type, industry value chain, and r&d performance: Evidence from high-tech firms in an emerging market," *Technol. Forecast. Soc. Change*, vol. 79, no. 2, pp. 326–340, 2012.

[56] C. K. Yau, A. Porter, N. Newman, and A. Suominen, "Clustering scientific documents with topic modeling," *Scientometrics*, vol. 100, no. 3, pp. 767–786, 2014.