Scientometrics (2021) 126:5431–5476
https://doi.org/10.1007/s11192-021-04001-1

Identifying emerging technologies to envision a future innovation ecosystem: A machine learning approach to patent data

Youngjae Choi[1] · Sanghyun Park[1] · Sungjoo Lee[1,2]

Received: 9 November 2019 / Accepted: 17 April 2021 / Published online: 16 May 2021

Abstract

With the rapid changes to technology as well as industry value chains, it has become essential for firms to identify emerging promising technologies that can better respond to external disruptive forces and be used to launch new businesses or improve current businesses. One of the most commonly used approaches in identifying emerging promising technologies is patent analysis. Patents have long been regarded as a useful source of data on technologies; accordingly, a number of previous studies have applied patents to define rising technologies. However, most previous studies have significantly relied on patent information in assessing promising technologies, whereas promisingness is determined by various other factors that are not explained by patent information. To overcome the limitation of previous approaches, this study proposes a hybrid approach considering both expert opinions and patent information to identify emerging promising technologies. For analysis, we firstly developed a set of criteria with which to evaluate potentially valuable patents, had experts evaluate only a portion of the patents from a larger patent portfolio of interest based on the criteria, and finally used the evaluation results to identify other potentially valuable patents from the rest of patents in the portfolio. Here, an active semi-supervised learning technique was applied, in which a small amount of labeled data (patents evaluated by experts) was used with a large amount of unlabeled data (the other patents from the portfolio). An analysis model consists of two layers—patents and patent attributes—with patent attributes such as technology characteristics used to classify patents into promising and unpromising ones. The proposed approach was applied to the automobile industry sector, and its usability was verified; the analysis results indicated that semi-supervised learning combined with active learning has potential in effectively searching for emerging promising technologies or filtering non-promising technologies with less human input. With only a small set of labeled patents, a large set of patents could be labeled, which saves time and effort when experts evaluate patents. Methodologically, this is an early attempt to introduce active semi-supervised learning in the context of patent analysis. Practically, the research findings enable expert opinions to be used effectively in identifying promising technologies and envisioning a future innovation ecosystem, making a balance between data- and expert-driven decision-making.

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (NRF-2019R1F1A1063032).

Extended author information available on the last page of the article

Keywords Emerging technologies · Innovation ecosystem · Patent · Semi-supervised learning · Active learning

Introduction

Current technology environments are characterized by the emergence of new and disruptive technologies as well as by converging technologies (Bröring et al., 2006; Geum et al., 2016; Momeni & Rost, 2016). Organizations need to cope with such rapidly changing technologies to sustain their business and gain further competitive advantages, which has made them take more futuristic perspectives in developing innovation strategies (Caviggioli, 2016; Lee et al., 2003; Veryzer, 2005). Indeed, identifying emerging technologies has become essential to envision a future innovation system and position a firm within that system.

Among various data sources, patents have long been regarded as a useful source of data on technologies due to their wide coverage in terms of technological areas, countries, and data accessibility (Ernst, 2003; Pilkington et al. 2009; Zhang, 2011). Accordingly, a number of previous studies have applied patent analysis to investigate technology trends, particularly to identify emerging technologies such as those that will have great impacts on subsequent technologies (Altuntas et al. 2015; Kim & Bae, 2017; Kyebambe et al. 2017; Verhoeven et al. 2016), those required to satisfy customer needs (Lee et al. 2008; Livotov, 2015), those likely to be licensed or transferred (Arora & Fosfuri, 2003; Bekkers et al. 2011; Giuri et al. 2013; Hsieh, 2013; Kang & Bekkers, 2015), those with distinguishing features (Fischer & Leidinger, 2014; Hall et al. 2005; Lee et al. 2012), and those with a mixture of previously mentioned features.

Despite their value, however, previous studies are subject to some limitations. First, although previous attempts have been made to automate the data-analysis process by minimizing human inputs, experts' insights are commonly required during the analysis to ensure high-quality results due to the inherent nature of patent analysis (Choi & Park, 2009; Daim et al. 2006; Mitchell, 1992). On one hand, uncertainties are inevitable when identifying emerging technologies that address future issues, which necessitate experts' insight to consider the complexity that will shape the future. On the other hand, domain knowledge is needed for effective data cleansing. For example, experts tend to play a significant role in selecting meaningful keywords in patent documents for further analysis. Despite the significance of experts' insights in identifying emerging technologies, few previous studies have focused on how to better collect and use such information during patent analysis. Second, while recent advances in data analytics have enabled technological trends to be more effectively investigated, only a few studies (e.g., Kyebambe et al. 2017; Lai et al. 2018) have made the best use of those advances in patent analysis so far. More attempts are needed to benefit from the development of analytical approaches.

To overcome the limitations of existing studies, this study proposes an approach with which to effectively incorporate experts' insights during patent analysis for identifying emerging promising technologies, using a machine learning approach—with both semi-supervised learning and active learning (i.e., active semi-supervised learning). In the proposed approach, firstly, experts evaluate a small set of patents from a larger patent portfolio of interest to classify promising ones and the others (labeled data), which are used to identify other potentially valuable patents from the rest of patents in the portfolio (unlabeled data). During this process, patent characteristics from multiple perspectives are investigated to assign labels to the unlabeled data; patents similar to those in the promising group will be classified as promising ones. Here, experts' inputs are interactively provided to obtain the desired outputs from the unlabeled data, particularly for the patents with low predictability. Finally, only the patents predicted as promising and recently published are analyzed so as to envision the future innovation system using co-classification analysis, through which key technologies areas along with the key organizations leading those areas can be identified from the prospective perspectives. Focusing on the recently published patents enables us to detect emerging technologies given the time required for those patents to be commercialized.

In developing the approach, however, a patent indicator, that is, the forward citation frequency, was used as a proxy for promisingness instead of experts' evaluations for the following two reasons. First, stable evaluation results are needed in developing a methodology and patent quality measures instead of experts' opinions are expected to provide more robust valuations of patents during the experiment. Second, among various dimensions, only a single dimension, technological impact on subsequent technologies, is chosen as a measure of patent quality in this study on the assumption that the potential users are likely to identify the patents with high technological impacts and also for the comparative analysis of performance with the existing study by Lee et al. (2018). Indeed patent quality can be measured from various dimensions including patent scope, paten family size, grant lag, backward citations, citations to non-patent literature, claims, forward citations, breakthrough inventions, generality index, originality index, radicalness index, and patent renewal (Squicciarini et al., 2013).

The proposed approach was applied to the automobile industry sector [specifically, to electric vehicle batteries (EVBs)], and its usability was verified. The analysis results showed high prediction accuracy, and the experts' interaction during the process (i.e., in our experiments, telling weather a patent is promising or not based on its forward citation frequency for the patents classified neither as promising nor non-promising) significantly increased the prediction accuracy. Accordingly, with only a small set of labeled patents, a large set of patents could be labeled, which can save time and effort when evaluating patents. Methodologically, this is an early attempt to introduce semi-supervised learning in the context of patent analysis. Practically, the research findings enable us to effectively use expert opinions in identifying emerging promising technologies with which to envision a future innovation system, making a balance between data- and expert-driven decision-making.

The rest of this paper consists of four sections. In "Background" Section, the existing approaches to identifying emerging technologies from patents are reviewed, with more recent approaches investigated in greater detail, and the methodological background is also explained. Section 3 describes the proposed approach in terms of research ideas and process, based on which the analysis results are shown in "Analysis results" Section. Finally, after the discussion on the analysis results in "Discussion" Section, the contributions along with this study's limitations are summarized to offer future research directions in Conclusions" Section.

Table 1 Comparisons of the traditional and recent approaches to patent analysis

	Traditional approach	Recent approach
Perspective	Retrospective	Prospective
Data	Ex-post	Ex-ante
Analysis techniques	Unsupervised	Supervised

Background

Data-mining techniques applied to patents

Patent analysis has long been used to identify emerging technologies, and various analysis techniques have been applied to patent data. Table 1 summarizes the differences between the traditional and recent approaches to patent analysis.

In traditional approaches, the focus has been on analyzing past technological trends by taking a *retrospective perspective*. The patents published in the last 10–20 years were examined to produce a snapshot of the technological landscape through static analysis or to identify the evolutionary path and dynamics of technology evolution through time-series analysis. The analysis method has been advanced predominantly by using *ex-ante patent information* (the information produced after patent application), such as patent citation or renewal information. Furthermore, data-mining techniques have been used to explore patent data and ultimately understand the technological characteristics revealed in the data.

Accordingly, *unsupervised machine learning approaches* have frequently been adopted, particularly to reduce the dimensions of patent data and map them into a two-dimensional space for visualization. These needs become stronger when the description parts of patent documents are analyzed because text data are usually transformed into numeric vectors so that the frequency of a word appearing in a document can be measured. For example, Yoon et al. (2002) adopted a self-organization feature map to position similar patents with respect to technological content onto two-dimensional space. Similarly, Lee et al. (2005) used principal component analysis to visualize technology vacuums—the areas with relatively low attention—and proposed those vacuums as candidates for new technology opportunities. Of course, some studies have applied other data-mining techniques to bibliographic patent data to investigate the relationships between patents or technological areas. For example, Kim et al. (2011) employed association rule mining to examine technological areas (i.e., international patent classification) that frequently appear together in patents to understand the drivers of technology convergence. Choi et al. (2014) and Jun et al. (2012) employed clustering analysis to group patents with similar characteristics. Indeed, these data-mining techniques have contributed significantly to the advances in patent analysis.

Furthermore, the recent progress in data analytics, particularly the emergence of machine learning approaches, has opened up a new possibility for patent analysis by taking a *prospective perspective*; emerging approaches have emphasized the possibilities of predicting the future statuses of technologies by adopting supervised machine learning approaches. In these studies, the target technologies to be identified are twofold, (1) those with great impacts on subsequent technologies, named hot patents, and (2) those converging with other technologies, named converging technologies.

First, one of the earliest attempts to predict promising technologies is the work by Lee et al. (2011), who applied *supervised machine learning approaches* to identify patents expected to have high forward citations. They used a decision tree model and conducted a performance test on 50 technologies in Gartner's hype cycle for emerging technologies from 2009 to 2010, for which they obtained an accuracy of 84%. The invention characteristics, as a part of the *ex-ante patent information* (the information that was obtainable at the point of patent application), were used as predictors. Later, Lee et al. (2018) also carried out a machine learning approach, but unlike the previous work, which extracted only hot patents, they classified patents into four groups depending on the degree of technological impacts and assigned patents to one of those four groups through prediction. They utilized a feed-forward neural network, random forest, and a support vector machine model to identify emerging technologies. The highest accuracy they obtained was 91%.

Second, recent studies on predicting technology convergence have started to propose a new approach with which to predict technology convergence, addressing the need for projecting the future status of technology convergence rather than monitoring its past trends. Kim & Lee (2016) applied neural network analysis to construct a future patent citation matrix, after arguing that neural network analysis has shown better performance than curve-fitting; the projected matrix was used to identify technologies that will converge in the future. Yoon and Magee (2018) were the first to introduce link-prediction analysis to patent analysis, by which a support vector machine was used to predict technologies that are likely to be linked to each other in the near future. Park and Yoon (2018) applied a link-prediction method to predict potential technological knowledge flows based on patent citations between biotechnologies and information technologies, and they suggested technological opportunities for convergence.

This study is in line with the recent research on patent analysis but was an attempt to overcome the existing studies' limitations through approach to effectively incorporate experts' insights into patent analysis with support from up-to-date data-analysis approaches.

Machine learning approaches

Semi-supervised learning

The conventional supervised learning technique needs a large scale of labeled data; if the data set is not labeled, then people should provide the labels, which may result in huge time usage and costs as well as the need for experts with domain knowledge. On the other hand, unlabeled data are relatively easy to obtain, compared to labeled data. A semi-supervised learning technique builds a classifier using some labeled data with a large amount of unlabeled data. In most semi-supervised learning processes, a classifier is developed using labeled data, based on which unlabeled data are gradually labeled; both labeled and unlabeled data are trained at the same time to develop an effective classifier. Various techniques have been proposed for semi-supervised learning, such as self-training (Li & Zhou, 2005; Rosenberg et al., 2005; Tanha et al., 2017; Triguero et al., 2014; Yarowsky, 1995), co-training (Chen & Deng, 2015; Tanha et al., 2011), and graph-based semi-supervised learning (Belkin et al., 2006; Zhu et al., 2005).

In particular, self-training, which was adopted in this study, is one of the simplest and most commonly used techniques; accordingly, it has been applied to a wide range of areas, including natural language processing (Ando & Zhang, 2005; McClosky et al., 2006; Yarowsky, 1995), object detection (Rosenberg et al., 2005), and document classification (Maulik & Chakraborty, 2011). First, a small set of labeled data is trained to develop a classifier, based on which a large set of unlabeled data are classified to assign labels (predicted values). Then, among the unlabeled data, those with reliable prediction results are added to the training set without human intervention to construct a new classifier with a new set of data. This process is repeated until the test set reaches a satisfying (the required) quality. In the self-training, it is assumed that the prediction results from using the initial labeled data set to classify unlabeled data are the most reliable (Triguero et al., 2015). However, as the labeled data for the initial training are sparse, it is sometimes difficult to obtain the target reliability. Furthermore, the unlabeled data patterns are gradually considered during the training processes, which may decrease classifier performance. To overcome this limitation, active learning was combined with self-training, which helped us to construct a more accurate data set for training.

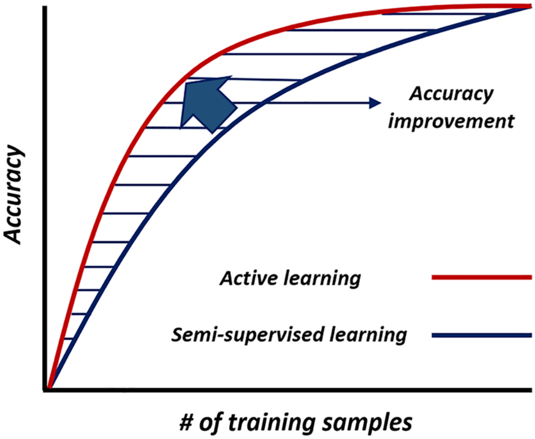
This self-training can be implemented based on several machine learning algorithms, among which three classifiers—random forest (RF), support vector machine (SVM), and extreme gradient boosting (XGBoost)—were adopted in this study. First, RF is one of the most well-known ensemble methods. It is an extended decision-tree method; instead of relying on a single tree, it aggregates multiple trees, with multiple predictions being combined to improve the classifier's performance (Liaw & Wiener, 2002). A tree that consists of RF is built based on randomly selected training data along with input variables. While the accuracy of an individual decision tree may not be high enough, the aggregated prediction results of multiple decision trees (i.e., a forest) become reliable and robust. Second, SVM creates a hyperplane that separates the data into classes, for which the vectors defining the hyperplane are called support vectors. The optimal hyperplane maximizes the margin between classes (i.e., the greatest distance from the data) (Suykens et al., 1999). By integrating the kernel-mapping concept and optimization techniques into the principles of statistical learning, predicted classification results can be returned according to the coordinate positions for the input vector values once the hyperplane is found. Finally, XGBoost, known to be fast and extendable, has gained enormous popularity and attention and has been frequently chosen by winners of data analysis competitions (Liu et al., 2016). It is also a type of ensemble method, in which a gradient boosting technique is applied to a tree model in order to progressively boost the weak learning models through iterative learning, to build a more powerful classifier (Chen & Guestrin, 2016). More specifically, it builds an optimized model by controlling the tree complexity in order to prevent overfitting while minimizing training loss.

Active learning

Unlike traditional semi-supervised learning, in which predicted labels are given to unlabeled data without human intervention, active learning allows such intervention; humans (i.e., experts or analysts) label some of the unlabeled data chosen by algorithms (Crawford et al., 2013; Tuia et al., 2009, 2011). The aim is to construct a small but rich newly labeled data set by assigning an accurate label to the candidate data for labeling. Thus, active learning can increase the model's performance and also accelerate the speed of convergence (Leng et al., 2013). At the same time, the performance of active learning may not increase without accurate labeling by humans.

Nevertheless, we recognized the potential of active learning in identifying promising technologies and applied it to this study. That is, a model is built by training the initial small set of labeled data; then, the model is used to assign reliable labels to unlabeled data. Here, humans assign accurate labels to identify and correct relatively less reliable data, which significantly increases the model's performance compared to simple

Fig. 1 Comparison of learning curves for active learning and semi-supervised learning



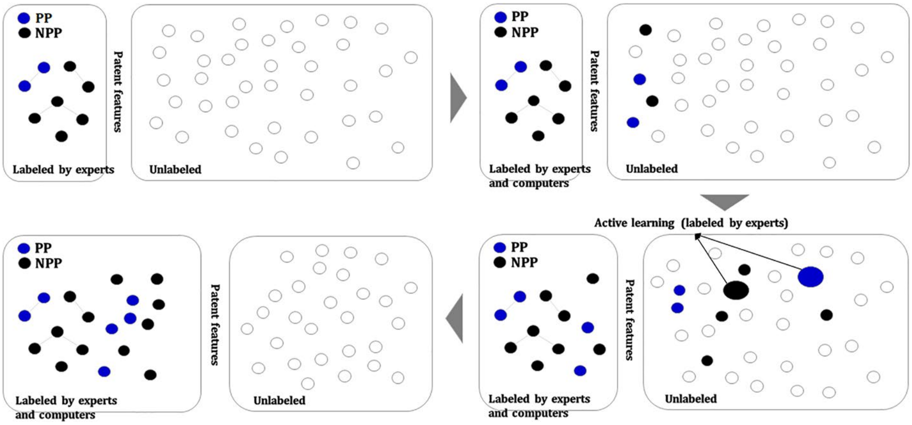

PP
NPP
Patent features
Labeled by experts
Unlabeled
Labeled by experts and computers
Active learning (labeled by experts)

Fig. 2 Basic research ideas

semi-supervised learning. This process is repeated to expand the data, and the expanded range of data prevents overfitting (Cohn et al., 1999; Riccardi et al., 2005). Accordingly, a model performing at least the same or better was expected by adding accurate labels to some of the unlabeled data that play a significant role in determining the class boundaries—that is, those with ambiguous probability values (e.g., with values between 0.4 and 0.6 in the case of binomial classification) (see Fig. 1).

Research framework

Research ideas

Figure 2 describes the basic ideas of the proposed approach used to identify emerging promising technologies from patent data. The term, promising technologies, may indicate different things in different contexts; moreover, distinguishing such technologies from the others using an enormous amount of patent data may not be easy and may rely solely on experts. Hence, in order to identify promising technologies using patent data, we firstly defined technologies as a set of relevant patents, and then try to seek for promising patents to be grouped together based on similar terms and concepts. More specifically, promising patents (PPs) are distinguished from non-promising patents (NPPs), for which a machine learning–based classifier with a binary target variable (1 for PP vs. 0 for NPP) and a set of patent characteristics used as input variables are proposed. This model is developed based on an initial small set of data evaluated by experts, which is then used to assign labels to the patents along with their predicted values for the dependent variable being above or below a cut-off value (above for PP, below for NPP). Then, among the unlabeled patents, those with most uncertain values with which to be classified as ET or NET (i.e., those having predicted values for the dependent variable of around 0.5) are labeled by experts, to convert the most unreliable data into reliable data. The patents that were newly labeled during the previous two stages are added to the training set until no unlabeled data are available. Here, it should be noted again that a forward citation frequency, as a proxy for promisingness, was used in the experiment instead of experts' evaluation of the patents, to ensure robustness of the experimental results. In a real case study, various other dimensions can be considered and the target year (e.g., promisingness in the next five years) should be given to obtain reliable results from experts. If the proposed approach is applied to the latest patents, the emerging promising patents can be identified to be grouped into several emerging promising technologies.

Overall research process

The overall research process used to implement the research idea consists of four steps (see Fig. 3). The first step focuses on data collection and pre-processing, during which patent data in the field of interest are collected, the variables for the analysis are defined, and data pre-processing is implemented, such as feature scaling and feature selection. In the next step, six models combining two approaches—semi-supervised learning and active learning—and three classifiers—RF, SVM, and XGBoost—are designed. The third step involves evaluating and comparing the six models with respect to three indicators: accuracy, F1-score and AUROC. In the final step, the best performing model is adopted to identify promising technologies using recent models, based on which the key technological themes (innovation fields) along with the key organizations leading those areas (innovation actors) are investigated. The analysis results will be presented in the form of multi-layered network that can describe the features of the future innovation system. The detailed procedures will be explained in the following section.

Detailed procedures

Data collection and preprocessing

It is essential to select an appropriate database for research purposes in order to obtain reliable results. The most frequently adopted databases include those published from the United States Patent and Trademark Office (USPTO) database, European Patent Office (EPO), and Japanese Patent Office (JPO) (Kim & Lee, 2015). Among these databases, the USPTO database contains the largest amount of information and is easy to access; thus, it is regarded as the most represented patent database in innovation studies (Archibugi &

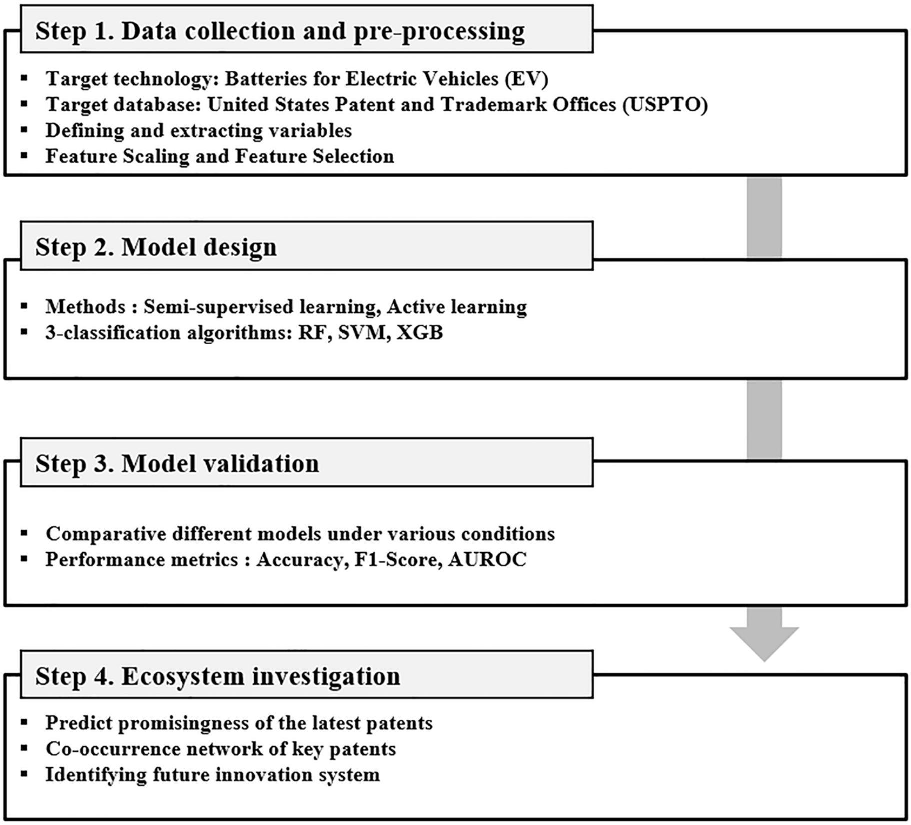

Step 1. Data collection and pre-processing
▪ Target technology: Batteries for Electric Vehicles (EV)
▪ Target database: United States Patent and Trademark Offices (USPTO)
▪ Defining and extracting variables
▪ Feature Scaling and Feature Selection
Step 2. Model design
▪ Methods : Semi-supervised learning, Active learning
▪ 3-classification algorithms: RF, SVM, XGB
Step 3. Model validation
▪ Comparative different models under various conditions
▪ Performance metrics : Accuracy, F1-Score, AUROC
Step 4. Ecosystem investigation
▪ Predict promisingness of the latest patents
▪ Co-occurrence network of key patents
▪ Identifying future innovation system

Fig. 3 Overall research process

Planta, 1996). Accordingly, this paper also involved collecting patents from the USPTO database.

Then, the patents need to be preprocessed for analysis, whereby the values of the input variables and a target variable are produced. Regarding the target variable, promisingness can be measured from various perspectives; generally, market and/or technology potential have been considered in evaluating promisingness (Park et al. 2016; Song et al. 2018). Given that a patent with significant market and technology potential would have great impacts on subsequent technologies, resulting in a high frequency of patent forward citations, the patent forward citation frequency was adopted as a proxy for patent promisingness in this study. The top 10% of patents as promising patents based on their total citation frequencies where the comparison was made among the patents published in the same year; patent citations are sensitive to the year in which the patent was granted since they tend to increase over time.

On the other hand, the input variables used to identify promising technologies concern four types of patent features: invention, assignee, class, and keyword features. First, invention features indicate the characteristics of the patent itself (i.e., ex-ante patent information) and are thus obtained from the target patent's bibliographic information. These include the number of claims, inventors, classes, and so on and have long been used to evaluate patent

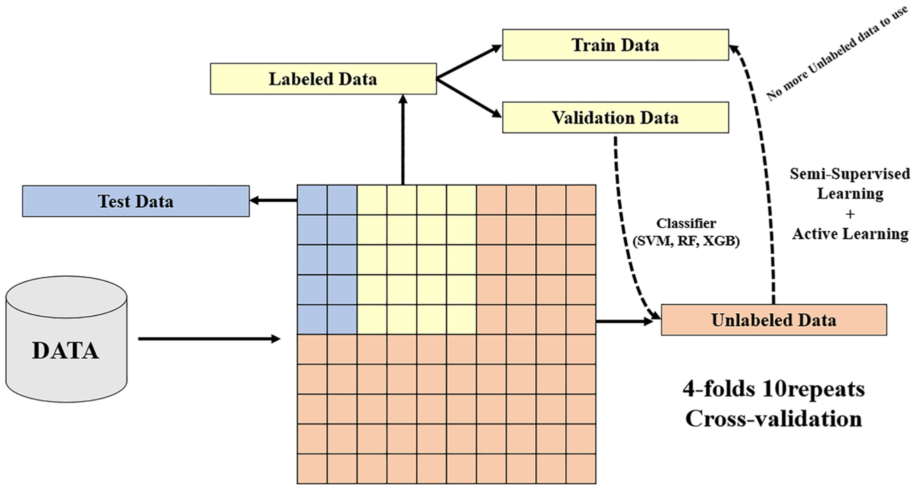


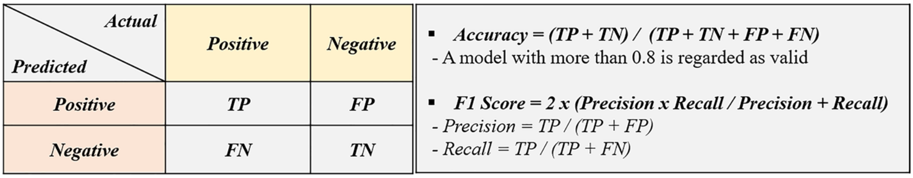
Fig. 4 Process of designing semi-supervised learning and active learning model

potential. Second, assignee features correspond to the representative assignee's characteristics, indicating the excellence of the patent owners, on the assumption that patents published by better assignees are more likely to be promising than the others. Third, class features are related to the characteristics of the technological fields to which the target patent belongs, given that such technological fields are likely to affect whether the patent will be highlighted in the future or not. Finally, keywords features are similar to the class features, in that they both are aimed at examining the characteristics of technological fields. However, unlike the class features, the keyword features collect information from the keywords in the patent titles. In total, 34 input variables were designed, and the patent data were preprocessed to be suitable for obtaining the values of these variables.

Model design

The model designed in this study is described in Fig. 4. We first set aside 10% of the data to be used as a test set for model validation. Then, the remaining data were divided into two categories—labeled data and unlabeled data, according to the labeling proportion—and the labels for the patents in the unlabeled data set were removed for further analysis. Third, the labeled data were again partitioned into a training set and a validation set to train the model. Here, three classification algorithms—RF, SVM, and XGBoost—were applied, and the 10-repeat 4-folds cross-validation was used to fit the model.[1] The next step was to predict the labels for one batch of the unlabeled data, which correspond to 200 units, based on the initial trained model. Taking into account the distribution of target variable values, only some of the unlabeled data (i.e., those with estimated probability values greater than

[1] K-fold cross-validation is a statistical technique used to prevent model overfitting by splitting the training data set into K sub-sets during training. The training data set is first partitioned into K clusters of the same size. K iterations of training and valuation are performed, with one fold used for validation and the other K − 1 folds used for training. Finally, the training results are put together to accurately estimate the model's performance.



Actual *Predicted*	***Positive***	***Negative***
Positive	***TP***	***FP***
Negative	***FN***	***TN***

- ***Accuracy = (TP + TN) / (TP + TN + FP + FN)***

- A model with more than 0.8 is regarded as valid

- ***F1 Score = 2 x (Precision x Recall / Precision + Recall)***

- Precision = TP / (TP + FP)
- Recall = TP / (TP + FN)

Fig. 5 Example of a confusion matrix

a cut-off value) were labeled and added, to become a new labeled data set. From this new data set, a new training set was formed and used to fit the model. The process was iterated until no unlabeled data were available. Also, the model's performance was evaluated using the test set for each iteration. We used R packages for this analysis, specifically Caret Package to design a prediction model, and Ranger Package for the RF, Xgboost Package for XGBoost, and the e1071 Package for SVM, to implement the classification algorithms.

The active semi-supervised learning process was quite similar to the process described above. However, unlike in traditional semi-supervised learning, analysts labeled the unlabeled data that failed to obtain a label based on the model and were given a less confident estimation. That is, in addition to the data newly labeled by a model, the unlabeled data were also used for the training in the following step. In this study, the unlabeled data with estimated probabilities ranging from 0.4 to 0.6, being regarded as the least confident, were labeled by the analysts.

Model validation

Two criteria were used to evaluate the model's performance. The first is a confusion matrix that compares the actual and predicted values and produces the accuracy and F1 scores (see Fig. 5). Accuracy, as one of the most commonly used performance measures, is measured as the number of correct predictions divided by the number of all data points. A model is regarded as valid when its value is greater than 0.8. However, for an imbalanced dataset, the accuracy measure may produce biased results (Kim et al., 2017), as is the case in this study; only 10% of the total data were promising patents, whereas the remaining 90% were non-promising patents. Accordingly, we also introduced the F1 score, obtained from the harmonic average of precision (the percentage of true positives from all predicted positives) and recall (the percentage of predicted positives from all true positives), to evaluate the model's performance for such an imbalanced dataset.

The second criterion is based on the area under receiver operating characteristic curve (AUROC). It is a common performance metric that can be used to evaluate classification models and shows sensitivity against 1-specificity. The AUROC is between 0 and 1; as the AUROC becomes further away from 0.5, the model's performance is evaluated to be better (perfect accuracy for AUROC $=1$, high accuracy for AUROC ≥ 0.9, moderate accuracy for $0.9>$ AUROC ≥ 0.7, low accuracy for $0.7>$ AUROC ≥ 0.5, and null model for $0.5>$ AUROC) (Fischer et al., 2003). Hence, we focused on the F1 score and the AUROC in evaluating the model's performance.

Ecosystem development

The ultimate goal of the proposed approach is to envision the innovation system from the prospective perspectives by establishing the main themes and key actors of emerging

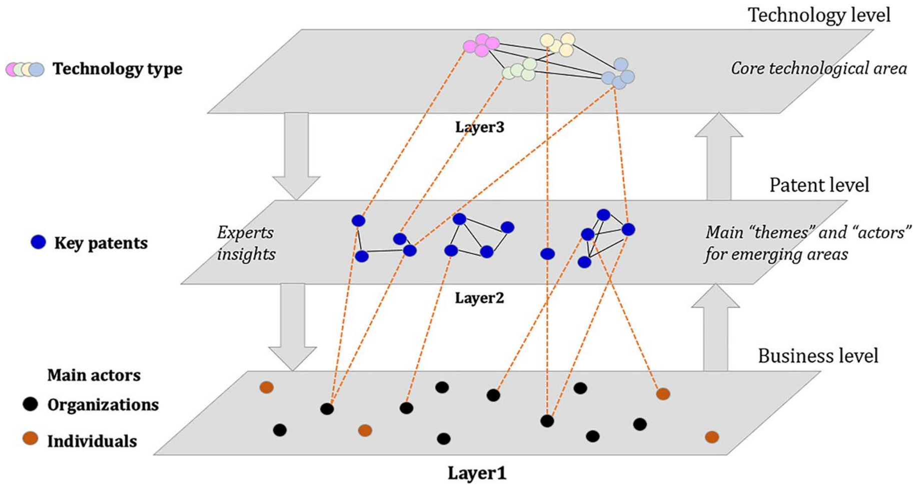


Fig. 6 A multi-layered network innovation system

promising technologies. As the main themes and key actors of emerging promising technologies are based on a set of patent recently published and are expected to be promising in the future, the performance of this innovation system mapping will depend on the performance of the identification of promising patents.

The innovation system takes the form of a multi-layered network consisting of three layers (see Fig. 6). The first layer corresponds to the business level and is aimed at identifying the key actors in this innovation system. Thus, the holders of the promising patents identified by the proposed approach are listed, and their relationships are visualized based on the similarity of their patents, in terms of corporative patent classification (CPC) codes at the sub-class level. Those possessing more promising patents are expected to be more active in creating new technologies. The second layer is for the patent level, where the promising patents are positioned to link the first and third layers. Thus, this layer could be a hidden layer. Finally, the third layer is for the technology level and presents the key technology themes in the future innovation systems. The list of CPCs at the sub-class level is constructed and their relationships are established based on the CPC co-occurrence analysis using the promising patents. If two CPCs frequently appear in the promising patents, they are regarded as closely related to each other. Accordingly, we can obtain a network with CPCs as the nodes and the degree of relationships as the edges. Here, the degree of centrality is calculated to select the top 10 key CPCs for visualization.

Analysis results

Data collection and preprocessing

The target area was set as EVBs. Using the Wisdomain website for patent search (https://www.wisdomain.com), relevant patents published between 1st January 1976 and 31st July 2019 were found. The following search terms were selected and applied to titles, abstracts and claims, during which the patent classification information was added to avoid noisy patents:

> TAF = ((electr* W/1 automot*) OR (electr* W/1 automobil*) OR (electr* W/1 vehicle*) OR (electr* W/1 car*) OR (hybrid* W/1 automot*) OR (hybrid* W/1 automobil*) OR (hybrid* W/1 vehicle*) OR (hybrid* W/1 car*)) AND (battery) AND (IC = (B60D* OR B60K* OR B60L* OR B60R* OR B62D* OR B62M* OR C25* OR G06F* OR G06K* OR H01H* OR H01L* OR H01M* OR H01R* OR H02J* OR H02K* OR H02M* OR H02P* OR H04L*))

We also used Web crawler to collect patent information (i.e., the number of drawings and the citations of non-patents) from Google Patents that were not available from the Widsdomain website. As a result, 6,287 total patents were collected.

Then, the data were preprocessed for further analysis. As citations accumulate over time (Breitzman & Thomas, 2015; Yoon & Park, 2004), the citation data may not be available for young patents, while the data may be biased for old patents. Thus, to ensure the reliability of the analysis, we firstly deleted relatively old or young patents, and the period of analysis was limited from 1990 to 2013. Then, the values for the target variable and input variables were obtained.

Table 2 summarizes their operational definitions. The promising patent (target variable) was defined as a patent with total forward citation number in the top 10%, based on the forward citation distribution for the patents published in the same year, following the work by Kaplan & Vakilia (2015). However, unlike their work, in which only the top 5% were considered as breakthrough patents, we expanded the scope of promising technologies to 10% because non-breakthrough patents can also be promising. On the other hand, the patent features (input variables) included invention features (patent level) along with technology, assignee, and keywords (aggregate level) features. For feature selection and scaling, the highly skewed variables were log-transformed, and the variables with missing values were removed from the dataset. Consequently, the scopes of inventor countries and of applicant countries were deleted, and finally the 32 input variables and 1 target variable for 2,698 patents remained. More detailed information for basic descriptive dataset and pre-processing is summarized in "Appendices" 1 and 2.

Model design

Semi-supervised learning

How analysts design a model has a significant impact on the mode performance for the three machine learning techniques of semi-supervised learning. Thus, a proper tuning process for setting appropriate hyperparameters is required. In this study, a grid search technique was adopted, as shown in Table 3; for RF and XGBoost, custom grid search was applied, for which users set the hyperparameters, while for SVM, the optimal hyper-parameters were automatically determined by cross-validation for different tuneLength values. In addition, of the 2698 patents, only 276 (10.2%) patents corresponded to promising patents, indicating the imbalanced distribution of target variable; thus, random oversampling examples (ROSE) was applied to solve this problem. ROSE, proposed by Menardi and Torelli (2014), is a technique for synthetic generation of training data without duplication and has shown better performance than other sampling techniques.

Table 2 Summary of the input variables and target variable

Type	Category	No	Variables	Operational definition (references)
Input variables	Patent level: invention features	1	Scope of patent rights	No. of independent claims (Tong et al., 1994; Su et al., 2012)
		2	Degree of novelty	No. of backward citations by other patents published in the USPTO (Trajtenberg, 1990; Harhoff et al., 2003; Hirschey et al., 2001)
		3	Scope of inventors	No. of inventors (Nelson, 1961; Su et al., 2012)
		4	Scope of inventor's country	No. of inventor countries (Su et al., 2012)
		5	Scope of applicants	No. of patent applicants (Reitzig, 2004; Guellec et al., 2000)
		6	Scope of applicants' country	No. of countries for patent applicants (Su et al., 2012)
		7–8	Scope of technology	No. of IPCs assigned to the patent (Lerner, 1994; Scotchmer, 1991; Su et al., 2012) No. of CPCs assigned to the patent
		9–10	Diversity of technology	No. of IPCs with the same four digits (same subclasses) No. of CPCs with the same four digits
		11	Scope of patent market	No. of family patents (Harhoff et al., 2003; Putnam, 1996)
		12	Technological insight	No. of non-patent references (Su et al., 2012)

Table 2 (continued)

Type	Category	No	Variables	Operational definition (references)
		13	Scope of patent images	No. of patent images
	Aggregate level: technology characteristics	14	Technology recency	No. of top 10 CPCs in terms of patent numbers within the recent three years among those assigned to the patent
		15–16	Technology quality	Average no. of forward citations for all patents belonging to the CPCs assigned to the patent Average no. of backward citations for all patents belonging to the CPCs assigned to the patent
		17	Technology coverage	Average no. of patent families for all patents belonging to the CPCs assigned to the patent
		18	Recombinant novelty	Indicator of the impact of the new patent sub-class combination (Fleming et al., 2007)
		19	Component familiarity	How often the patent classification code has been used until recently (Fleming, 2001)
	Aggregate level: assignee characteristics	20	Assignee recency	Applicant growth rate over the most recent five years compared to the next five years
		21–22	Assignee quality	Average no. of forward citations for all patents published by the assignee (s) Average no. of backward citations for all patents published by the assignee (s)
		23	Assignee coverage	Average no. of family patents for all patents published by the assignee (s)
	Aggregate level: keyword characteristics	24	Nouns in title	No. of nouns within the patent's title
		25–34	10 distinguishable words	Frequency of the top 10 words within the patent's title
Target variable			Promisingness	Whether the patent is promising or not – promising for top 10% of patent citations (1 for ET or 0 for NET)

Table 3 Assigned hyperparameter by classifier

Classifier	Hyperparameter	Options
SVM	Kernel	"Radial"
	Cost	Optimal parameter value found by random grid search
	Gamma	Optimal parameter value found by random grid search
RF	Mtry	10, 15, 20
	Splitrule	"extratrees", "gini"
	Min.node.size	1, 2, 3
XGB	Objective	"binary:logistic"
	Eval_metric	"Error"
	Nrounds	50, 100, 150, 200, 250, 300, 350
	Eta	0.3, 0.4
	Gamma	0
	Max_depth	1, 2, 3, 4, 5, 6, 7
	Colsample_bytree	0.6, 0.8
	Min_child_weight	1
	Subsample	0.5, 0.583, 0.667, 0.75, 0.833, 0.917, 1

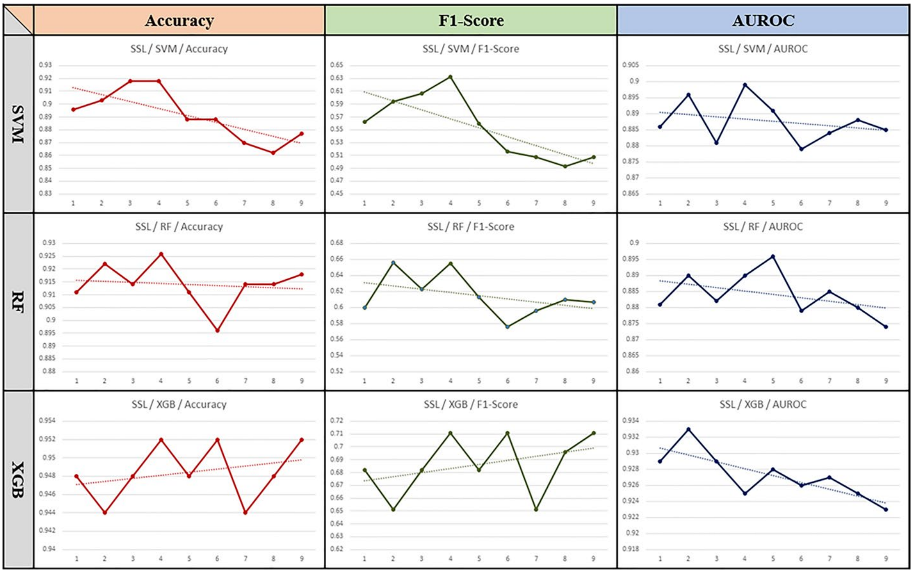


Fig. 7 Semi-supervised learning results during iteration

Figure 7 presents the semi-supervised learning results, where the accuracy, F1-score, and AUROC values for three classifiers—SVM, RF, and XGBoost—were produced over iterations. The accuracy measure is not reliable for an imbalanced dataset, so only the other two measures were considered in the performance evaluation. As to the F1 score (target value: ET), the performance increased for XGBoost but decreased for SVM and RF over iterations. However, for the final performance after the ninth iteration, both the RF and

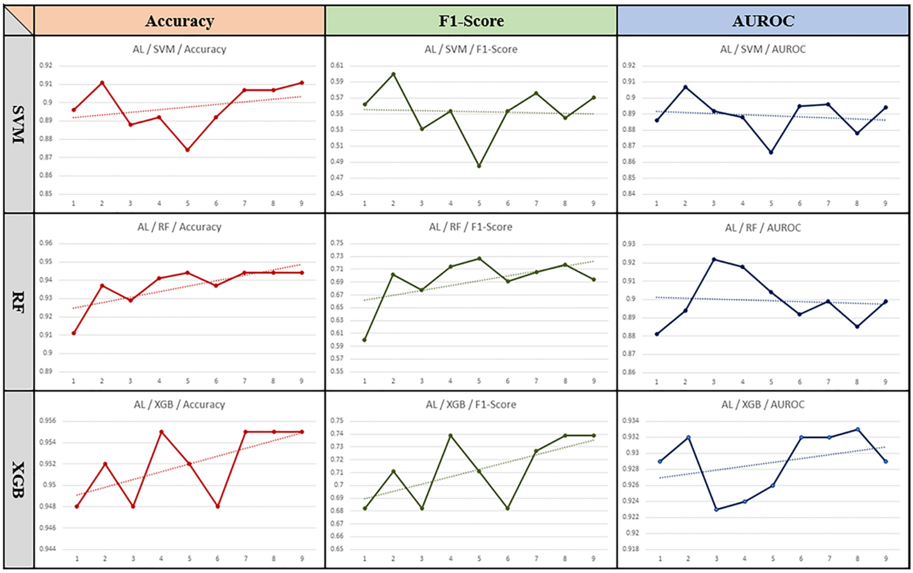


Fig. 8 Active learning results during iteration

XGB classifiers produced better performance than the initial model (Iteration 0), by 0.7% for RF and 2.9% for SVM. On the contrary, the performance of XGBoost was 5.5% lower. With regard to the AUROC, the performance after the ninth iteration decreased for all classifiers, as compared to the initial model (Iteration 0)—by 0.1% for SVM, 0.7% for RF, and 0.6% for XGBoost.

To summarize, no meaningful differences were found between supervised learning and unsupervised learning, possibly due to the limitations of self-training and the characteristics of the patent data. Self-training can be significantly affected by outliers, and having outliers in a training set will decrease the model's performance. Patent data are likely to have a number of outliers, given that the promisingness can be affected by various features other than the patent features. There may be two solutions: to stop the iterations when the best performance is reached or to consider active learning to improve the performance.

Active learning

For active semi-supervised learning, the true label was given to the patents with a probability between 0.4 and 0.6 of being promising, as the classifiers will fail to give labels to those patents. Figure 8 presents the changes to the accuracy, F1 score, and AUROC values over iterations. Again, interpretations were made only based on F1 score (target value: ET) and AUROC. First, regarding the F1 score, the performance tended to increase by iteration for RF and XGB. Furthermore, the final performance (Iteration 9) for all three classifiers was higher than that of the initial model (Iteration 0), leading to a 0.9% increase for SVM, 9.4% for RF, and 5.7% for XGB. Second, as for the AUROC, the performance of XGB classifier tended to increase over iterations, and the final performance increased for all three classifiers as well. The results indicated that the final performance (Iteration 8) was increased by 0.8% for SVM and 1.8% for RF, as compared to the initial model (Iteration 0). It stayed the

Table 4 Performance comparisons of all models

Classifier			Accuracy	F1 Score		AUROC
				Positive: PP	Positive: NPP	
a	Supervised learning	SVM	0.896	0.562	0.941	0.886
		RF	0.911	0.600	0.950	0.881
		XGB	0.948	0.682	0.972	0.929
b	Semi-supervised learning	SVM	0.877	0.507	0.930	0.885
		RF	0.918	0.607	0.954	0.874
		XGB	0.952	0.711	0.974	0.923
c	Active learning	SVM	0.911	0.571	0.950	0.894
		RF	0.944	0.694	0.969	0.899
		XGB	0.955	0.739	0.976	0.929

same for XGBoost. We can conclude that active learning contributed to overcoming the limitations of self-training, to some extent, by giving labels to uncertain data.

Model validation

Table 4 summarizes the final performance at the end of learning for three cases: (a) supervised learning; (b) semi-supervised learning (after the ninth iteration); and (c) active learning (after the ninth iterations). The performance was evaluated based on three measures—accuracy, F1 score, and AUROC, while the values for two additional measures, precision and recall, are provided in "Appendix" 3. Here, we calculated two types of F1 scores: one with PP as its target value and the other with NPP as its target value (see "Appendix" 4 for more details). If the PP is set to a target value, then the goal is to identify PPs. If the NPP is set to a target value, then the analysis is aimed at identifying NPPs, which will be greatly useful for screening purposes. The table shows that when semi-supervised learning is applied, the performance stays the same or sometimes even slightly decreases, as compared to the supervised learning. On the other hand, active learning seems to have superior performance to the others. As patent data are characterized by a large number of outliers compared to other types of data, and thus the performance increase seems to be small or even is not observed when a semi-supervised learning using only a small number of data is applied. Accordingly, we argue that active learning can be applied to improve the performance (see "Appendix" 5 for more details).

First, for F1 scores with a target PP value, active learning's performance was greater than that of semi-supervised learning for all classifiers, resulting in 0.571 for SVM, 0.694 for RF, and 0.739 for XGB. The performance difference between supervised learning and active learning varies by classifier, corresponding to 0.9% for SVM, 9.4% for RF, and 5.7% for XGBoost. The latest algorithm, XGBoost, showed good basic performance and the largest performance increase when active learning was adopted. Additionally, when a target value was set to NPP, we obtained quite high F1 scores—0.950 for SVM, 0.969 for RF, and 0.976 for XGBoost—signifying that it is relatively easy to filter out NPPs. Thus, if the focus of the analysis is to reduce the number of patents to further review in identifying promising technologies, then the target value should be set to NPP. On the other hand, if it is to identify core promising technologies, then setting the target value to PP could

Table 5 List of applicants for promising patents

No	Applicant	Frequency	Type	Country
1	Emerging automotive LLC	17	Organizations	US
2	Chargepoint Inc	5	Organizations	US
3	Fallbrook intellectual property co LLC	3	Organizations	US
4	Qualcomm Inc	3	Organizations	US
5	Future motion Inc	2	Organizations	US
6	Scheucher; Karl f	2	Individuals	US
7	Wong; Alexander	2	Individuals	US
8	Zeco systems Pte Ltd	2	Organizations	SG
9	Aerovironment Inc	1	Organizations	US
10	Allison transmission Inc	1	Organizations	US
11	Bosch automotive service solutions LLC	1	Organizations	US
12	Civilized cycles Inc	1	Organizations	US
13	Deka products LP	1	Organizations	US
14	Geo line co Ltd	1	Organizations	KR
15	Hitachi ltd	1	Organizations	JP
16	Invently llc	1	Organizations	US
17	Knickerbocker; CECIL	1	Individuals	US
18	Melrok llc	1	Organizations	US
19	Optimization technologies Inc	1	Organizations	US
20	Pedersen; Robert d	1	Individuals	US
21	Power technology holdings LLC	1	Organizations	US
22	v2 green inc	1	Organizations	US
23	Zhou; Andrew H B|Zhou; Dylan T X|Zhou; Tiger T G	1	Individuals	US

be preferable. Second, as for AUROC, similar results were obtained. In general, the performance of active learning was superior to that of semi-supervised learning, with a 0.8% increase for SVM and 1.8% for RF. For XGBoost, the AUROC value stayed the same; the small size of the data sample could produce great performance.

Ecosystem development

The proposed approach was applied to recent patents (997 patents published between 2014 and 2018) in the same area (EVB) for the experiments, to identify emerging promising technologies and ultimately envision a future innovation system based on the technologies. Active learning with XGBoost was adopted, which was found to have the best performance; accordingly, 51 patents were classified as promising, whereas the remaining 946 were determined to be non-promising.

A list of the applicants for the promising patents is provided in Table 5. Two types of applicants were observed: organizations and individuals, with an 18:5 ration. Quite naturally, organizations have more key patents than individuals. By country, the US had the most patents (20), while Japan, South Korea, and Singapore had one patent each. Then, the CPC-co-occurrence network was developed at the sub-group level, as shown in Fig. 9, for which only 14 applicants with relatively strong relationships with others were visualized

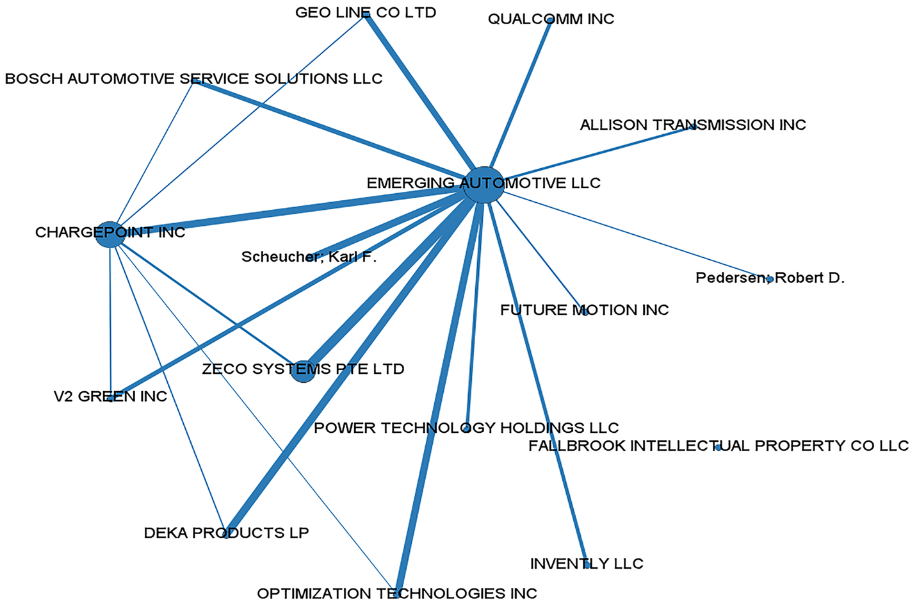


Fig. 9 Main actors network—business layers

for readability. In the figure, the node size is proportional to the betweenness centrality, while the width of edge is determined by the edge weight with the cut-off value of 4.35% (50) for edge weight.

In the next step, the relationships between patents were analyzed. Assuming that the first CPC was the most important code, a pair of patents obtained a value of 1 when they shared the same CPC codes at the main group level; otherwise, they obtained a value of 0. If necessary, all CPCs assigned to patents could be considered in defining the degree of classifications. Figure 10 describes the promising patents network where different colors are used to easily distinguish different clusters.

Finally, the main technological themes centered on the promising patents were investigated. Similarly to the other networks, a CPC co-occurrence matrix at the subgroup level was developed to build a network of CPCs. That is, CPC became a node, while the weight of its edge was defined by the number of co-classified patents; if CPC i and CPC j have k number of patents, all of which are assigned both CPCs, the weights of CPC i and CPC j would be k. For visualization, the top 10 CPCs in terms of their degrees of centrality were identified (see Table 6 and "Appendix" 7 for more details). The information and communications technologies for operating EVBs and the technologies for electric charging stations were ranked high according to the degree of centrality, while the battery monitoring and controlling technologies were ranked low. Figure 11 describes the network of main themes (CPCs) for emerging promising patents. In the figure, the node size is proportionate to the betweenness centrality with the cut-off value of 0.5% (16) for edge weight, while the edge color is determined by the cluster. Key themes included Y02T90/00 (enabling technologies or technologies with a potential or indirect contribution to GHG emission mitigation), B60L53/00 (methods of charging batteries, specially adapted for electric vehicles; charging stations or onboard charging equipment therefor; exchange of energy storage elements in electric vehicles), Y02T10/00 (road transport of goods or passengers), and Y04S30/00

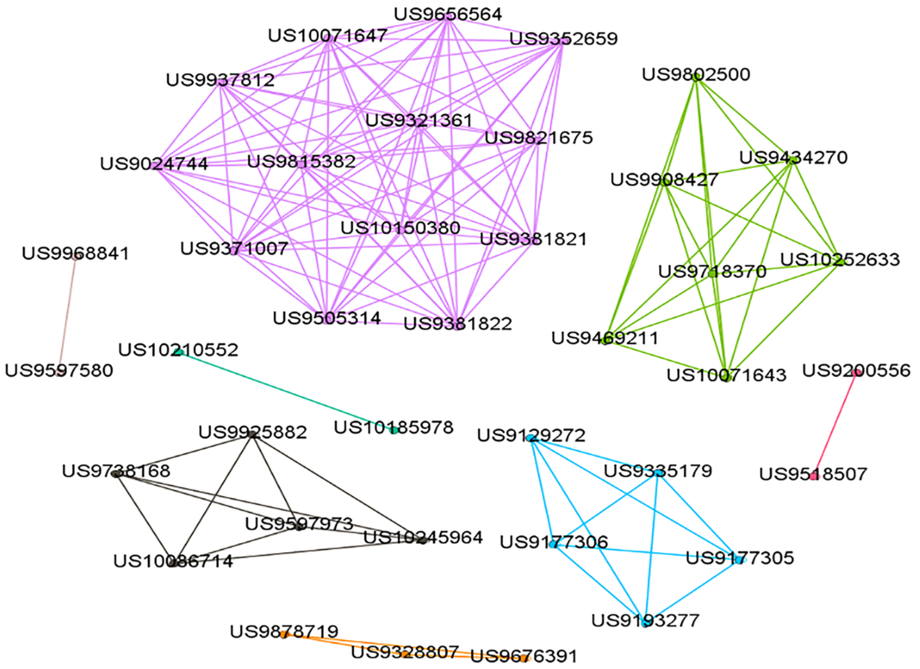


Fig. 10 Promising patents network—patent layer

(communication or information technology-specific aspects supporting electrical power generation, transmission, distribution or end-user application management).

Based on the analysis of the three different layers, a multilayered network showing the relationships both within and across layers was developed to envision the future innovation ecosystem (see Fig. 12). In the figure, the main actors network (business level) was positioned in the bottom layer, the emerging promising patents network (patent level) was positioned in the middle layer, and the themes network (technology level) was positioned in the top layer. This network enabled our understanding of the emerging promising technologies, their key R&D agents, and the main topics consisting of those technologies. Furthermore, the patents that the proposed approach expected to be promising played a significant role in linking innovation actors and areas of innovation. For example, QUALCOMM, one of the leading companies in EVB technologies, had three patents—US9381821, US9656564, and US9505314—that were expected to be promising. These patents were in the areas of Y02T90/121, Y02T90/163, Y02T10/7088, Y02T90/128, B60L53/305, and Y02T90/16, which are also closely related to each other (see Fig. 13).

Discussion

Methodological perspective

Several methodological issues need to be addressed for an effective utilization of the proposed approach. First, it is worth to discuss the model proposed in this study being

Table 6 Parts of degree centrality rankings of EVB technology

Rank	CPC	Description	Degree centrality
1	Y02T90/16	Information or communication technologies improving the operation of electric vehicles	18
2	Y02T90/128	Energy exchange control or determination	14
3	Y02T90/121	Electric charging stations by conductive energy transmission	13
4	Y02T90/163	Information or communication technologies related to charging of electric vehicles	13
5	Y02T10/7291	Optimization of vehicle performance by route optimization processing	12
6	Y02T90/169	Aspects supporting the interoperability of electric or hybrid vehicles (e.g., recognition, authentication, identification, or billing)	12
7	Y04S30/14	Details associated with the interoperability (e.g., vehicle recognition, authentication, identification, or billing)	12
8	B60L53/665	Data transfer between charging stations and vehicles methods related to measuring, billing, or payment	12
9	Y02T10/7088	Charging stations	10
10	B60L53/305	Constructional details of charging stations communication interfaces	9
…	…	…	…
22	G06Q20/18	Payment architectures involving self-service terminals, vending machines, kiosks, or multimedia terminals	4
23	B60L2240/70	Interactions with external data bases, e.g. traffic centres	2

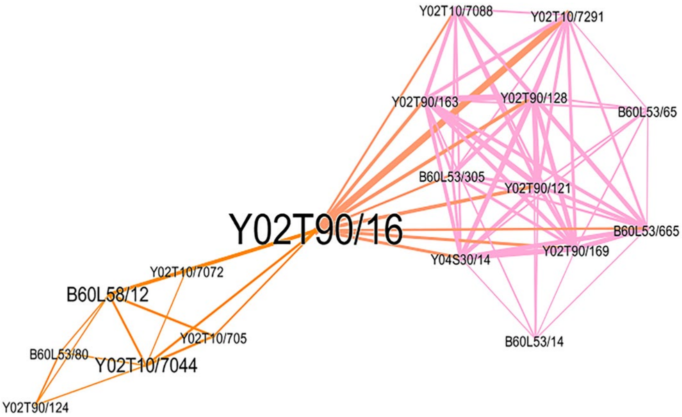


Fig. 11 Main themes network—technology layer

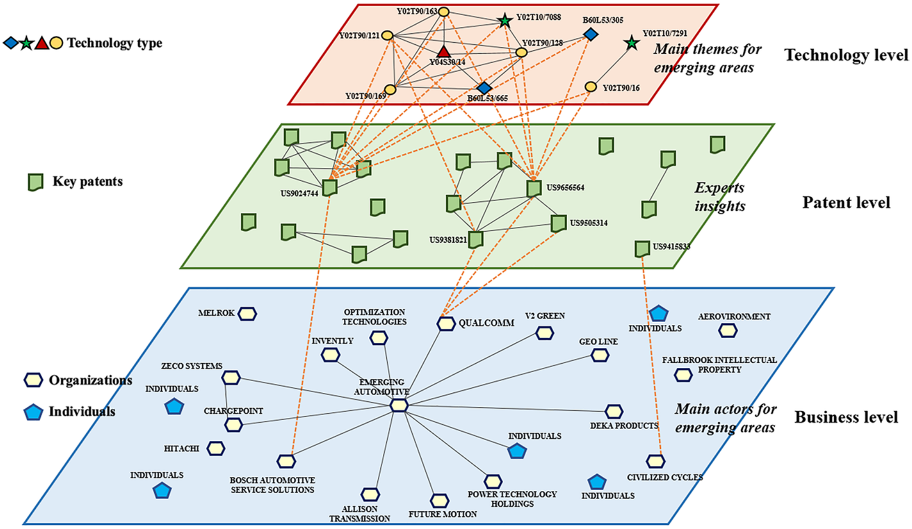


Fig. 12 A multilayered network to describe a future innovation ecosystem

compared with those in the other studies. However, since the purpose of this study is not to develop a machine learning algorithm with the higher performance but to apply it for technology mapping from the prospective perspective, the comparative study was aiming to review the literature for understanding the existing approaches to emphasize the contribution of this study. Accordingly, the model comparison was made in terms of target variable, input variables, analytic methods, and performance for the studies with similar purposes of identifying emerging technologies from patent data. The analysis results indicate that this study is different from the other studies in that it proposed the ways to involve experts during patent analysis (see "Appendix" 6 for details): (1) the use of a small set of labeled data for model training enables expert engagement in search for emerging technologies, (2)

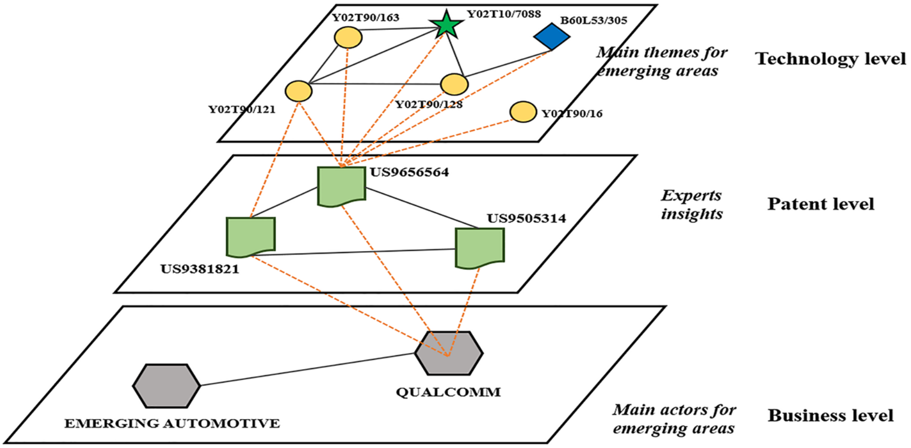


Fig. 13 Position of QUALCOMM within the innovation ecosystem

the expert engagement allows to identify emerging technologies from various viewpoints, and (3) the prediction performance could be improved by active learning. This study also adopted the characteristics of technological field in which a patent belongs on top of the characteristics of patent as input variables unlike the existing studies focusing mostly on the characteristics of patent. Considering both the patent and its relevant technological fields is expected to improve the performance of predicting emerging technology given that the growth of individual technology is likely to be affected by technological life cycle. In-depth analysis of variable section needs to be conducted along with the development of machine learning algorithms with better performance, being customized to patent data.

Second, the number of labeled data required for semi-supervised learning needs to be chosen carefully. The analysis results may be affected by the proportion of labeled data in the initial data set. In this study, the proportion was set to 30%, while the results with proportions of 10%, 20%, and 40% are summarized in "Appendix" 3. Thus, the proportion of labeled data needs to be determined before the analysis. Further studies will be needed to find the optimal proportion of labeled data to total data.

Third, for active learning, the criteria for labeling unlabeled data need to be defined clearly. In this study, the probability of being a promising patent was calculated and used for those criteria; the patents with probabilities between 0.4 and 0.6, which failed to obtain labels by the classifiers, obtained labels by human intervention. Yet various other criteria could be designed. For example, the data with probabilities slightly lower or higher than the cut-off values (and thus failing to obtain the labels based on the classifiers) could be labeled by human intervention. Another approach could be a random selection of patents that failed to obtain labels by the classifiers. Identifying the dataset that can better contribute to improving the model's performance is worth the extra work, especially in cases where the characteristics of patent data need to be considered.

Finally, this study adopted a machine learning technique using three classifiers, though there are other approaches also available for the problem in this study. For example, a regression model can be considered a basic model where the root mean squared error, mean absolute percentage error, and mean absolute squared error are used to evaluate model performance. In particular, recent advances in explainable machine learning models that enables to examine the behavior of the model, such as LIME (Local Interpretable Model-Agnostic Explanation) and SHAP (Shapley Additive exPlanations), can elaborate the proposed approach and enhance the analysis results significantly (Lundberg & Lee, 2017; Ribeiro et al., 2016). Machine learning models are divided into two types of approaches, which are black-box and explainable while-box approaches (Loyola-Gonzlez, 2019). A block-box approach was adopted in this study given that the purpose of this study was not to interpret the model. Nevertheless, valuable implications will be obtained if the key factors that made the technologies promising can be explained in the model. As the practices of identifying promising technologies are generally linked to financial investments, understanding such factors will increase the reliability of the analysis results.

Practical perspective

In addition to those from a methodological perspective, there are other practical issues to consider. First, the focus of this study was to accurately identify emerging promising patents from a large collection of patent documents. Consequently, the interpretations of the F1-scores emphasized the performance measures where the target value was set to PP. That is, the focus of the evaluation was the ratio of true promising patents to predicted promising patents. We found that these F1-values ranged from 0.562 to 0.739 because the potential promise of the patents can be affected by various factors apart from the technological features that can be explained by patent characteristics. By incorporating various perspectives in defining input variables, we tried to resolve this issue, but there is still room for improvement for the F1-score values. On the other hand, if none of the promising patents should be missed, the target value could be set to the NPP. The F1-score values for NPP are quite high, ranging from 0.930 to 0.970; thus, identifying NPPs are relatively accurate. After the NPPs are predicted by the proposed approach, the remaining patents can go through another round of review to evaluate their promise, which may decrease the time and effort required to evaluate patents. Additionally, it may fully consider experts' insights, missing few promising patents compared to conventional patent analysis approaches. Consequently, the proposed approach needs to be customized to the purpose of analysis along with the context of the analysis.

Second, to test the validity of the proposed approach in a real case setting, a mini experiment was designed using expert insights. However, as a more sophisticated approach would be needed to involved experts in the interactive experiments to avoid human errors. Hence, the experiment focused only on the usability of semi-supervised approach using experts' evaluations on promising patents. To obtain the labeled data, we randomly selected 200 patents from the 6,287 patents collected for our case study on EVBs. Then the patents were sent to two experts (one in academia and the other in practice) that have been working on the EVB technologies for more than five years. The promisingness of the patents was evaluated from the perspectives of technological value in the following five years; if any of the experts designated the patent as promising, it was coded to PP, whereas if none of the two experts designated the patents as promising, it was coded to NPP. Then the data were preprocessed and, as a result, only 86 data points among 200 were available; the experiment was conducted using 86 patents as labeled dataset and 2612 patents (the data available from 6087 patents, which is the size of unlabeled data before preprocessing) as unlabeled data. From the 86 patents, 35% (31 patents) were used for a test set with the remaining 65% (55 patents) being used for a training set. The Random Forest was adopted for the model fitting for supervised learning and semi-supervised learning methods. The analysis

Table 7 Performance evaluation using expert dataset

Iteration	PP				NPP			AUROC
	Accuracy	Precision	Recall	F1-score	Precision	Recall	F1-score	
Iter0	0.587	0.466	0.425	0.444	0.653	0.690	0.671	0.556
Iter1	0.646	0.450	0.505	0.474	0.754	0.713	0.733	0.602
Iter2	0.663	0.414	0.550	0.464	0.800	0.713	0.752	0.607
Iter3	0.700	0.375	0.681	0.463	0.879	0.721	0.790	0.per627

results for predicting both PPs and NPPs are presented in Table 7. The table showed that the highest performance was observed at the iteration 3 for a semi-supervised learning with the AUROC of 0.627. The F1-score for predicting PPs was 0.444, whereas the score for predicting NPPs was relatively higher, producing the value of 0.790. With only a small set of data (86 patents), we expect to identify NPPs effectively, which can reduce the time and efforts to evaluate each patent. A further study to involve experts during the experiments will be needed to test the performance of active learning.

Conclusions

This study proposed a novel approach to identifying emerging promising technologies by applying semi-supervised learning and active learning to patents. Furthermore, it explored the possibilities of efficiently integrating experts' insights into the patent analysis process. The analysis results helped us envision future innovation systems based on the promising patents identified by the proposed approach. Unlike previous studies that relied on a simple supervised-learning process, this study introduced semi-supervised and active learning to patent analysis. To do this, patent data on electric vehicle batteries were collected from the USPTO database. The model had 32 input variables and 1 target variable, and three algorithms—SVM, RF and XGBoost—were adopted for the training. Finally, the model was applied to more recent patents to identify emerging promising technologies (which were used to derive key innovation agents) and key innovation areas as well as their relationships. The research findings indicated that active learning can be a useful tool to incorporate experts' insights into the patent analysis process and verified its value in searching for promising patents and emerging technologies, and ultimately for describing future innovation systems.

Thus, we believe that this study can contribute to existing knowledge in three ways. First, we proposed a systematic approach to combine experts' insights and patent data in envisioning a future innovation system. In particular, the use of semi-supervised learning and active learning based on 32 ex-ante patent indices enabled us to increase the efficiency of analysis in identifying emerging promising technologies. Second, the proposed approach can be useful not only in identifying PPs but also screening NPPs; the latter might be a more realistic approach in practice. Finally, the proposed approach can be a basis for further study by proposing a way to balance expert-driven decision-making and data-driven decision-making as well as the concept of multilayered networks to describe future innovation systems.

Despite the meaningful contributions of this study, it is more for a feasibility test of active learning in identifying promising technologies. Accordingly, the proposed approach has several limitations, and future studies are needed. First, though promising patents (technologies) can be defined in a number of ways in different contexts, this study considered them as patents with a great impact on subsequent technologies. Thus, further analysis is needed to validate whether the proposed approach can produce as good a performance as that of in this study when applied to a wide range of other possibilities. Furthermore, this study adopted the forward citation frequency as a proxy for experts' inputs for patent evaluation. In reality, however, promisingness is often determined by various other factors that are not explained by patent information. Accordingly, in addition to the experiment implemented in this study, a real case study involving interactive feedback from experts is required to acquire external validity. Second, the proposed model can be expanded to include more variables or adopt better algorithms for semi-supervised learning. For example, a graph-based semi-supervised learning system is available that uses co-citation or co-classification information as input features and other advanced analysis algorithms (i.e., Generative Adversarial Network, Reinforcement learning) that could be helpful for identifying promising patents with higher accuracy. Combining several learning algorithms to make the best decisions is worth considering. Combining a composite patent quality index with experts' insights is also worth to consider. Finally, more specific guidelines are needed for semi-supervised learning and active learning for easier application of the proposed approach in practice. For example, the amount of labeled data required to generate a satisfactory performance or the types of unlabeled data required to get feedback from experts are the key issues to be addressed in future research. Further studies will address these issues and the GitHub address of the code is provided (https://github.com/pphanho/emerging) for reproducibility and continuous development in the methodology.

Appendix 1: Basic information about dataset

See Tables 8, 9, 10, 11, 12.

Table 8 Descriptive statistics for variables

No.	Variables	N	Mean	Std. dev	Min	Max
1	Scope of patent right	2698	2.51	1.68	1.00	22.00
2	Degree of novelty	2698	29.03	59.88	0.00	882.00
3	Scope of inventors	2698	2.82	1.89	1.00	15.00
4	Technological insight	2698	2.64	10.64	0.00	178.00
5	Scope of patent images	2698	8.85	8.17	0.00	92.00
6	Scope of patent market	2698	4.05	2.68	1.00	29.00
7	Assignee quality (backward citation)	2698	29.01	57.69	1.00	751.25
8	Assignee quality (forward citation)	2698	39.57	54.60	0.00	715.00
9	Assignee coverage	2698	4.03	2.17	1.00	27.00
10	Technology quality (backward citation)	2698	40.05	15.03	11.57	195.71
11	Technology quality (forward citation)	2698	55.89	23.80	10.36	329.42
12	Technology coverage	2698	4.33	0.60	2.08	11.56
13	Technology recency	2698	1.65	2.29	0.00	19.00
14	Title noun words	2698	2.31	1.90	0.00	18.00
15	Recombinant novelty	2698	0.06	0.14	0.00	1.00
16	Component familiarity	2698	206.37	229.59	0.00	1398.58
17	Scope of applicants	2698	1.08	0.33	1.00	10.00
18	Scope of applicant countries	2698	1.00	0.15	0.00	2.00
19	Scope of inventor countries	2698	1.03	0.17	1.00	2.00
20	Scope of technology (IPC)	2698	3.44	3.37	1.00	29.00
21	Scope of technology (CPC)	2698	17.40	13.03	1.00	98.00
22	Diversity of technology (IPC)	2698	1.90	1.21	1.00	11.00
23	Diversity of technology (CPC)	2698	4.16	1.95	1.00	16.00
24	Assignee recency	2698	0.55	0.50	0.00	1.00
25	Word1	2698	0.22	0.41	0.00	1.00
26	Word2	2698	0.14	0.35	0.00	1.00
27	Word3	2698	0.08	0.27	0.00	1.00
28	Word4	2698	0.04	0.21	0.00	1.00
29	Word5	2698	0.05	0.21	0.00	1.00
31	Word6	2698	0.04	0.20	0.00	1.00
32	Word7	2698	0.04	0.20	0.00	1.00
33	Word8	2698	0.03	0.17	0.00	1.00
34	Word9	2698	0.03	0.18	0.00	1.00
35	Word10	2698	0.03	0.18	0.00	1.00
36	Promisingness	2698	0.10	0.30	0.00	1.00

Table 9 Means and standard deviations of labeled data: SVM

Algorithm: support vector machine		SL		SSL		AL	
No	Variables	Mean	Std. dev	Mean	Std. dev	Mean	Std. dev
1	Scope of patent right	1.19	0.42	1.17	0.41	1.18	0.41
2	Degree of novelty	2.89	0.81	2.88	0.84	2.90	0.83
3	Scope of inventors	2.78	1.89	2.70	1.84	2.79	1.90
4	Technological insight	0.58	0.84	0.59	0.86	0.60	0.88
5	Scope of patent images	2.04	0.66	2.01	0.65	2.01	0.67
6	Scope of patent market	1.52	0.47	1.54	0.47	1.53	0.48
7	Assignee quality (backward citation)	3.04	0.62	3.04	0.66	3.05	0.66
8	Assignee quality (forward citation)	3.37	0.76	3.33	0.81	3.35	0.81
9	Assignee coverage	1.55	0.37	1.57	0.37	1.57	0.37
10	Technology quality (backward citation)	3.66	0.29	3.67	0.30	3.67	0.31
11	Technology quality (forward citation)	3.99	0.34	3.98	0.36	3.99	0.36
12	Technology coverage	1.67	0.10	1.67	0.10	1.67	0.11
13	Technology recency	0.78	0.63	0.76	0.65	0.76	0.65
14	Title noun words	2.36	2.09	2.30	1.99	2.33	2.00
15	Recombinant novelty	0.04	0.10	0.04	0.10	0.04	0.11
16	Component familiarity	213.56	241.59	205.77	232.58	202.65	232.89
17	Scope of applicants	0.73	0.14	0.73	0.13	0.73	0.13
20	Scope of technology (IPC)	1.27	0.61	1.27	0.61	1.27	0.61
21	Scope of technology (CPC)	2.71	0.71	2.68	0.73	2.69	0.72
22	Diversity of technology (IPC)	1.88	1.20	1.88	1.22	1.89	1.25
23	Diversity of technology (CPC)	4.21	1.99	4.21	2.15	4.23	2.14

Table 10 Means and standard deviations of labeled data: random forest

Algorithm: random forest		SL		SSL		AL	
No	Variables	Mean	Std. dev	Mean	Std. dev	Mean	Std. dev
1	Scope of patent right	1.19	0.42	1.17	0.41	1.19	0.42
2	Degree of novelty	2.89	0.81	2.90	0.85	2.93	0.85
3	Scope of inventors	2.78	1.89	2.75	1.85	2.84	1.96
4	Technological insight	0.58	0.84	0.59	0.89	0.61	0.90
5	Scope of patent images	2.04	0.66	2.06	0.64	2.08	0.66
6	Scope of patent market	1.52	0.47	1.54	0.46	1.55	0.48
7	Assignee quality (backward citation)	3.04	0.62	3.06	0.67	3.07	0.67
8	Assignee quality (forward citation)	3.37	0.76	3.41	0.77	3.46	0.77
9	Assignee coverage	1.55	0.37	1.58	0.36	1.58	0.38
10	Technology quality (backward citation)	3.66	0.29	3.67	0.30	3.67	0.30
11	Technology quality (forward citation)	3.99	0.34	4.00	0.35	4.01	0.35
12	Technology coverage	1.67	0.10	1.67	0.10	1.67	0.10
13	Technology recency	0.78	0.63	0.78	0.65	0.80	0.65
14	Title noun words	2.36	2.09	2.24	1.97	2.29	1.96
15	Recombinant novelty	0.04	0.10	0.04	0.10	0.04	0.11
16	Component familiarity	213.56	241.59	212.13	238.37	202.77	233.91
17	Scope of applicants	0.73	0.14	0.73	0.13	0.73	0.13
20	Scope of technology (IPC)	1.27	0.61	1.27	0.62	1.26	0.62
21	Scope of technology (CPC)	2.71	0.71	2.73	0.71	2.75	0.71
22	Diversity of technology (IPC)	1.88	1.20	1.89	1.22	1.88	1.24
23	Diversity of technology (CPC)	4.21	1.99	4.29	2.10	4.37	2.12

Table 11 Means and standard deviations of labeled data: XGBoost

Algorithm: XGBoost		SL		SSL		AL	
No	Variables	Mean	Std. dev	Mean	Std. dev	Mean	Std. dev
1	Scope of patent right	1.19	0.42	1.17	0.41	1.18	0.42
2	Degree of novelty	2.89	0.81	2.87	0.83	2.91	0.87
3	Scope of inventors	2.78	1.89	2.74	1.84	2.76	1.83
4	Technological insight	0.58	0.84	0.62	0.89	0.64	0.91
5	Scope of patent images	2.04	0.66	2.02	0.65	2.05	0.67
6	Scope of patent market	1.52	0.47	1.53	0.48	1.54	0.49
7	Assignee quality (backward citation)	3.04	0.62	3.04	0.66	3.06	0.69
8	Assignee quality (forward citation)	3.37	0.76	3.29	0.85	3.33	0.87
9	Assignee coverage	1.55	0.37	1.57	0.38	1.58	0.39
10	Technology quality (backward citation)	3.66	0.29	3.65	0.31	3.66	0.31
11	Technology quality (forward citation)	3.99	0.34	3.93	0.39	3.94	0.39
12	Technology coverage	1.67	0.10	1.68	0.12	1.68	0.12
13	Technology recency	0.78	0.63	0.75	0.64	0.75	0.65
14	Title noun words	2.36	2.09	2.33	2.01	2.37	2.00
15	Recombinant novelty	0.04	0.10	0.04	0.11	0.05	0.12
16	Component familiarity	213.56	241.59	190.24	222.27	190.87	224.62
17	Scope of applicants	0.73	0.14	0.73	0.13	0.73	0.13
20	Scope of technology (IPC)	1.27	0.61	1.25	0.61	1.26	0.61
21	Scope of technology (CPC)	2.71	0.71	2.60	0.74	2.62	0.75
22	Diversity of technology (IPC)	1.88	1.20	1.81	1.20	1.83	1.23
23	Diversity of technology (CPC)	4.21	1.99	4.04	2.15	4.09	2.15

Table 12 Correlation coefficient matrix

Input variables	Correlation coefficient values									
Scope of patent right	1.000									
Degree of novelty	0.039	1.000								
Scope of inventors	−0.019	0.021	1.000							
Technological insight	0.031	0.559	0.050	1.000						
Scope of patent images	0.048	0.229	0.197	0.142	1.000					
Scope of patent market	−0.035	0.297	−0.014	0.185	0.174	1.000				
Assignee quality (backward citation)	0.032	0.963	0.015	0.546	0.204	0.282	1.000			
Assignee quality (forward citation)	0.090	0.465	0.000	0.204	0.160	0.270	0.483	1.000		
Assignee coverage	−0.013	0.337	−0.006	0.198	0.160	0.808	0.350	0.336	1.000	
Technology quality (backward citation)	0.005	0.480	−0.009	0.226	0.132	0.205	0.469	0.503	0.248	1.000

Table 12 (continued)

Input variables	Correlation coefficient values																			
Technology quality (forward citation)	0.056	0.316	−0.042	0.104	0.090	0.146	0.308	0.595	0.186	0.793	1.000									
Technology coverage	−0.049	0.225	0.052	0.124	0.056	0.382	0.221	0.274	0.377	0.475	0.315	1.000								
Technology recency	0.000	0.386	0.033	0.206	0.214	0.197	0.359	0.437	0.212	0.438	0.352	0.175	1.000							
Title noun words	0.125	−0.052	0.057	0.016	−0.033	−0.075	−0.043	−0.064	−0.059	−0.111	−0.102	−0.079	−0.026	1.000						
Recombinant novelty	0.035	0.046	0.013	0.014	0.046	0.007	0.045	0.046	−0.001	−0.009	−0.012	−0.018	0.015	0.013	1.000					
Component familiarity	−0.098	0.047	−0.013	0.049	0.052	0.056	0.032	−0.101	0.008	0.070	0.004	−0.053	0.330	0.010	−0.003	1.000				
Scope of applicants	−0.063	0.026	0.103	0.005	−0.024	0.010	0.027	0.027	0.012	0.042	0.023	0.007	0.085	−0.035	−0.009	0.019	1.000			
Scope of applicant countries	−0.030	−0.064	0.015	0.024	−0.039	−0.021	−0.059	−0.124	−0.040	−0.103	−0.143	−0.041	−0.074	0.048	−0.006	0.024	0.200	1.000		
Scope of inventor countries	−0.014	0.040	0.187	0.007	0.071	0.056	0.039	0.038	0.076	0.021	0.009	0.043	−0.019	0.023	0.001	−0.020	0.036	0.077	1.000	
Scope of technology (IPC)	−0.085	0.119	0.023	0.125	0.097	0.128	0.103	−0.007	0.083	0.113	0.019	0.053	0.340	0.036	0.166	0.640	0.018	0.035	0.006	1.000

Table 12 (continued)

Input variables	Correlation coefficient values																						
Scope of technology (CPC)	0.017	0.335	0.004	0.165	0.261	0.205	0.300	0.426	0.211	0.366	0.403	0.089	0.700	−0.049	0.019	0.241	0.026	−0.099	−0.005	0.291	1.000		
Diversity of technology (IPC)	−0.063	0.091	−0.022	0.062	0.064	0.097	0.073	0.009	0.040	0.095	0.068	−0.021	0.251	−0.032	0.399	0.658	0.053	0.021	−0.010	0.748	0.228	1.000	
Diversity of technology (CPC)	**0.055**	0.276	−0.010	0.110	0.193	0.210	0.249	0.410	0.210	0.340	0.453	0.098	0.437	−0.056	0.116	0.194	0.025	−0.098	0.003	0.215	0.701	0.309	1.000

Appendix 2: Data preprocessing results

See Table 13.

Table 13 Data pre-processing result for skewness and near-zero variance

No	Variables	Skewness	Near zero variance
1	Scope of patent right	**3.154548**	False
2	Degree of novelty	**7.976391**	False
3	Scope of inventors	1.528660	False
4	Technological insight	**9.753616**	False
5	Scope of patent images	**3.621313**	False
6	Scope of patent market	**2.303232**	False
7	Assignee quality (backward citation)	**8.492116**	False
8	Assignee quality (forward citation)	**8.119434**	False
9	Assignee coverage	**2.830502**	False
10	Technology quality (backward citation)	**4.441738**	False
11	Technology quality (forward citation)	**4.765845**	False
12	Technology coverage	**3.976486**	False
13	Technology recency	**4.056870**	False
14	Title noun words	1.409008	False
15	Recombinant novelty	**2.885501**	False
16	Component familiarity	1.627266	False
17	Scope of applicants	**10.15448**	False
18	Scope of applicant countries	– 1.017930	**True**
19	Scope of inventor countries	**5.362227**	**True**
20	Scope of technology (IPC)	**2.153643**	False
21	Scope of technology (CPC)	**2.158063**	False
22	Diversity of technology (IPC)	1.785868	False
23	Diversity of technology (CPC)	1.319517	False

Bold values indicate skewness greater than 2 or True near-zero variance

Appendix 3: Results based on percentage of labeled data

See Table 14.

Table 14 Total performance evaluation

Measure		Accuracy				Precision				Recall				F1-Score				AUROC			
		10%	20%	30%	40%	10%	20%	30%	40%	10%	20%	30%	40%	10%	20%	30%	40%	10%	20%	30%	40%
SL	SVM	0.799	0.907	0.896	0.888	0.299	0.526	0.486	0.462	0.741	0.741	0.667	0.667	0.426	0.615	0.562	0.545	0.822	0.906	0.886	0.881
	RF	0.907	0.929	0.911	0.918	0.528	0.618	0.545	0.593	0.704	0.778	0.667	0.593	0.603	0.689	0.600	0.593	0.882	0.926	0.881	0.875
	XGB	0.918	0.948	0.948	0.959	0.778	0.882	0.882	0.900	0.259	0.556	0.556	0.667	0.389	0.682	0.682	0.766	0.899	0.954	0.929	0.969
SSL	SVM	0.799	0.903	0.877	0.885	0.286	0.513	0.425	0.444	0.667	0.741	0.630	0.593	0.400	0.606	0.507	0.508	0.803	0.912	0.885	0.889
	RF	0.818	0.944	0.918	0.937	0.304	0.750	0.586	0.727	0.630	0.667	0.630	0.593	0.410	0.706	0.607	0.653	0.795	0.903	0.874	0.901
	XGB	0.922	0.941	0.952	0.963	0.667	0.824	0.889	0.905	0.444	0.519	0.593	0.704	0.533	0.636	0.711	0.792	0.91	0.944	0.923	0.967
AL	SVM	0.851	0.911	0.911	0.926	0.333	0.560	0.552	0.613	0.481	0.519	0.593	0.704	0.394	0.538	0.571	0.655	0.847	0.894	0.894	0.917
	RF	0.903	0.929	0.944	0.941	0.520	0.722	0.773	0.762	0.481	0.481	0.630	0.593	0.500	0.578	0.694	0.667	0.847	0.879	0.899	0.889
	XGB	0.937	0.952	0.955	0.963	0.750	0.850	0.895	1.000	0.556	0.630	0.630	0.630	0.638	0.723	0.739	0.773	0.938	0.946	0.929	0.973

Table 15 Classification results by classifier

		SVM	RF	XGB
SL	Recall – A	0.667	0.667	0.556
	Recall – B	0.921	0.938	0.992
	Precision – A	0.486	0.545	0.882
	Precision – B	0.961	0.962	0.952
	F1 score – A	0.562	0.600	0.682
	F1 score – B	0.941	0.950	0.972
SSL	Recall – A	0.630	0.630	0.593
	Recall – B	0.905	0.950	0.992
	Precision – A	0.425	0.586	0.889
	Precision – B	0.956	0.958	0.956
	F1 score – A	0.507	0.607	0.711
	F1 score – B	0.930	0.954	0.974
AL	Recall – A	0.593	0.630	0.630
	Recall – B	0.921	0.979	0.992
	Precision – A	0.552	0.773	0.895
	Precision – B	0.961	0.960	0.960
	F1 score – A	0.571	0.694	0.739
	F1 score – B	0.941	0.969	0.976

A—EP and B—NPP

Appendix 4: Performance results in different classes (NPP)

See Table 15.

Appendix 5: ROC plots and performance results for each iteration

See Fig. 14, Table 16.

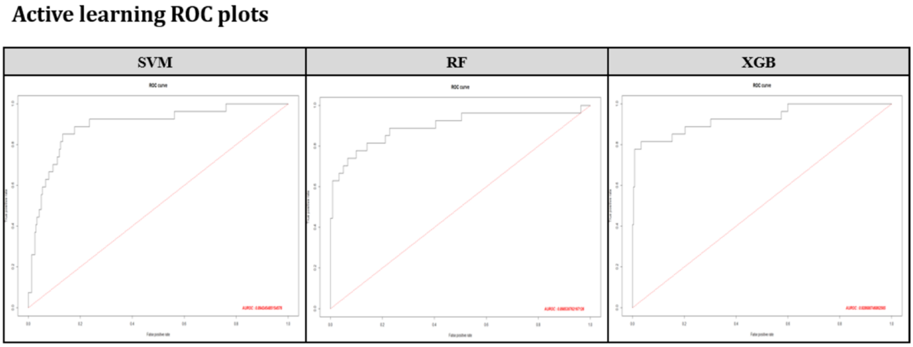


Fig. 14 ROC plots according by classifier

Table 16 Total performance over time

Active learning Labeling percentage 30%					
	Accuracy	Precision	Recall	F1-Score	AUROC
SVM					
Iteration0	0.896	0.486	0.667	0.562	0.886
Iteration1	0.911	0.545	0.667	0.6	0.907
Iteration2	0.888	0.459	0.63	0.531	0.892
Iteration3	0.892	0.474	0.667	0.554	0.888
Iteration4	0.874	0.41	0.593	0.485	0.866
Iteration5	0.892	0.474	0.667	0.554	0.895
Iteration6	0.907	0.531	0.63	0.576	0.896
Iteration7	0.907	0.536	0.556	0.545	0.878
Iteration8	0.911	0.552	0.593	0.571	0.894
RF					
Iteration0	0.911	0.545	0.667	0.6	0.881
Iteration1	0.937	0.667	0.741	0.702	0.894
Iteration2	0.929	0.625	0.741	0.678	0.922
Iteration3	0.941	0.69	0.741	0.714	0.918
Iteration4	0.944	0.714	0.741	0.727	0.904
Iteration5	0.937	0.679	0.704	0.691	0.892
Iteration6	0.944	0.75	0.667	0.706	0.899
Iteration7	0.944	0.731	0.704	0.717	0.885
Iteration8	0.944	0.773	0.63	0.694	0.899
XGB					
Iteration0	0.948	0.882	0.556	0.682	0.929
Iteration1	0.952	0.889	0.593	0.711	0.932
Iteration2	0.948	0.882	0.556	0.682	0.923
Iteration3	0.955	0.895	0.63	0.739	0.924
Iteration4	0.952	0.889	0.593	0.711	0.926
Iteration5	0.948	0.882	0.556	0.682	0.932
Iteration6	0.955	0.941	0.593	0.727	0.932
Iteration7	0.955	0.895	0.63	0.739	0.933
Iteration8	0.955	0.895	0.63	0.739	0.929

Appendix 6: Performance comparison with previous studies

See Table 17.

Table 17 Performance comparison with previous literatures

Title	Aim	Target variables	Input variables	Main analytical method	Performance
Forecasting emerging technologies: a supervised learning approach through patent analysis (Kyebambe et al., 2017)	To forecast emerging technologies	Promisingness	7 patent indicators: number of claims, number of citations, etc	Support vector machine	Accuracy: 70.60 F-measure: 0.72
				Artificial neural network	Accuracy: 54.20 F-measure: 0.54
				Random forest	F-measure: 0.55
Early identification of emerging technologies: A machine learning approach using multiple patent indicators (Lee et al., 2018)	To identify emerging technologies	Potential impact (forward citations over 3 years)	18 patent indicators: number of dependent claims, number of non-patent literature references, median age of cited patents, etc	Artificial neural network	Average accuracy: 0.91 Precision: 0.77 Recall: 0.37
				Random forest	Average accuracy: 0.91 Precision: 0.59 Recall: 0.34
				Support vector machine	Average accuracy: 0.91 Precision: 0.57 Recall: 0.38
Forecasting Forward Patent Citations: Comparison of Citation-Lag Distribution, Tobit Regression, and Deep Learning Approaches (Noh et al., 2020)	To forecast the value of patents	The impact on subsequent technological advancements (forward patents citations)	7 patent indicators: number of IPCs, value of Recombinant novelty, etc	Tobit regression	MAE: 0.987 R-square: 0.185
				Feed forward neural network	MAE: 0.983 R-square: 0.148

Table 17 (continued)

Title	Aim	Target variables	Input variables	Main analytical method	Performance
The proposed approach	To predict emerging promising technologies	Promisingness (high forwarded citation patents)	32 patent indicators: number of independent claims, number of inventors, frequency of the top 10 words, etc	Support vector machine	Accuracy: 0.91 Precision: 0.55 Recall: 0.52 AUROC: 0.89
				Random forest	Accuracy: 0.94 Precision: 0.77 Recall: 0.48 AUROC: 0.90
				Extreme gradient boosting	Accuracy: 0.96 Precision: 0.90 Recall: 0.63 AUROC: 0.93

Appendix 7: Results of degree centrality ranking of EVB technology

See Table 18.

Table 18 Results of degree centrality ranking of EVB technology

Rank	CPC	Description	Degree centrality
1	Y02T90/16	Information or communication technologies improving the operation of electric vehicles	18
2	Y02T90/128	Energy exchange control or determination	14
3	Y02T90/121	Electric charging stations by conductive energy transmission	13
4	Y02T90/163	Information or communication technologies related to charging of electric vehicle	13
5	Y02T10/7291	Optimisation of vehicle performance by route optimisation processing	12
6	Y02T90/169	Aspects supporting the interoperability of electric or hybrid vehicles, e.g. recognition, authentication, identification or billing	12
7	Y04S30/14	Details associated with the interoperability, e.g. vehicle recognition, authentication, identification or billing	12
8	B60L53/665	Data transfer between charging stations and vehicles methods related to measuring, billing or payment	12
9	Y02T10/7088	Charging stations	10
10	B60L53/305	Constructional details of charging stations communication interfaces	9
11	B60L53/65	Monitoring or controlling charging stations involving identification of vehicles	8
12	B60L53/14	Conductive energy transfer	8
13	B60L53/80	Exchanging energy storage elements, e.g. removable batteries	7
14	Y02T10/7044	Controlling the battery or capacitor state of charge	7
15	Y02T90/124	Electric charging stations by exchange of energy storage elements	7
16	Y02T10/7072	Electromobility specific charging systems or methods for batteries, ultracapacitors, supercapacitors or double-layer capacitors	6
17	B60L58/12	Methods or circuit arrangements for monitoring or controlling batteries or fuel cells, specially adapted for electric vehicles for monitoring or controlling batteries responding to state of charge	6
18	Y02T10/705	Controlling vehicles with one battery or one capacitor only	6
19	Y02E60/721	Systems characterised by the monitored, controlled or operated power network elements or equipments the elements or equipments being or involving electric vehicles [EV] or hybrid vehicles [HEV], i.e. power aggregation of EV or HEV, vehicle to grid arrangements [V2G]	4

Table 18 (continued)

Rank	CPC	Description	Degree centrality
20	Y04S10/126	Systems characterised by the monitored, controlled or operated power network elements or equipment the elements or equipment being or involving energy generation units, including distributed generation [DER] or load-side generation the energy generation units being or involving electric vehicles [EV] or hybrid vehicles [HEV], i.e. power aggregation of EV or HEV, vehicle to grid arrangements [V2G]	4
21	G06Q50/06	Electricity, gas or water supply	4
22	G06Q20/18	Payment architectures involving self-service terminals, vending machines, kiosks or multimedia terminals	4
23	B60L2240/70	Interactions with external data bases, e.g. traffic centres	2

Acknowledgements We appreciate Prof. Daniel Stefan Hain and Prof. Roman Jurowetzki for their suggestions of applying active learning to this study. We also appreciate Prof. ZHOU Yuan, Joseph for his constructive comments on the previous version of this paper.

References

Altuntas, S., Dereli, T., & Kusiak, A. (2015). Forecasting technology success based on patent data. *Technological Forecasting and Social Change, 96*, 202–214

Ando, R. K., & Zhang, T. (2005). A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research, 6*, 1817–1853

Archibugi, D., & Planta, M. (1996). Measuring technological change through patents and innovation surveys. *Technovation, 16*(9), 451–519

Arora, A., & Fosfuri, A. (2003). Licensing the market for technology. *Journal of Economic Behavior AND Organization, 52*(2), 277–295

Bekkers, R., Bongard, R., & Nuvolari, A. (2011). An empirical study on the determinants of essential patent claims in compatibility standards. *Research Policy, 40*(7), 1001–1015

Belkin, M., Niyogi, P., & Sindhwani, V. (2006). Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of machine learning research, 7*(11), 2399–2434

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

Bröring, S., Martin Cloutier, L., & Leker, J. (2006). The front end of innovation in an era of industry convergence: Evidence from nutraceuticals and functional foods. *R&D Management, 36*(5), 487–498

Caviggioli, F. (2016). Technology fusion: Identification and analysis of the drivers of technology convergence using patent data. *Technovation, 55*, 22–32

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785–794.

Chen, X., & Deng, N. (2015). A semi-supervised machine learning method for Chinese patent effect annotation. In 2015 international conference on cyber-enabled distributed computing and knowledge discovery, pp. 243–250.

Choi, C., & Park, Y. (2009). Monitoring the organic structure of technology based on the patent development paths. *Technological Forecasting and Social Change, 76*(6), 754–768

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

Cohn, D. A., Ghahramani, Z., & Jordan, M. I. (1996). Active learning with statistical models. *Journal of Artificial Intelligence Research, 4*, 129–145

Cohn, J. F., Zlochower, A., Lien, J., & Kanade, T. (1999). Automated face analysis by feature point tracking has high concurrent validity with manual FACS coding. *Psychophysiology, 36*, 35–43.

Crawford, M. M., Tuia, D., & Yang, H. L. (2013). Active learning: Any value for classification of remotely sensed data? *Proceedings of the IEEE, 101*(3), 593–608

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

Ernst, H. (2003). Patent information for strategic technology management. *World Patent Information, 25*(3), 233–242

Fischer, J. E., Bachmann, L. M., & Jaeschke, R. (2003). A readers' guide to the interpretation of diagnostic test properties: Clinical example of sepsis. *Intensive Care Medicine, 29*(7), 1043–1051

Fischer, T., & Leidinger, J. (2014). Testing patent value indicators on directly observed patent value—An empirical analysis of Ocean Tomo patent auctions. *Research Policy, 43*(3), 519–529

Fleming, L. (2001). Recombinant uncertainty in technological search. *Management Science, 47*(1), 117–132

Fleming, L., Mingo, S., & Chen, D. (2007). Collaborative brokerage, generative creativity, and creative success. *Administrative Science Quarterly, 52*(3), 443–475

Geum, Y., Kim, M. S., & Lee, S. (2016). How industrial convergence happens: A taxonomical approach based on empirical evidences. *Technological Forecasting and Social Change, 107*, 112–120

Giuri, P., Munari, F., & Pasquini, M. (2013). What determines university patent commercialization? Empirical evidence on the role of IPR ownership. *Industry and Innovation, 20*(5), 488–502

Guellec, D., & de la Potterie, B. V. P. (2000). Applications, grants and the value of patent. *Economics letters, 69*(1), 109–114

Hall, B. H., Jaffe, A., & Trajtenberg, M. (2005). Market value and patent citations. *The RAND Journal of Economics, 36*(1), 16–38

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

Hirschey, M., & Richardson, V. J. (2001). Valuation effects of patent quality: A comparison for Japanese and US firms. *Pacific-Basin Finance Journal, 9*(1), 65–82

HLT-NAACL, 152–159. Training and assessing classification rules with imbalanced data

Hsieh, C. H. (2013). Patent value assessment and commercialization strategy. *Technological Forecasting and Social Change, 80*(2), 307–319

Jun, S., Sung Park, S., & Sik Jang, D. (2012). Technology forecasting using matrix map and patent clustering. *Industrial Management & Data Systems, 112*(5), 786–807

Kang, B., & Bekkers, R. (2015). Just-in-time patents and the development of standards. *Research Policy, 44*(10), 1948–1961

Kaplan, S., & Vakili, K. (2015). The double-edged sword of recombination in breakthrough innovation. *Strategic Management Journal, 36*(10), 1435–1457

Kim, C., Lee, H., Seol, H., & Lee, C. (2011). Identifying core technologies based on technological cross-impacts: An association rule mining (ARM) and analytic network process (ANP) approach. *Expert Systems with Applications, 38*(10), 12559–12564

Kim, G., & Bae, J. (2017). A novel approach to forecast promising technology through patent analysis. *Technological Forecasting and Social Change, 117*, 228–237

Kim, H., Hong, S., Kwon, O., & Lee, C. (2017). Concentric diversification based on technological capabilities: Link analysis of products and technologies. *Technological Forecasting and Social Change, 118*, 246–257

Kim, J., & Lee, S. (2015). Patent databases for innovation studies: A comparative analysis of USPTO, EPO, JPO and KIPO. *Technological Forecasting and Social Change, 92*, 332–345

Kim, J. S., Lee, Y. Y., & Kim, T. H. (2016). A review on alkaline pretreatment technology for bioconversion of lignocellulosic biomass. *Bioresource technology, 199*, 42-48.

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

Lai, C., Hwang, S., & Wei, C. (2018). On the patent claim eligibility prediction using text mining techniques. Proceedings of the 51st Hawaii International Conference on System Sciences, 587–596

Lee, C., Cho, Y., Seol, H., & Park, Y. (2012). A stochastic patent citation analysis approach to assessing future technological impacts. *Technological Forecasting and Social Change, 79*(1), 16–29

Lee, C., Kim, J., Kwon, O., & Woo, H. G. (2016). Stochastic technology life cycle analysis using multiple patent indicators. *Technological Forecasting and Social Change, 106*, 53–64

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, D. S., Park, J. M., & Vanrolleghem, P. A. (2005). Adaptive multiscale principal component analysis for on-line monitoring of a sequencing batch reactor. *Journal of Biotechnology*, *116*(2), 195–210.

Lee, J., Kim, J., Lee, S., Seo, D., Jung, H., & Sung, W. K. (2011). Towards discovering emerging technologies based on decision tree. In 2011 International Conference on Internet of Things and 4th International Conference on Cyber, Physical and Social Computing, 529–532

Lee, S., Lee, S., Seol, H., & Park, Y. (2008). Using patent information for designing new product and technology: keyword based technology roadmapping. *R&d Management, 38*(2), 169–188

Lee, Y., & Colarelli O'Connor, G. (2003). The impact of communication strategy on launching new products: The moderating role of product innovativeness. *Journal of Product Innovation Management, 20*(1), 4–21

Leng, Y., Xu, X., & Qi, G. (2013). Combining active learning and semi-supervised learning to construct SVM classifier. *Knowledge-Based Systems, 44*, 121–131

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

Li, M., & Zhou, Z. H. (2005). SETRED: Self-training with editing. In Pacific-Asia Conference on Knowledge Discovery and Data Mining, 611–621.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news, 2*(3), 18–22

Liu, G., Nguyen, T. T., Zhao, G., Zha, W., Yang, J., Cao, J., Wu, M., Zhao, P., & Chen, W. (2016). Repeat buyer prediction for e-commerce. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 155–164.

Livotov, P. (2015). Using patent information for identification of new product features with high market potential. *Procedia engineering, 131*, 1157–1164

Loyola-Gonzalez, O. (2019). Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. *IEEE Access, 7*, 154096–154113

Lundberg, S., & Lee, S. I. (2017). A unified approach to interpreting model predictions. arXiv preprint arXiv:1705.07874.

Maulik, U., & Chakraborty, D. (2011). A self-trained ensemble with semisupervised SVM: An application to pixel classification of remote sensing imagery. *Pattern Recognition, 44*(3), 615–623

McClosky, D., Charniak, E., & Johnson, M. (2006). Effective Self-Training for Parsing. In Proceedings of

Menardi, G., & Torelli, N. (2014). Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery, 28*(1), 92–122

Mitchell, V. W. (1992). Using Delphi to forecast in new technology industries. *Marketing Intelligence & Planning, 10*(2), 4–9

Momeni, A., & Rost, K. (2016). Identification and monitoring of possible disruptive technologies by patent-development paths and topic modeling. *Technological Forecasting and Social Change, 104*, 16–29

Nelson, R. R. (1961). Uncertainty, learning, and the economics of parallel research and development efforts. *The Review of Economics and Statistics*, 351–364.

Noh, H., & Lee, S. (2020). Forecasting Forward Patent Citations: Comparison of Citation-Lag Distribution, Tobit Regression, and Deep Learning Approaches. IEEE Transactions on Engineering Management.

Oommen, T., Baise, L. G., & Vogel, R. M. (2011). Sampling bias and class imbalance in maximum-likelihood logistic regression. *Mathematical Geosciences, 43*(1), 99–120

Park, I., Park, G., Yoon, B., & Koh, S. (2016). Exploring promising technology in ICT sector using patent network and promising index based on patent information. *ETRI Journal, 38*(2), 405–415.

Park, I., & Yoon, B. (2018). Technological opportunity discovery for technological convergence based on the prediction of technology knowledge flow in a citation network. *Journal of Informetrics, 12*(4), 1199–1222

Pilkington, A., Lee, L. L., Chan, C. K., & Ramakrishna, S. (2009). Defining key inventors: A comparison of fuel cell and nanotechnology industries. *Technological Forecasting and Social Change, 76*(1), 118–127

Putnam, J. (1997). The value of international patent rights. Yale University, Ph.D. Thesis, pp. 2589–2589.

Reitzig, M. (2004). Improving patent valuations for management purposes—validating new indicators by analyzing application rationales. *Research policy, 33*(6–7), 939–957

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1135–1144.

Riccardi, G., & Hakkani-Tur, D. (2005). Active learning: Theory and applications to automatic speech recognition. *IEEE Transactions on Speech and Audio Processing, 13*(4), 504–511

Rosenberg, C., Hebert, M., & Schneiderman, H. (2005). Semi-Supervised Self-Training of Object Detection Models. WACV/MOTION, 2.

Scotchmer, S. (1991). Standing on the shoulders of giants: Cumulative research and the patent law. *Journal of Economic Perspectives, 5*(1), 29–41

Song, K., Kim, K., & Lee, S. (2018). Identifying promising technologies using patents: A retrospective feature analysis and a prospective needs analysis on outlier patents. *Technological Forecasting and Social Change, 128*, 118–132.

Squicciarini, M., Dernis, H., & Criscuolo, C. (2013). Measuring patent quality: Indicators of technological and economic value.

Su, H. N., Lee, P. C., Chen, C. M. L., & Chiu, C. H. (2012). Assessing the values of global patents. In 2012 Proceedings of PICMET'12: technology management for emerging technologies, pp. 966–974.

Suykens, J. A., & Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural Processing Letters, 9*(3), 293–300

Tanha, J., van Someren, M., & Afsarmanesh, H. (2011). Disagreement-based co-training. In 2011 IEEE 23rd international conference on tools with artificial intelligence, pp. 803–810.

Tanha, J., van Someren, M., & Afsarmanesh, H. (2017). Semi-supervised self-training for decision tree classifiers. *International Journal of Machine Learning and Cybernetics, 8*(1), 355–370

Tong, X., & Frame, J. D. (1994). Measuring national technological performance with patent claims data. *Research Policy, 23*(2), 133–141

Trajtenberg, M. (1990). A penny for your quotes: patent citations and the value of innovations. The Rand Journal of Economics, pp. 172–187.

Triguero, I., García, S., & Herrera, F. (2015). Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study. *Knowledge and Information systems, 42*(2), 245–284

Triguero, I., Sáez, J. A., Luengo, J., García, S., & Herrera, F. (2014). On the characterization of noise filters for self-training semi-supervised in nearest neighbor classification. *Neurocomputing, 132*, 30–41

Tuia, D., Pasolli, E., & Emery, W. J. (2011). Using active learning to adapt remote sensing image classifiers. *Remote Sensing of Environment, 115*(9), 2232–2242

Tuia, D., Ratle, F., Pacifici, F., Kanevski, M. F., & Emery, W. J. (2009). Active learning methods for remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing, 47*(7), 2218–2232

Verhoeven, D., Bakker, J., & Veugelers, R. (2016). Measuring technological novelty with patent-based indicators. *Research Policy, 45*(3), 707–723

Veryzer, R. W. (2005). The roles of marketing and industrial design in discontinuous new product development. *Journal of Product Innovation Management, 22*(1), 22–41

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In 33rd annual meeting of the association for computational linguistics, pp. 189–196.

Yoon, B., & Magee, C. L. (2018). Exploring technology opportunities by visualizing patent information based on generative topographic mapping and link prediction. *Technological Forecasting and Social Change, 132*, 105–117

Yoon, B., & Park, Y. (2004). A text-mining-based patent network: Analytical tool for high-technology trend. *The Journal of High Technology Management Research, 15*(1), 37–50

Yoon, B., Yoon, C., & Park, Y. (2002). On the development and application of a self–organizing feature map–based patent map. *R&D Management, 32*(4), 291–300

Zhang, L. (2011). Identifying key technologies in Saskatchewan, Canada: Evidence from patent information. *World Patent Information, 33*(4), 364–370

Zhu, X., Lafferty, J., & Rosenfeld, R. (2005). Semi-supervised learning with graphs (Doctoral dissertation, Carnegie Mellon University, language technologies institute, school of computer science).

Authors and Affiliations

Youngjae Choi[1] · Sanghyun Park[1] · Sungjoo Lee[1,2]

✉ Sungjoo Lee
sungjoo@ajou.ac.kr

Youngjae Choi
pphanho@ajou.ac.kr

Sanghyun Park
miyal42@ajou.ac.kr

[1] Department of Artificial Intelligence, Ajou University, 206 Worldcup-ro, Yeongtong-gu, Suwon-si, Gyeonggi-do, South Korea

[2] Department of Industrial Engineering, Ajou University, 206 Worldcup-ro, Yeongtong-gu, Suwon-si, Gyeonggi-do, South Korea