Scientometrics (2020) 123:1–29
https://doi.org/10.1007/s11192-020-03351-6

Forecasting emerging technologies using data augmentation and deep learning

Yuan Zhou[1] · Fang Dong[1] · Yufei Liu[1,3] · Zhaofu Li[2] · JunFei Du[2] · Li Zhang[2]

Received: 28 July 2018 / Published online: 25 January 2020
© The Author(s) 2020

Abstract
Deep learning can be used to forecast emerging technologies based on patent data. However, it requires a large amount of labeled patent data as a training set, which is difficult to obtain due to various constraints. This study proposes a novel approach that integrates data augmentation and deep learning methods, which overcome the problem of lacking training samples when applying deep learning to forecast emerging technologies. First, a sample data set was constructed using Gartner's hype cycle and multiple patent features. Second, a generative adversarial network was used to generate many synthetic samples (data augmentation) to expand the scale of the sample data set. Finally, a deep neural network classifier was trained with the augmented data set to forecast emerging technologies, and it could predict up to 77% of the emerging technologies in a given year with high precision. This approach was used to forecast emerging technologies in Gartner's hype cycles for 2017 based on patent data from 2000 to 2016. Four out of six of the emerging technologies were forecasted correctly, showing the accuracy and precision of the proposed approach. This approach enables deep learning to forecast emerging technologies with limited training samples.

Keywords Emerging technologies forecasting · Data augmentation · Deep learning · Supervised learning

Introduction

Forecasting emerging technologies is important for governments and enterprises to identify strategic opportunities in the face of technological change. Existing forecasting studies use either normative or extrapolative methods, while the latter mainly involves the analyses of bibliometrics and patents. For example, Daim et al. (2006) used patents and curve fitting techniques to forecast emerging technologies; Chang et al. (2010) proposed an approach

✉ Yufei Liu
liuyufei0418@qq.com

[1] School of Public Policy and Management, Tsinghua University, Beijing 100084, China

[2] School of Mechanical Science and Engineering, Huazhong University of Science and Technology, Wuhan 430074, China

[3] Center for Strategic Studies, Chinese Academy of Engineering, Beijing 100088, China

employing bibliometrics and patent network analysis to forecast emerging technologies; and Breitzman and Thomas (2015a) assessed the value of patents by considering inventors' team-size, in order to identify high-value emerging technologies. However, these approaches were mostly concerned with bibliometric/patent indicators and ignored rich text information contained by these patents. In addition to these studies, other recent studies attempted to explore text mining and cluster analysis. For instance, Chang and Breitzman (2009) used the clustering of patents to identify emerging and high-impact technology clusters and trends; Chiavetta and Porter (2013) proposed the basic idea of tech mining to forecast emerging technologies using text and data mining, Choi and Jun (2014) developed a Bayesian model for patent clustering to forecast emerging technologies; Breitzman and Thomas (2015b) proposed the "Emerging Clusters Model" based on patent citations to identify emerging technologies across multiple patent systems; and Zhou et al. (2019a, b) developed a framework through citation network and topology clustering to reveal the convergence process of scientific knowledge to forecast emerging technologies. All these approaches utilized unsupervised learning to probe valuable text-based data. However, unsupervised learning methods cannot incorporate external domain knowledge during the machine learning process, and the results need to be professionally interpreted by domain experts that are usually rare, costly, and sometimes biased.

As a remedy, supervised learning methods can generate forecasting results by embedding external knowledge into the model using labeled samples. Some recent studies have explored the use of supervised learning. For example, Kreuchauff and Korzinov (2015) developed a support vector machine model based on robotics patents to detect the early development of an emerging technology in patent data. Kyebambe et al. (2017) used labeled data based on new classes established in the United States Patent Classification (USPC) system to train supervised learners to forecast emerging technologies. Lee et al. (2018) employed a feed-forward multilayer neural network to capture the complex nonlinear relationships between the input and output indicators to identify emerging technologies in early stages. Zhou et al. (2019a, b) developed a semi-supervised topic clustering model and generated a sentence-level semantic technological topic description to identify emerging technologies. Supervised learning requires high-quality labeled samples to prevent overfitting and ensure the accuracy of the forecasting model. However, large labeled samples are difficult to obtain.

As an advanced supervised learning process, deep learning has a relatively complex model structure and exhibits better performance (Liu et al. 2019). Some recent studies have explored the applications of deep learning in bibliometrics and patent analysis, such as patent classification (Li et al. 2018), citation classification (Hassan et al. 2018), and natural language processing (Zhang et al. 2018). These studies showed that deep learning exhibits superior performances and great potential for forecasting emerging technologies compared to the traditional supervised learning methods. However, these studies used large-scale labeled sample-data to fully optimize the model parameters and lead to superior performance. Existing studies suggests that the sample-size of a dataset would significantly affect the deep learning performance (Goodfellow et al. 2016). However, large training samples (e.g., emerging technologies in history) are difficult to obtain due to data/resource constraints. To cope with this issue, we utilize a generative adversarial network (GAN) method, which recently emerged in computer science, as a data augmentation method to enlarge the data scale for emerging technologies samples.

Superior to basic deep learning, an integrated GAN-based deep learning can help to develop new approaches to address the problem of lacking emerging technology samples. A GAN consists of two deep-architecture functions for the generator and the discriminator, which can learn simultaneously from the trained data in an adversarial fashion (Radford et al. 2015). Most recent studies have shown that a GAN can effectively augment training samples. For example, Fiore et al. (2019) used a GAN to generate synthetic illicit transaction records and merged them into an augmented training set to improve the effectiveness of credit card fraud detection, and Pascual et al. (2017) proposed a speech-enhancement framework based on GAN. Prior to these publications, Santana and Hotz (2016) proposed an approach for generating images with the same distribution as real driving scenarios. In summary, GANs provide an opportunity to overcome the problem of the lacking training samples for applying deep learning to forecast emerging technologies.

This paper, therefore, proposes a novel approach that integrates data augmentation and deep learning to forecast emerging technologies. First, we built training and testing sets by labeling emerging technology (ET) and non-emerging technology (NET) samples, collecting patent data for each technology, and extracting patent features. Second, a data augmentation method based on a GAN was employed to generate a large amount of synthetic data to train the forecasting model using a deep learning classifier. Finally, we evaluated the performance of the forecasting model with a testing set. We adopted Gartner's emerging technology hype cycles (GETHC) and the Thomson Innovation patent database to forecast technologies that emerged in 2017. The results show that our approach could forecast an ET 1 year before it emerged with high precision. Our proposed approach overcomes the problem of lacking emerging technology samples by combining a GAN with basic deep learning, and the integrated new model was proven to be effective given limited training samples in patents.

This paper is organized as follows. In "Related work" section, related work is presented, and then the research process and methodology are explained in "Methodology" section. Guidelines for the implementation and evaluation of our approach are presented in "Results" section. Finally, our conclusions are provided in "Conclusions" section.

Related work

Forecasting ET based supervised learning

Supervised learning uses a set of known categories of samples to optimize the parameters of the classifier, enabling the classifier to accurately fit the relationships between the features of data samples and the sample categories (Jung and Pedram 2010). Supervised learning approaches include Support Vector Machines (SVMs), Naive Bayes (NB), and Random Forests (RFs). Supervised learning has the following advantages over unsupervised learning: (1) the classifier can effectively introduce external knowledge to increase the reusability and external scalability through the learning of the labeled samples. (2) The trained classifier can automatically and quickly give a sample's category, which greatly reduces the degree of manual participation and prevents subjective biases caused by manual participation (Love 2002; Kyebambe et al. 2017). However, compared with unsupervised learning, supervised learning suffers from the major limitation that it relies on many high-quality labeled samples (Zhu et al. 2006).

According to Kyebambe et al. (2017), using supervised learning to forecast emerging technologies requires the forecasting problem to be transformed into the construction of classifiers. This first involves selecting historical emerging technologies as labeling samples and determining the time when they began to emerge. Second, the historical ET data prior to emergence are collected, and the corresponding data features are extracted. Third, a classifier is built to discover the relationships between the historical data features and the ET or NET categories. Finally, the classifier is used to forecast whether it will become an ET in the future based on historical data for a certain technology (shown in Fig. 1). The general process of constructing the classifier mainly has three steps: labeling samples, constructing data sets, and training and testing classifiers (Fu and Aliferis 2010; Kong et al. 2017; Lee et al. 2018).

Deep learning in bibliometrics

Deep learning, developed by Hinton and Salakhutdinov (2006), has become the key technology of big data intelligence (Zhuang et al. 2017) and has led to major breakthroughs in many fields. Recent studies have also explored the value of deep learning in bibliometrics. Li et al. (2018) proposed an effective patent classification algorithm based on deep learning to solve the large-scale and multiclass patent classification problem and suggested that deep learning has several advantages in large-scale patent classification, including being free of handcrafted features, utilizing straightforward models, and being easy to implement without tedious feature engineering compared with traditional supervised learning algorithms. Hassan et al. (2018) compared deep learning and the classical statistical supervised learning models for classifying the importance of a citation using the same dataset, and the results showed that deep learning with all 64 features had a higher accuracy than SVMs and RFs using the 29 best features. This study also proved the modeling power of deep learning with complex features and large-scale data. Zhang et al. (2018) utilized word embedding, as one such application of deep learning in natural language processing, to map words from vocabulary to vectors and created a method to discover the latent semantics in large-scale text. This study showed the superior performance of deep learning in handling topic extraction tasks in large-scale text data.

The existing studies in bibliometrics have shown that deep learning, as the core of the big data intelligence method, exhibits high powerful modeling power and better performances than classical statistical supervised learning methods, and thus, it holds great potential for forecasting emerging technologies. However, large-scale labeled sample data in existing studies of bibliometrics are required to fully optimize the model parameters and achieve superior performance. A previous study suggested that the quality of a dataset can significantly affect deep learning (Goodfellow et al. 2016). Since there is limited historical emerging technology

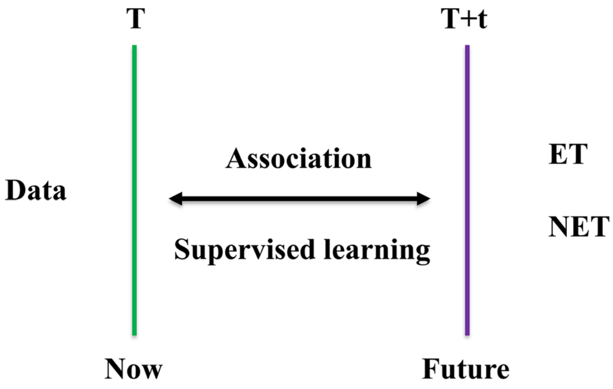


Fig. 1 Schematic diagram of forecasting ET based on supervised learning

data, existing methods for forecasting emerging technologies using deep learning can be easily overfit during the training process, lowering the effectiveness of the forecasting. In other words, under the existing framework, the performance of the deep learning model is decided by the scale and quality of the labeled data. With a small sample size, the high performance of deep learning is generally restricted.

Data augmentation in small-scale samples

In previous studies, to address the insufficiency of the data scale, several oversampling methods were proposed. Their main advantage is that they are self-sufficient. In the early stage, the training set can be enlarged by duplicating the training examples of the minority class if the examples of different classes are imbalanced or by creating a new data set by adding artificial noise (DeRouin et al. 1991). Chawla et al. (2002) proposed a classic oversampling method known as the synthetic minority oversampling technique (SMOTE), which involves the creation of a synthetic minority class data set. Barua et al. (2014) proposed a majority-weighted minority sampling technique with the aim of generating valid synthetic samples. The existing research has mainly focused on imbalanced learning in which better performances can be achieved by adding oversampling instances to the minority class data set. However, data sets in many realms remain insufficient rather than imbalanced in every class. With the development of artificial intelligence and deep learning, GANs have provided opportunities to create new approaches to solve the problem of a lack of samples.

A GAN is a powerful type of generative model (Wang et al. 2017) introduced in 2014 by Goodfellow (Goodfellow et al. 2014). The GAN consists of two deep-architecture functions for the generator and the discriminator, which can learn simultaneously from the trained data in an adversarial fashion compared to oversampling technology (Radford et al. 2015). In the learning process, the generator captures the potential distribution of the real data and generates synthetic samples, while the discriminator discriminates between real samples and synthetic samples as accurately as possible. Recent work has shown that GANs have been successfully applied. Hwang et al. (2018) used a GAN for disease prediction and compared the predictive performance of an auxiliary classifier GAN (AC-GAN) with existing models, such as SVMs and adaptive boosting (AdaBoost), which are widely used in research involving medical data, such as disease prediction. Their results showed that the combination of stacked autoencoders and AC-GAN performed significantly better than existing algorithms for the problem of disease prediction. Fiore et al. (2019) trained a GAN to output mimicked minority-class examples, which were then merged with training data into an augmented training set to improve the classification effectiveness of credit card fraud detection. Zhu et al. (2018) proposed several evaluation methods using three benchmark data sets to validate GAN performance. The empirical results showed that a 5–10% increase in classification accuracy could be obtained after employing GAN-based data augmentation. GANs have been successfully applied in many fields, and they have the potential to solve the problem of limited samples in forecasting emerging technologies.

Methodology

Overall process

Figure 2 shows the overall process of the proposed approach. Given the complexities involved, the proposed approach was designed to be executed in five discrete steps: ET and NET sample labeling, data set construction, data augmentation, forecasting model construction, and evaluation. A detailed discussion of how we performed each step is presented below.

Labeling ET and NET samples

In this study, GETHCs were used to identify ET and NET samples. The GETHC was proposed by Gartner (www.gartner.com), which was established in 1979 as the world's first information technology research and analysis company. The GETHC aims to describe a specific stage of development of an emerging technology (Jun 2012). "Appendix 1" presents the GETHCs from 2008 to 2017.

The foundation of deep learning to forecast emerging technologies is to find the relationships between emerging technologies and their historical data. Thus, when we label the ET and NET samples, we must specify the time when a technology became an ET or NET. Knowing the specific time point, we can find the historical data corresponding to each sample. The GETHC released yearly can effectively solve this problem. According

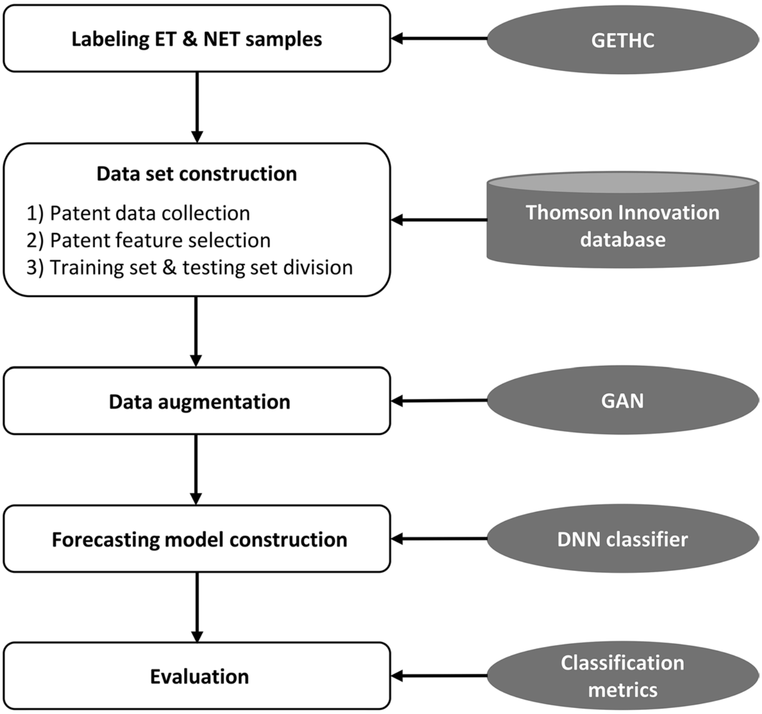

Labeling ET & NET samples
GETHC
Data set construction
1) Patent data collection
2) Patent feature selection
3) Training set & testing set division
Thomson Innovation database
Data augmentation
GAN
Forecasting model construction
DNN classifier
Evaluation
Classification metrics

Fig. 2 Overall process of the proposed approach

to the characteristics of the GETHC, if a technology entered the GETHC in a certain year, this indicates that this technology began to gain attention and showed the characteristics of emerging technology in this year. If a technology dropped out of the GETHC in a certain year, this technology was in last year's GETHC and disappeared in this year's GETHC, which indicates that this technology's influence began to decline and its characteristics as an emerging technology began to disappear. In our method, a technology that entered the GETHC for the first time in a certain year is labeled as an ET sample for that year, and a technology that dropped out of the GETHC in a certain year is labeled as a NET sample for that year. Technology evolves dynamically over time. A technology may enter the GETHC and become an ET in a certain year, and as the technology evolves, it may disappear from the GETHC and become a NET after several years. Thus, the same technology may be labeled as an ET sample in 1 year and may be labeled as a NET sample in another year.

The following are examples of labeling ET and NET samples. *Cloud Computing* entered the GETHC of 2008 for the first time, which indicated that *Cloud Computing* began to garner attention and show the characteristics of an emerging technology in 2008. Thus, we labeled *Cloud Computing* as an ET sample in 2008. As the technology evolved, *Cloud Computing* dropped out of the GETHC of 2015, which means this technology was in the GETHC of 2014 and disappeared in the GETHC of 2015. This indicates that *Cloud Computing's* influence began to decline, and its characteristics as an emerging technology began to disappear in 2015. Thus, we labeled *Cloud Computing* as the NET samples in 2015. The same method was used for labeling ET and NET samples in other years.

Data set construction

After labeling ET and NET samples, patents for the corresponding technology were retrieved from the patent database. We subsequently extracted patent features and used them to create a patent feature vector of each ET and NET sample. Finally, all the patent feature vectors of the ET and NET samples were used to construct the training set and testing set for the training and evolution of the forecasting model.

Patent data collection

When collecting historical patent data corresponding to each ET and NET sample from the patent database, patent data should be collected according to the time point when each sample become an ET or a NET. The purpose of the proposed method is to forecast whether a certain technology will become an emerging technology 1 year later. Therefore, we need the deep learning model to find the relationships between an ET or NET sample that becomes an ET or a NET in year T and all the corresponding historical patent data until year $T-1$. We collected historical patent data until year $T-1$ for each ET and NET sample in year T. The patent data were collected from the Thomson Innovation (TI) patent database, which contains comprehensive and high-quality patent data.

According to the sample labeling method in "Labeling ET and NET samples" section, technology may be labeled as an ET sample and a NET sample in different years. However, when collecting the corresponding patents, the historical patent data of this technology as a sample of ET and NET is different because of the different years when this technology became an ET and a NET. When the same technology is labeled as an ET sample and a NET sample in different years, for the ET sample in year T, we collect historical patent data until year $T-1$, and for the NET sample in year T, we collect historical patent data from the year it becomes an ET sample to the year $T-1$.

The following are examples of collecting patents. *Cloud Computing* is used as an example for the labeling of ET and NET samples. This technology was labeled as an ET sample in 2008. Thus, we collected the corresponding patents until 2007 for *Cloud Computing*. As the technology evolved, *Cloud Computing* was labeled as an NET sample in 2015. Thus, we collected the corresponding patents from 2008 to 2014 for *Cloud Computing*. The same method was used to collect corresponding patents of each ET and NET sample in other years.

Patent feature selection

Patent feature selection is crucial to this study, because forecasting accuracy greatly depends on the relevance of the patent features for emerging technologies. Earlier reports have presented a variety of patent features to capture the characteristics of emerging technologies. This study employed a total of seven features from previous studies to capture the key characteristics of emerging technologies, which are divided into five sub-categories (shown in Table 1): low-novelty, science-intensity, growth speed, scope and coverage, and development effort and capabilities (Lee et al. 2018). These features can be extracted from patent databases immediately after the relevant patents are collected.

(1) *Low-novelty* Earlier studies have explored the relationship between patent novelty and backward citations. Some studies suggested that the breadth of patent backward citations can capture the absorption of previous technological innovations from many different fields and the novelty of a patent increasing with the breadth of the backward citations (Hall and Helmers 2013; Hall et al. 2013). Some other studies suggested that patent backward citations can capture the related prior publications and patents in the patent applications and patents with large numbers of backward citations have a relatively low novelty and low monetary value (Harhoff et al. 2003; Lee et al. 2018). In this study, we consider that the number of backward citations tend to directly measure the number of related prior publications and patents and the breadth of the backward citations requires other metrics.

(2) *Growth speed* Since the growth speed cannot be fully observed at early stages of technology development without continuous monitoring and surveillance, the *technology cycle time (TCT)* was employed in this study, which captures the degree of newness of prior knowledge or the pace of technology progress (Bierly and Chakrabarti 1996; Kayal and Waters 1999) in this sub-category.

(3) *Science-intensity* Earlier studies have suggested more scientific knowledge contained in the patented invention may lead to the development of more innovative and influential technology (Cozzens et al. 2010; Day and Schoemaker 2000). The *number of non-patent citations* (Trajtenberg 1990) was employed in this study to capture the closeness to scientific knowledge.

(4) *Scope and coverage* This sub-category contains three patent features to represent the scope and coverage of patents. The first feature is the *number of claims*, which indicates the scope of the legal protection conferred by patent claims (Lanjouw and Schankerman 2004). The second feature is the *family patent size*, which corresponds to the number of times the same invention has been developed in different countries and represents the economic and technical importance of the invention (OuYang and Weng 2011). The third feature is the *number of International Patent Classification (IPC)*, which represents the scope of the technological fields of a patent (Lee et al. 2009).

Table 1 Patent feature descriptions

Categories	Feature	Description
Low-novelty	Backward citations	The number of patents cited by this patent
Growth speed	Technology cycle time (TCT)	The interval between two adjacent generations of technology is represented by the median value of the patent citation interval. The formula is $\mathrm{TCT}_i = \mathrm{median}_j\left\{\left\|T_i - T_j\right\|\right\}$, where i represents a patent and j represents the patent cited by i
Science-intensity	Number of non-patent citations	The number of non-patent citations in patents reflects the relationship between technological innovation and research in this field
Scope and coverage	Number of claims	The number of rights protections required by a patent at the time of application, which reflects the importance of the patent
	Family patent size	The number of applications for the same invention in different countries represents the economic and technical importance of the invention
	Number of IPC	The number of IPC for a patent represents the technical complexity of a single patent in the field
Development capabilities	Forward citations	The number of times this patent is cited by other patents

(5) *Development capabilities* Many empirical studies have found that there is a significant positive relationship between forward citations and development capabilities (Lerner 1994; Martin 1995). This study employs *forward citations*, which captures the development capabilities.

At this point, each ET and NET sample comprises many patents. However, for the subsequent data augmentation and deep learning analysis, we must represent each ET and NET sample as a single feature vector that has the same features as the individual patents. As such, drawing on previous research (Kyebambe et al. 2017), for a given ET or NET sample, we compute the feature value by taking the arithmetic mean of values for the constituent patents. For example, if an ET or NET sample comprises 100 patents, we first compute the seven features of each patent, after which we compute the average of each feature for all 100 patents, and finally we construct a 7-dimensional feature vector representing this sample with the average value of each feature.

Training set and testing set construction

The feature vector of each ET and NET is calculated according to the patent features described in "Patent feature selection" section. We obtain a set of patent feature vectors labeled as ET or NET. To precisely evaluate the performance of the forecasting model, the set of patent feature vectors is randomly divided into a training set and a test set at a ratio of 7:3. Due to the small scale of the training samples, we utilize a data augmentation method based on a GAN to generate a large number of synthetic samples to train the forecasting model and improve the forecasting accuracy. In the next section, we briefly explain the process used for data augmentation based on the GAN.

Data augmentation based on GAN

To solve the problem of a lack of emerging technology samples, the approach of data augmentation using a GAN is proposed to enlarge the data scale of emerging technology samples. After the ET and NET training sets are built, we use the original ET and NET training sets to build the corresponding GAN and generate synthetic samples. The workflow of generating ET or NET synthetic samples based on GAN involves two steps (shown in Fig. 3):

(1) The generator of the GAN begins to generate the original synthetic samples when the loss functions of the generator and discriminator of the GAN converge after being trained using the ET or NET training sets several thousand times.

(2) The trained generator of the GAN is used to generate the synthetic samples and the discriminator is used to filter these samples. In the synthesized ET or NET samples created by the GAN generator, samples that fool the discriminator are selected as the final synthesized samples. According to the adversarial idea in GANs (McDaniel et al. 2016), the generator attempts to generate synthetic samples that can fool the discriminator while the discriminator tries to distinguish between real samples and synthetic samples. This means that when ET or NET synthetic samples were discriminated as real by the discriminator, the synthetic samples were more akin to the distribution of the real ET or NET training sets.

Training of the GAN involves finding the parameters of a discriminator (D) that maximize its classification accuracy and finding the parameters of a generator (G) that maximally

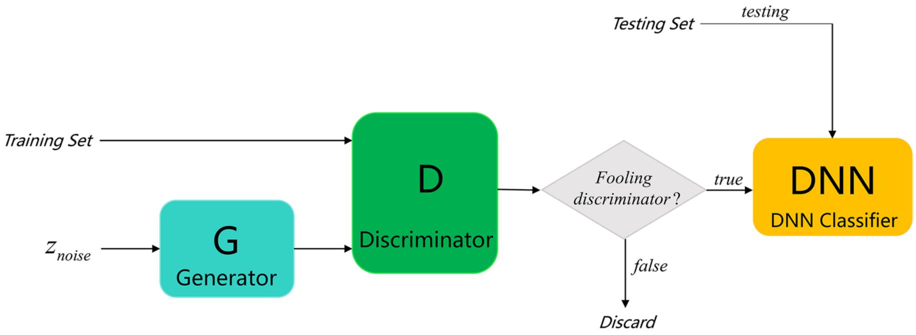


Fig. 3 Workflow of data augmentation using GAN

confuses the discriminator. The cost of training is evaluated using a value function, which is defined in Eq. 1, that depends on D and G. During training, G and D play a minimax game with the value function, D and G are updated, and the iteration stops until a Nash equilibrium is achieved. In greater detail, $D(s)$ is the probability that s comes from the real data, $G(z)$ is the synthetic sample that is generated by the generator, $D(G(z))$ is the probability that the synthetic sample is discriminated as real by the discriminator. Equation 1 is as follows:

$$\max_{D}\min_{G} V(D,G) = E_{P_{\text{data}}(s)}[\log D(s)] + E_{P_z(Z)}[\log(1 - D(G(z)))]. \tag{1}$$

The hyperparameters of the GAN include the hyperparameters of the generator and discriminator. The generator and discriminator of the GAN both have deep neural network (DNN) structures. The input of the generator is white noise, the number of input units is equal to the dimension of the white noise, the output is the synthetic sample, and the number of output units are equal to the number of selected patent features, which is seven. The number of hidden layers in the network and the optimal number of units per layer must be experimentally determined. The input of the discriminator is a real or synthetic sample, and the output is the category of the real or synthetic samples. The number of input units is equal to the number of selected patent features, which is seven, and the output unit is one no-activation-function unit. The optimal number of hidden layers in the neural network and the number of units per layer must be experimentally determined.

Forecasting model construction based on DNN classifier

After the augmentation of the original ET and NET samples, the GAN will generate a large number of synthetic samples. To make full use of the advantages of big data, we constructed a DNN classifier based on deep learning to forecast emerging technologies. DNN classifiers based on deep learning are complex and have larger model capacities. After extensive training on large-scale labeled samples, they can exhibit superior performances (Goodfellow et al. 2016). Meanwhile, the multilayered neural network structure of a DNN can learn the multilevel abstract features of sample data in which high-level features are constructed by low-level-feature combinations, which can more effectively express the distribution characteristics of the data and produce a better learning result than the classical statistical supervised learning model (Bengio and Lecun 2007).

The construction of the DNN classifier includes training and testing. First, the DNN classifier is trained with many synthetic ET and NET samples generated by the GAN. The DNN classifier is subsequently tested with partially independent real ET and NET samples. Testing DNN classifiers with real independent ET and NET samples simulates the real forecasting environment and can effectively reflect the general performance of a classifier. The input of the DNN classifier is a synthetic sample, and the output is the ET or NET classification. The number of input units is equal to the number of selected patent features, which is seven. The number of output units is equal to the number of categories, which is two. The number of neural network layers and the number units per layer must be determined experimentally.

When using the trained DNN classifier to forecast emerging technologies, it is necessary to collect the corresponding patent data for the technology to be forecasted and extract the patent features to construct the feature vector. By inputting the vector of one technology to be forecasted into the DNN classifier, the DNN classifier can directly and automatically forecast whether the technology will become an emerging technology in the next year.

Evaluation

To test the performance of the DNN classifier, we used three classification metrics based on a confusion matrix (Table 2): accuracy, F1, and G-mean. Accuracy is the proportion of predictions that were correct, F1 is the harmonic mean of the precision and recall, and the G-mean indicates the geometric mean of the recall (Sun et al. 2007). Accuracy, F1, and G-mean are defined as follows:

$$\text{Accuracy} = \frac{\text{TN} + \text{TP}}{\text{TN} + \text{TP} + \text{FP} + \text{FN}}, \tag{2}$$

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{3}$$

$$\text{G - mean} = \sqrt{\frac{\text{TN}}{\text{TN} + \text{FP}} \cdot \frac{\text{TP}}{\text{TP} + \text{FN}}}. \tag{4}$$

In these equations, TP, TN, FP, and FN are the number of true positive samples, true negative samples, false positive samples, and false negative samples, respectively. Further, recall and precision are defined as follows:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{5}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{6}$$

Table 2 Diagram of confusion matrix

	Predicted positive sample	Predicted negative sample
True positive sample	TP	FN
True negative sample	FP	TN

Results

Analysis result of the proposed approach

Based on the GETHCs from 2008 to 2017, 57 ET and 48 NET samples were extracted. Tables 3 and 4 list the samples of each ET and NET. We retrieved and downloaded the patents for all the ET and NET samples and calculated the patent feature vector for each technology to construct the data set. The details of the data set are shown in “Appendix 2”.

In “Appendix 3”, patent feature descriptive statistics are reported. The results of the descriptive statistics showed that different patent features have different distributions, and the same patent features between ET and NET samples also had different distributions. In “Appendix 4”, from the results of Pearson correlation analysis, *Number of IPC* had significant positive correlations with *Number of claims* and *Family patent size*. These three features were used to represent the scope and coverage of the patent, and the correlation

Table 3 Results of ET samples

Name of technology	Year	Name of technology	Year
5G	2017	Consumer telematics	2012
Cognitive computing	2017	Hybrid cloud computing	2012
Deep learning	2017	Big data	2011
Digital twin	2017	Group buying	2011
Edge computing	2017	Image_Content recognition	2011
Deep reinforcement learning	2017	NFC payments	2011
4D printing	2016	QR_Color code	2011
802.11ax	2016	IOT	2011
Blockchain	2016	Social TV	2011
Neuromorphic hardware	2016	Activity streams	2010
Nanotube electronics	2016	Autonomous vehicles	2010
Digital dexterity	2015	Broadband over power lines	2010
Machine learning	2015	Consumer generated media	2010
Micro data centers	2015	Media tablets	2010
IOT platforms	2015	Tangible user interfaces	2010
Data science	2014	Terahertz waves	2010
Digital security	2014	Interactive TV	2010
Connected home	2014	Virtual assistants	2010
Affective computing	2013	Ebook readers	2009
Biochips	2013	Home health monitoring	2009
Electrovibration	2013	Human augmentation	2009
Mobile health monitoring	2013	Video search	2009
Virtual reality	2013	Wireless power	2009
3D scanners	2012	Internet TV	2009
Application stores	2012	Cloud computing	2008
Complex event processing	2012	Solid state drives	2008
Crowdsourcing	2012	Surface computers	2008
HTML5	2012	Green IT	2008
Volumetric and holographic displays	2012		

Table 4 Results of NET samples

Name of technology	Year	Name of technology	Year
802.11ax	2017	Crowdsourcing	2013
Micro data centers	2017	Home health monitoring	2013
3D bioprinting	2016	HTML5	2013
Affective computing	2016	Image_Content recognition	2013
Biochips	2016	Internet TV	2013
Cryptocurrencies	2016	Media tablets	2013
Digital security	2016	Private cloud computing	2013
Hybrid cloud computing	2016	Social analytics	2013
IOT	2016	Wireless power	2013
3D scanners	2015	Ebook readers	2012
Activity streams	2015	Group buying	2012
Big data	2015	QR_Color code	2012
Cloud computing	2015	Social TV	2012
Complex event processing	2015	3D flat panel displays	2011
Data science	2015	Broadband over power lines	2011
Gamification	2015	Consumer generated media	2011
In memory analytics	2015	Interactive TV	2011
Machine to machine comm. services	2015	Microblogging	2011
Mobile health monitoring	2015	Terahertz waves	2011
NFC payments	2015	Video search	2011
Prescriptive analytics	2015	Green IT	2010
Predictive analytics	2014	IP Video_Internet video	2010
Application stores	2013	Surface computers	2010
BYOD	2013	Solid state drives	2009

analysis results were consistent with previous literature studies. Furthermore, we found that *Number of claims* had significant positive correlations with *Backward citations*, *Number of non-patent citations,* and *Forward citations* have significant positive correlations with *Backward citations*, *Number of non-patent citations*, *Number of claims*. Although the features to capture emerging technologies in previous research were divided into five different sub-categories. The correlation analysis results showed that the sub-categories were not completely independent and had certain correlations. However, there were no strong correlations between the different sub-categories, with the correlation coefficients not exceeding 0.6. Since the multi-layer nonlinear structure of deep learning does not require strict independence of features (Valmadre et al. 2017), we assumed that the correlations between the selected features would not have a significant impact on the performance of the deep learning.

According to the data augmentation method, we first use the ET and NET samples in the training set to train the corresponding GAN and then used the trained GAN to generate synthetic samples. The GAN consisted of two deep-architecture functions for the generator and the discriminator, as many hyperparameters could influence the performance of a GAN. The number of layers in the generator and discriminator was a fundamental hyperparameter. Too few layers would hinder the ability of the network to build a representation at a level of abstraction to adequately capture the data complexity, and too many layers would complicate the training substantially and likely result in overfitting (Fiore et al. 2019). As a reasonable tradeoff, networks with two and three hidden layers were tested in the generator and discriminator. The optimal parameter was determined by the convergence value of the loss function of the generator and discriminator. Since training and tuning a GAN is an expensive operation, we conducted a limited number of experiments, in which 4, 8, 12, 16, and 32 nodes with two and three layers were tested. Through a series of experiments on the GAN, the best-performing hyperparameters were determined. The generator had two hidden layers containing four ReLU units, seven softmax units are used as the output layer, and the dimension of the noise vector z was set to four. The discriminator also had two hidden layers containing four ReLU units, and one no-activation-function unit was used as the output layer. The ET and NET GAN had the same hyperparameters. In each iteration of the GAN training, the discriminator first iterated 100 times, and then the generator iterated once. The GAN development environment was TensorFlow 1.1 with Python 3.5.2, and it was trained through a GPU.

After the synthetic samples were generated, they were used to train the DNN classifier. The DNN classifier was subsequently tested using the test set samples. Hyperparameters for the DNN classifier were empirically determined using a similar procedure as that used to determine the hyperparameters of the GAN. We also conducted a limited number of experiments, in which 4, 8, 12, 16, and 32 nodes with two and three hidden layers were tested. The optimal parameter was determined by the accuracy, F1, and G-mean. Through a series of experiments on the DNN classifier, the best-performing hyperparameters of the DNN classifier were determined. The classifier had two hidden layers, each containing 32 ReLU units. Two softmax units were used as the output layer, the dimension of the classifier's input was seven, and cross entropy was used as the loss function. The number of iterations was set to 3000. The DNN classifier's development environment was TensorFlow 1.1 with Python 3.5.2, and it was trained through GPU.

Existing studies on data augmentation have shown that the number of synthetic samples significantly affects the performance of supervised learning models (Natten 2017; Fiore et al. 2019). To evaluate the effect of the synthetic training ET and NET sample size on the performance of the DNN classifier, we used a different number of synthetic ET and NET samples to train the DNN classifier and used test samples to evaluate the performance through the accuracy, F1, and G-mean. From 100 synthetic training samples of each class, more than 100 synthetic samples were generated each time to train the DNN classifier, and the changes in the accuracy, F1, and G-mean are shown in Fig. 4. When the number of ET and NET synthetic samples exceeded 1000, the accuracy, F1, and G-mean no longer increased but fluctuated within a certain range. In other words, when the synthetic ET and NET training sample size was 1000, the performance of the GAN began to converge, and an effective DNN classifier for forecasting ETs could be obtained.

Table 5 shows the forecasted results for the DNN classifier on the test set data when the number of synthetic ET and NET samples generated by the GAN was 1000. A total of 14 ETs and four technologies were identified as NETs with a precision of 71%. There was a total of 17 NETs, and three technologies were identified as ETs with a precision of 82%. There were seven technologies wrongly forecasted in the test set data, indicating that the forecast model could forecast ETs with an accuracy of 77% 1 year before their emergence.

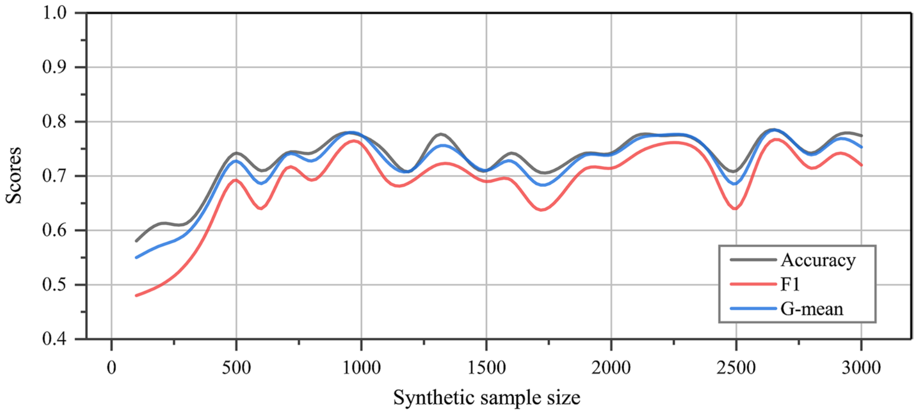


Fig. 4 Effect of synthetic training sample size on DNN classifier performance

Table 5 Confusion matrix of forecasting emerging technologies

	Prediction		Total
	ET	NET	
ET	10	4	14
NET	3	14	17
Total	13	18	31

Table 6 Performance of different classification strategies

Model	Accuracy	F1	G-mean
SVM	0.6522	0.4286	0.5477
NB	0.6957	0.3636	0.4830
RF	0.5652	0.3750	0.4804
GAN-DNN	0.7742	0.7407	0.7670

Evaluation of proposed approach

To evaluate the effectiveness of the proposed method, we used statistical supervised learning classifiers, SVM, NB, and RF, for comparison experiments. As classic supervised learning classifiers, SVM, NB, and RF exhibit higher classification accuracies and better general performances than other statistical supervised learning classifiers. The results of the comparative experiments in the two categories of ET and NET are shown in Table 6.

Table 6 shows that for the same data set, the accuracy, F1, and G-mean of the SVM, NB, and RF supervised learning classifiers were lower than those of our proposed GAN-DNN. The comparison of the evaluation indicators shows that the forecasting quality of the classical supervised learning classifiers was lower than that of the combined GAN-based data augmentation and DNN-based forecasting model proposed in this study. The results of the comparative experiments showed that our approach enabled us to obtain an effective forecasting model based on the GAN and DNN classifier without large-scale ET and NET samples.

Validation of forecasting model

To validate the forecasting effect of the proposed method in a real environment, the model must be trained based on the current data and forecast whether technology will become an emerging technology in the future. Therefore, we applied the method to the available data in 2016 to make predictions for 2017 and validated the forecasting effect based on the real data for 2017. According to the ET and NET sample labeling method in "Labeling ET and NET samples" section, whether technology will become an ET or NET in the next year is measured by whether the technology will enter the GETHC for the first time or disappear from the curve in the next year.

In this paper, we utilize the historical data from 2000 to 2016 to train forecasting model to forecast whether technology will become an emerging technology in 2017, and utilize the forecasting results of 2017 to validate the forecasting effect. Firstly, all the ET and NET samples from 2000 to 2017 were labeled from the GETHC according to the proposed method. Next, the patent feature vector corresponding to the ET and NET samples from 2000 to 2016 were adopted to train the forecasting model. The ET and NET samples from 2017 were not used in the training. Finally, the forecasting effect of the forecasting model was validated with the ET and NET samples labeled in 2017 from GETHC. The model parameters of the GAN and DNN classifier were consistent with those in "Analysis result of the proposed approach" section and the number of synthetic ET and NET samples was also 1000.

For 2017, six ET samples and two NET samples were labeled from the GETHC. The 6 ETs were *5G, deep learning, edge computing, cognitive computing, digital twin*, and *deep reinforcement learning*. The two NETs were *802.11ax* and *micro data centers. 5G* and *cognitive computing* were incorrectly forecasted, and the other four ETs were correctly forecasted by the proposed method. The forecasting results for the ET samples in 2017 showed that our method based on GAN and DNN could forecast emerging technologies 1 year before they emerged with high precision and few samples.

Conclusions

A novel approach for forecasting emerging technologies using data augmentation and deep learning was proposed in this study. The essence of this proposed approach was to integrate data augmentation (GAN) and deep learning, which enabled deep learning to effectively forecast emerging technology with limited training samples. Specifically, this paper constructed a sample data set of emerging technologies from the GETHC and TI patent database, and subsequently used a GAN to augment the sample data set and construct a forecasting model based on DNN classifiers. The test results showed that the forecasting accuracy reached 77% when the synthetic sample size was 1000. Finally, this approach was used to forecast technology in 2017. Four of the six emerging technologies were correctly forecasted. This verified that the model could, given limited samples, forecast emerging technologies 1 year before they emerged with high precision.

The contributions of this research are twofold. First, this study contributes to technology forecasting literature by proposing a novel approach that advances the basic deep learning method for forecasting emerging technology. In previous research, large-scale labeled sample data was required to fully optimize the parameters of the deep learning model and obtain a superior performance compared with the other traditional supervised learning methods. Our proposed approach utilized a GAN to overcome the problem of lacking training samples, and the integrated new model was proven to be effective, even without large training samples in the patents. Second, from a practical perspective, the proposed approach is more effective than previous unsupervised methods when embedding external knowledge into the forecasting model through deep learning classifier. After a forecasting-model-based deep learning classifier was constructed, we can obtain the forecasting results effectively on a real-time basis without requiring extra work for experts' interpretation, which is usually less-efficient and may lead to significant biases in technology forecasting.

The main objective of this paper was to use a GAN to overcome the problem of lacking training samples. Thus, we selected patent features that were simple and directly verified the effectiveness of the proposed method. In this study, the patent features we explored were all external features that had better consistency in theory, and they were selected through a review of prior literature. The empirical results showed that all these patent features had strong correlations with emerging technologies. Compared to the external features used in this paper, internal semantic patent features based on text mining and semantic analysis may elicit patent information more deeply and comprehensively. However, this requires more complex methods and would increase the uncertainty of the method in the feature extraction stage, creating difficulties for the verification of the effectiveness of the proposed method. Thus, in this study, we chose not to consider internal semantic patent features; however, it would be valuable to explore this concept in future research.

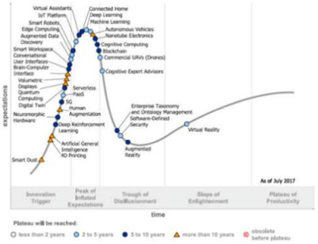
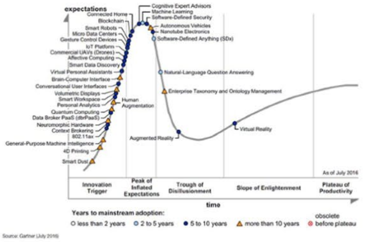
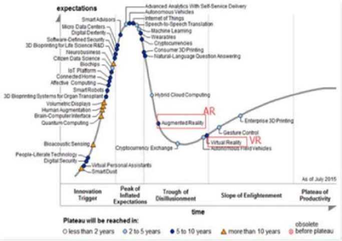
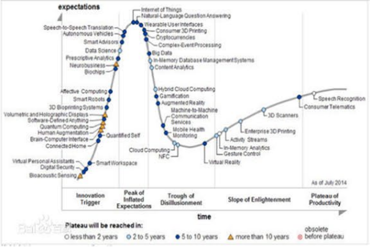
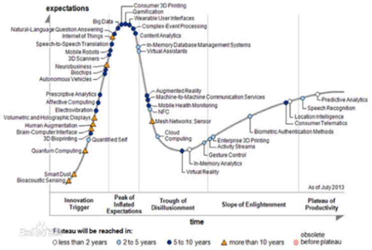
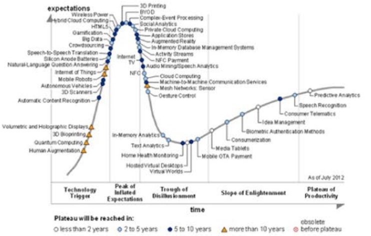
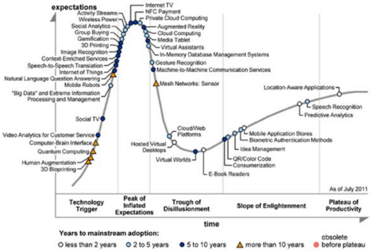
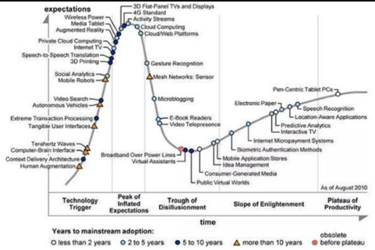
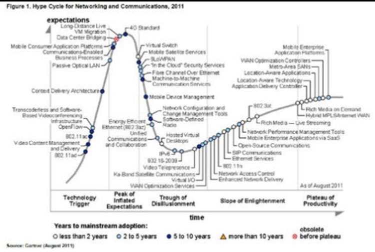
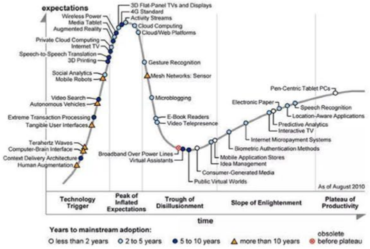
Acknowledgements This research was supported by the National Natural Science Foundation of China (91646102, L1824039, L1724034, 71872019, L1724030, L1824040, L1624045, and L1524015, 71203117), the MOE (Ministry of Education in China) Project of Humanities and Social Sciences (Engineering and Technology Talent Cultivation) (16JDGC011), the UK-China Industry Academia Partnership Programme (UK-CIAPP\260), the Volvo-Supported Green Economy and Sustainable Development Tsinghua University (20153000181), the Tsinghua Initiative Research Project (2016THZW), the Chinese Academy of Engineering's China Knowledge Centre for Engineering Sciences and Technology Project (CKCEST-2019-2-13), the Fundamental Research Funds for the Central Universities (2018RC41 and 2018XKJC04), the Beijing Natural Science Foundation (9182013), the Beijing Social Science Foundation (Grant Number: 17GLC058), and China Knowledge Centre for Engineering Sciences and Technology (CKCEST-2018-1-13).

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Appendix 1

See Table 7.

Table 7 Gartner's emerging technology hype cycles from 2008 to 2017

 2017	 2016
 2015	 2014
 2013	 2012
 2011	 2010
 2009	 2008

Appendix 2

See Tables 8 and 9.

Table 8 Results of ET samples and their patent features

Name of technology	Year	Number of patents	Backward citations	TCT	Number of non-patent citations	Number of claims	Family patent size	Number of IPC	Forward citations
5G	2017	14,869	6.421	7.137	1.881	16.619	4.621	6.571	6.78
Cognitive computing	2017	73	1.123	5.252	2.479	16.849	1.658	3.247	6.26
Deep learning	2017	568	1.011	2.378	0.921	11.739	1.468	2.864	2.424
Digital twin	2017	17	0.706	5.799	0.353	13.941	1.647	3.941	2.941
Edge computing	2017	217	8.083	4.639	1.779	16.585	2.12	2.949	6.433
Deep reinforcement learning	2017	5	0	3.8	0	13.8	1.4	4.2	0.6
4D printing	2016	29	2.966	6.436	0.069	10	2.276	4.724	3.862
802.11ax	2016	35	2	3.417	2.314	25.371	3.114	5.086	8.629
Blockchain	2016	44	2.818	2.967	6.773	18.432	2.386	3.568	10.455
Neuromorphic hardware	2016	5	0	3.692	12.2	23.4	1.4	3	3.6
Nanotube electronics	2016	66	10.758	4.866	1.939	16.955	3.258	6.197	5.258
Digital dexterity	2015	11	11.818	16.897	0.364	19.636	2.091	3	20.364
Machine learning	2015	3060	9.541	5.567	6.158	18.458	2.658	3.208	12.651
Micro data centers	2015	7	0.571	2.981	1.429	13	2.286	2.714	5.429
IOT platforms	2015	18	5.889	2.807	3.722	12.389	1.889	3.278	19.722
Data science	2014	23	4.565	5.962	6.565	12.783	2.174	2.304	23.348
Digital security	2014	209	16.124	5.675	2.086	15.048	2.967	3.852	13.555
Connected home	2014	683	6.848	4.36	0.559	10.93	2.654	3.587	5.032
Affective computing	2013	65	14.969	5.382	8.292	16.2	2.477	3.323	16.108

Table 8 (continued)

Name of technology	Year	Number of patents	Backward citations	TCT	Number of non-patent citations	Number of claims	Family patent size	Number of IPC	Forward citations
Biochips	2013	1023	8.245	6.668	2.956	16.805	4.811	7.413	7.868
Electrovibration	2013	6	26	3.379	1.333	11.5	1.333	2	10
Mobile health monitoring	2013	5	8.6	3.738	0.4	17	2.6	2.8	4.2
Virtual reality	2013	4598	21.373	5.442	2.558	14.684	2.711	3.776	10.529
3D scanners	2012	399	11.348	5.731	1.719	14.782	3	3.657	8.238
Application stores	2012	1755	22.031	5.118	3.818	18.914	2.921	3.678	18.293
Complex event processing	2012	84	15.929	4.585	25.333	19.405	2.738	2.655	34.214
Crowdsourcing	2012	17	18.824	3.883	3.294	22.235	2.176	3.059	13.824
HTML5	2012	49	9.571	3.825	2.082	16.082	2.796	2.959	7.898
Volumetric and holographic displays	2012	6	15.333	8.568	12	18.5	1.5	5.833	34.167
Consumer telematics	2012	1	10	11	0	22	2	3	3
Hybrid cloud computing	2012	7	35.286	5.284	2	20.143	2.714	2.857	15.714
Big data	2011	80	5.763	3.857	0.288	8.1	1.838	3.2	3.325
Group buying	2011	111	13.234	2.226	3.405	16.441	1.892	1.694	5.324
Image_Content recognition	2011	4971	8.351	4.146	1.027	9	2.24	3.824	4.638
NFC payments	2011	5	37	5.658	0.2	13.8	2	3.4	7.4
QR_Color code	2011	3246	11.398	8.233	1.314	14.408	2.557	3.597	9.084
IOT	2011	125	7.024	6.815	0.896	12.248	3.504	3.944	5.52
Social TV	2011	7	21.857	6.133	7.143	23.143	2.143	5.429	26.286
Activity streams	2010	106	25.283	4.358	2.66	16.887	3.368	3.547	12.896

Table 8 (continued)

Name of technology	Year	Number of patents	Backward citations	TCT	Number of non-patent citations	Number of claims	Family patent size	Number of IPC	Forward citations
Autonomous vehicles	2010	409	27.513	8.116	3.45	15.159	3.178	4.822	15.093
Broadband over power lines	2010	46	30.022	8.631	2.478	21.022	3.261	3.913	29.457
Consumer generated media	2010	30	20.767	4.382	1.767	13.267	2.633	3.433	8.9
Media tablets	2010	27	12.704	6.399	3.556	9.185	2.889	4.074	9.593
Tangible user interfaces	2010	17	8.765	4.476	0.706	13	2.588	3.882	2.412
Terahertz waves	2010	401	6.561	5.401	1.693	10.157	2.85	4.042	4.648
Interactive TV	2010	1037	40.473	5.546	3.175	25.288	3.596	5.091	24.144
Virtual assistants	2010	50	30.52	5.111	2.94	24.34	2.82	3.7	14.58
Ebook readers	2009	11	27.545	5.294	22.636	26.636	3.545	3.545	79.818
Home health monitoring	2009	19	111.895	8.79	2.737	23.895	2.211	3.684	46.158
Human augmentation	2009	9	17.556	5.671	5.778	36	3.667	10	18.667
Video search	2009	281	17.391	3.956	1.868	13.665	2.747	4.267	8.594
Wireless power	2009	455	35.49	5.585	1.42	15.844	4.105	4.492	12.393
Internet TV	2009	438	12.683	3.397	0.804	11.258	2.215	3.034	8.514
Cloud computing	2008	4	13	4.25	2	10	2.5	4.75	12
Solid state drives	2008	96	21.646	6.146	1.052	16.708	4.896	4.448	10.417
Surface computers	2008	358	15.718	6.34	1.802	11.651	2.972	4.05	6.045
Green IT	2008	71	10.324	8.384	1.563	11.028	2.127	3.099	5.535

Table 9 Results of NET samples and their patent features

Name of technology	Year	Number of patents	Backward citations	TCT	Number of non-patent citations	Number of claims	Family patent size	Number of IPC	Forward citations
802.11ax	2017	61	0.279	2.073	3.738	22.213	1.557	3.59	6.803
Micro data centers	2017	15	0.067	2.639	0.467	12.267	1.467	3.267	3.6
3D bioprinting	2016	8	0.75	5.519	0.875	19.125	2.875	6.75	3.125
Affective computing	2016	22	1.727	8.065	2.955	12.864	2.773	3.273	12.045
Biochips	2016	242	3.302	9.054	9.136	13.397	6.165	8.521	11.227
Cryptocurrencies	2016	8	3	3.443	1.625	15.625	1.25	2.5	10.25
Digital security	2016	60	2.217	6.064	3.433	15.2	3.867	4.717	42.833
Hybrid cloud computing	2016	48	6.813	3.962	1.979	18.583	2.104	3.354	7.833
IOT	2016	1487	2.243	3.929	1.255	13.578	2.176	3.513	6.504
3D scanners	2015	420	4.96	6.046	2.219	11.579	2.888	3.876	9.624
Activity streams	2015	166	8.223	6.062	6.741	19.729	3.289	4.151	35.259
Big data	2015	426	3.934	5.029	1.779	16.122	2.491	2.85	7.66
Cloud computing	2015	8273	8.259	4.442	2.94	14.072	2.209	2.928	12.783
Complex event processing	2015	101	5.188	4.979	6.99	15.98	2.901	2.832	13.713
Data science	2015	7	1.143	5.415	5.286	8.857	1.857	2.143	7.857
Gamification	2015	66	4.591	5.784	2.455	18.53	2.091	3.455	14.121
In Memory analytics	2015	7	2.857	7.639	4.429	13.143	3.571	4.571	12.714
Machine to machine comm. services	2015	7	1.714	3.9	1.286	11.857	2.429	3.143	7.714
Mobile health monitoring	2015	6	3.5	5.423	0.167	8.667	1.667	2.667	5.833
NFC payments	2015	73	4.877	4.384	0.726	13.452	2.014	3.288	9.699
Prescriptive analytics	2015	3	0	4.056	2.667	16	3.667	4	6
Predictive analytics	2014	82	13.988	6.612	8.902	20.537	2.671	2.585	19.134
Application stores	2013	434	7.959	5.842	2.631	17.857	3.657	4.32	14.615
BYOD	2013	9	61.111	4.744	3.333	19.444	2	3.444	79.556

Table 9 (continued)

Name of technology	Year	Number of patents	Backward citations	TCT	Number of non-patent citations	Number of claims	Family patent size	Number of IPC	Forward citations
Crowdsourcing	2013	47	6.872	5.112	6.511	23.064	2.574	3.277	17.277
Home health monitoring	2013	9	4.333	6.438	0.333	16.333	2.111	3	7
HTML5	2013	89	6.629	4.323	1.258	13.618	2.18	2.618	7.438
Image_Content recognition	2013	1287	3.713	5.648	1.106	10.983	3.044	4.039	7.793
Internet TV	2013	396	7.944	5.42	1.179	15.321	3.02	4.313	12.093
Media tablets	2013	21	4.286	6.924	0.429	15.81	5	4.762	6.476
Private cloud computing	2013	26	13.231	5.079	2.654	15.692	3.269	3.308	8.269
Social analytics	2013	4	3.75	2.346	1.75	20.25	1	3.75	30.25
Wireless power	2013	2267	20.425	4.907	3.054	14.921	4.124	4.471	19.316
Ebook readers	2012	62	13.339	5.898	10.468	22.645	2.339	2.726	33.903
Group buying	2012	24	3.333	3.542	0.542	15.833	1.792	1.625	4.25
QR_Color code	2012	654	5.217	6.732	2.139	13.047	3.147	3.74	9.687
Social TV	2012	2	88.5	5.375	0.5	19	4	13	8
3D Flat panel displays	2011	1	7	3.5	1	15	14	17	5
Broadband over power lines	2011	7	1.714	7.357	0.571	14.286	4	4.143	3.571
Consumer generated media	2011	7	10.286	7.197	0.571	12.714	2.143	2.714	8.571
Interactive TV	2011	130	27.9	7.879	3.115	23.292	6	6.931	36.562
Microblogging	2011	96	15.792	4.488	6.75	20.76	3.094	3.604	11.51
Terahertz waves	2011	147	5.088	6.369	1.367	9.429	3.265	3.639	4.347
Video search	2011	86	11.605	4.196	4.616	15.733	3.256	3.477	7.849
Green IT	2010	22	5.136	4.182	2.182	7.364	2.591	3.136	2.864
IP Video_Internet video	2010	72	12.972	6.216	0.819	17.361	3.458	4.042	11.542
Surface computers	2010	118	25.568	6.85	5.432	17.797	4.331	4.898	23.483
Solid State drives	2009	286	23.038	5.697	2.612	19.594	4.57	4.867	14.51

Appendix 3

See Tables 10 and 11.

Table 10 Patent feature descriptive statistics of ET samples

Features	Obs	Min	Max	Mean	SD
Backward citations	57	0.000	111.895	15.77604	16.491800
Number of non-patent citations	57	2.226	16.897	5.51818	2.339450
Number of claims	57	0.000	25.333	3.43393	4.734364
TCT	57	8.100	36.000	16.42658	5.299484
Forward citations	57	1.333	4.896	2.63484	0.804564
Family patent size	57	1.694	10.000	3.89932	1.343552
Average number of IPC	57	0.600	79.818	12.85679	12.783264

Table 11 Patent feature descriptive statistics of NET samples

Features	Obs	Min	Max	Mean	SD
Backward citations	48	0.000	88.500	10.00777	15.384141
Number of non-patent citations	48	2.073	9.054	5.34958	1.503196
Number of claims	48	0.167	10.468	2.89525	2.521401
TCT	48	7.364	23.292	15.80260	3.846595
Forward citations	48	1.000	14.000	3.16550	1.942798
Family patent size	48	1.625	17.000	4.22538	2.604960
Average number of IPC	46	2.864	79.556	13.56674	13.625225

Appendix 4

See Table 12.

Table 12 Pearson correlation analysis of patent features

Features	Backward citations	TCT	Number of non-patent citations	Number of claims	Family patent size	Number of IPC	Forward citations
Backward citations	1	0.184	0.057	0.350	0.081	0.219	0.478
TCT	0.184	1	0.008	0.113	0.139	0.105	0.179
Number of non-patent citations	0.057	0.008	1	0.403	0.010	−0.040	0.561
Number of claims	0.350	0.113	0.403	1	0.091	0.213	0.473
Family patent size	0.081	0.139	0.010	0.091	1	0.777	0.030
Number of IPC	0.219	0.105	−0.040	0.213	0.777	1	0.002
Forward citations	0.478	0.179	0.561	0.473	0.030	0.002	1

References

Barua, S., Islam, M. M., Yao, X., & Murase, K. (2014). MWMOTE–majority weighted minority oversampling technique for imbalanced data set learning. *IEEE Transactions on Knowledge and Data Engineering, 26*(2), 405–425.

Bengio, Y., & LeCun, Y. (2007). Scaling learning algorithms towards AI. *Large-Scale Kernel Machines, 34*(5), 1–41.

Bierly, P., & Chakrabarti, A. (1996). Determinants of technology cycle time in the US pharmaceutical industry'. *R&D Management, 26*(2), 115–126.

Breitzman, A., & Thomas, P. (2015a). Inventor team size as a predictor of the future citation impact of patents. *Scientometrics, 103*(2), 631–647.

Breitzman, A., & Thomas, P. (2015b). The emerging clusters model: A tool for identifying emerging technologies across multiple patent systems. *Research Policy, 44*(1), 195–205.

Chang, C. K., & Breitzman, A. (2009). Using patents prospectively to identify emerging, high-impact technological clusters. *Research Evaluation, 18*(5), 357–364.

Chang, P. L., Wu, C. C., & Leu, H. J. (2010). Using patent analyses to monitor the technological trends in an emerging field of technology: A case of carbon nanotube field emission display. *Scientometrics, 82*(1), 5–19.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: Synthetic minority oversampling technique. *Journal of Artificial Intelligence Research, 16*(1), 321–357.

Chiavetta, D., & Porter, A. (2013). Tech mining for innovation management. *Technology Analysis & Strategic Management, 25*(6), 617–618.

Choi, S., & Jun, S. (2014). Vacant technology forecasting using new Bayesian patent clustering. *Technology Analysis & Strategic Management, 26*(3), 241–251.

Cozzens, S., Gatchair, S., Kang, J., Kim, K. S., Lee, H. J., Ordóñez, G., et al. (2010). Emerging technologies: Quantitative identification and measurement. *Technology Analysis & Strategic Management, 22*(3), 361–376.

Daim, T. U., Rueda, G., Martin, H., & Gerdsri, P. (2006). Forecasting emerging technologies: Use of bibliometrics and patent analysis. *Technological Forecasting and Social Change, 73*(8), 981–1012.

Day, G. S., & Schoemaker, P. J. (2000). Avoiding the pitfalls of emerging technologies. *California Management Review, 42*(2), 8–33.

DeRouin, E., Brown, J., Beck, H., Fausett, L., & Schneider, M. (1991). Neural network training on unequally represented classes. New York: ASME Press.

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

Fu, L. D., & Aliferis, C. F. (2010). Using content-based and bibliometric features for machine learning models to predict citation counts in the biomedical literature. *Scientometrics, 85*(1), 257–270.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Cambridge: MIT Press.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. *International Conference on Neural Information Processing Systems, 3,* 2672–2680.

Hall, B. H., & Helmers, C. (2013). Innovation and diffusion of clean/green technology: Can patent commons help? *Journal of Environmental Economics and Management, 66*(1), 33–51.

Hall, B. H., Helmers, C., Rogers, M., & Sena, V. (2013). The importance (or not) of patents to UK firms. *Oxford Economic Papers, 65*(3), 603–629.

Harhoff, D., Scherer, F. M., & Vopel, K. (2003). Citations, family size, opposition and the value of patent rights. *Research Policy, 32*(8), 1343–1363.

Hassan, S. U., Imran, M., Iqbal, S., Aljohani, N. R., & Nawaz, R. (2018). Deep context of citations using machine-learning models in scholarly full-text articles. *Scientometrics, 117*(3), 1645–1662.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science, 313*(5786), 504–507.

Hwang, U., Choi, S., & Yoon, S. (2018). Disease prediction from electronic health records using generative adversarial networks.

Jun, S. P. (2012). An empirical study of users' hype cycle based on search traffic: The case study on hybrid cars. *Scientometrics, 91*(1), 81–99.

Jung, H., & Pedram, M. (2010). Supervised learning based power management for multicore processors. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 29*(9), 1395–1408.

Kayal, A. A., & Waters, R. C. (1999). An empirical evaluation of the technology cycle time indicator as a measure of the pace of technological progress in superconductor technology. *IEEE Transactions on Engineering Management, 46*(2), 127–131.

Kong, D., Zhou, Y., Liu, Y., & Xue, L. (2017). Using the data mining method to assess the innovation gap: A case of industrial robotics in a catching-up country. *Technological Forecasting and Social Change, 119,* 80–97.

Kreuchauff, F., & Korzinov, V. (2015). A patent search strategy based on machine learning for the emerging field of service robotics. *Scientometrics, 111*(2), 1–30.

Kyebambe, M. N., Cheng, G., Huang, Y., He, C., & Zhang, Z. (2017). Forecasting emerging technologies: A supervised learning approach through patent analysis. *Technological Forecasting and Social Change, 125,* 236–244.

Lanjouw, J. O., & Schankerman, M. (2004). Patent quality and research productivity: Measuring innovation with multiple indicators. *The Economic Journal, 114*(495), 441–465.

Lee, C., Kwon, O., Kim, M., & Kwon, D. (2018). Early identification of emerging technologies: A machine learning approach using multiple patent indicators. *Technological Forecasting and Social Change, 127,* 291–303.

Lee, S., Yoon, B., Lee, C., & Park, J. (2009). Business planning based on technological capabilities: Patent analysis for technology-driven roadmapping. *Technological Forecasting and Social Change, 76*(6), 769–786.

Lerner, J. (1994). The importance of patent scope: An empirical analysis. *The RAND Journal of Economics, 25*(2), 319–333.

Li, S., Hu, J., Cui, Y., & Hu, J. (2018). DeepPatent: Patent classification with convolutional neural networks and word embedding. *Scientometrics, 117*(2), 721–744.

Liu, Y., Zhou, Y., Liu, X., Dong, F., Wang, C., & Wang, Z. (2019). Wasserstein GAN-based small-sample augmentation for new-generation artificial intelligence: A case study of cancer-staging data in biology. *Engineering, 5*(1), 156–163.

Love, B. C. (2002). Comparing supervised and unsupervised category learning. *Psychonomic Bulletin & Review, 9*(4), 829.

Martin, B. R. (1995). Foresight in science and technology. *Technology Analysis & Strategic Management, 7*(2), 139–168.

Mcdaniel, P., Papernot, N., & Celik, Z. B. (2016). Machine learning in adversarial settings. *IEEE Security and Privacy, 14*(3), 68–72.

Natten, J. (2017). *Generative adversarial networks for improving face classification*. Master's thesis, Universitetet i Agder; University of Agder.

OuYang, K., & Weng, C. S. (2011). A new comprehensive patent analysis approach for new product design in mechanical engineering. *Technological Forecasting and Social Change, 78*(7), 1183–1199.

Pascual, S., Bonafonte, A., & Serrà, J. (2017). Segan: Speech enhancement generative adversarial network.

Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *Computer Science*.

Santana, E., & Hotz, G. (2016). Learning a driving simulator. arXiv preprint arXiv:1608.01230.

Sun, Y., Kamel, M. S., Wong, A. K. C., & Wang, Y. (2007). Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition, 40*(12), 3358–3378.

Trajtenberg, M. (1990). *Economic analysis of product innovation: the case of CT scanners* (Vol. 160). Cambridge, MA: Harvard University Press.

Valmadre, J., Bertinetto, L., Henriques, J., Vedaldi, A., & Torr, P. H. (2017). End-to-end representation learning for correlation filter based tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2805–2813).

Wang, K., Gou, C., Duan, Y., Lin, Y., Zheng, X., & Wang, F. Y. (2017). Generative adversarial networks: introduction and outlook. *IEEE/CAA Journal of Automatica Sinica, 4*(4), 588–598.

Zhang, Y., Lu, J., Liu, F., Liu, Q., Porter, A., Chen, H., et al. (2018). Does deep learning help topic extraction? A kernel k-means clustering method with word embedding. *Journal of Informetrics, 12*(4), 1099–1117.

Zhou, Y., Dong, F., Kong, D., & Liu, Y. (2019a). Unfolding the convergence process of scientific knowledge for the early identification of emerging technologies. *Technological Forecasting and Social Change, 144,* 205–220.

Zhou, Y., Lin, H., Liu, Y., & Ding, W. (2019b). A novel method to identify emerging technologies using a semi-supervised topic clustering model: A case of 3D printing industry. *Scientometrics, 120*(1), 167–185.

Zhu, X., Goldberg, A. B., Brachman, R., & Dietterich, T. (2006). Introduction to semi-supervised learning. *Semi-Supervised Learning, 3*(1), 130.

Zhu, X., Liu, Y., Li, J., Wan, T., & Qin, Z. (2018). Emotion classification with data augmentation using generative adversarial networks. In *Pacific-Asia conference on knowledge discovery and data mining* (pp. 349–360).

Zhuang, Y. T., Wu, F., Chen, C., & Pan, Y. H. (2017). Challenges and opportunities: From big data to knowledge in AI 2.0. *Frontiers of Information Technology & Electronic Engineering, 18*(1), 3–14.