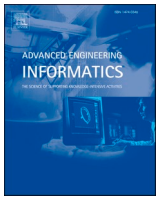
Advanced Engineering Informatics 62 (2024) 102944

Contents lists available at ScienceDirect

Advanced Engineering Informatics

journal homepage: www.elsevier.com/locate/aei



From technology opportunities to solutions generation via patent analysis: Application of machine learning-based link prediction

Ziliang Wang [a], Wei Guo [a,b], Hongyu Shao [a,*], Lei Wang [a,*], Zhixing Chang [a], Yuanrong Zhang [a], Zhenghong Liu [c]

[a] Key Laboratory of Mechanism Theory and Equipment Design of Ministry of Education, Tianjin University, Tianjin, China
[b] Tianjin Ren' ai College, Tianjin, China
[c] Guiyang University, Guiyang, China

ARTICLE INFO

Keywords:
Technology opportunity discovery
Technology convergence
Patent analysis
Machine learning
Link prediction

ABSTRACT

Technology convergence represents a significant mode of technological innovation that is widely prevalent across various industries. This innovative approach integrates multiple technologies to develop new integrated solutions, thereby fostering a competitive advantage for enterprises. Anticipating future potential technology convergence is of paramount importance for businesses. However, previous research has predominantly relied on the topological information of convergence networks, overlooking the nodal attributes and inter-nodal relationships that have an impact on the emergence of technology convergence. To enhance existing studies, this paper employs three types of features: node attributes and inter-node relationships based on the drivers of technology convergence, along with link prediction similarity indices. Additionally, we utilize Graph Convolutional Neural Network (GCN) for node embedding to leverage node attributes. Machine learning models are utilized for link prediction based on these features to identify potential technology opportunities. To guide research and development (R&D) efforts, we recommend high-value patents for each node using entropy weighting across five metrics that objectively quantify patent value, and transform patent abstracts into vectors using Doc2Vec. Patents with high similarity in abstract text between nodes are utilized to extract technical solutions and fuse ideas for technology convergence. A case study is conducted within the autonomous driving industry, leveraging comprehensive information including node attributes, inter-node relationships, and topology-based similarities to identify technology opportunities and guide the generation of R&D ideas through the convergence of technical solutions.

1. Introduction

With the significant acceleration of technological transformation, the rapid and effective capture of technology opportunities constitutes a crucial factor for enterprises to gain a competitive edge [1]. Technology opportunities represent intelligence that assists businesses in adapting to future market environments and technological changes [2]. The discovery of technology opportunities serves as the beginning of activities such as formulating R&D strategies and determining technological innovation directions [3]. Consequently, scholars have conducted extensive research to identify technology opportunities. To identify promising opportunities, predicting potential technology convergence has gained widespread attention, as new technologies often emerge through the convergence of multiple technologies [4], and technology convergence is recognized as a primary source driving innovation [5].

Technology convergence is defined as the phenomenon in which different technologies interact and transform into new integrated technologies, observed as a prevalent innovation pattern across various domains [6]. Many researchers highlight that technology convergence acts as a significant driver for corporate technological innovation and new product development [5]. Therefore, leveraging technology convergence to identify potential technology opportunities is crucial

* Corresponding authors at: Key Laboratory of Mechanism Theory and Equipment Design of Ministry of Education, Tianjin University, No.135 Yaguan Road, Haihe Education Park, Tianjin, 300350, PR China, Key Laboratory of Mechanism Theory and Equipment Design of Ministry of Education, Tianjin University, No.135 Yaguan Road, Haihe Education Park, Tianjin, 300350, PR China.
E-mail addresses: wzl07@tju.edu.cn (Z. Wang), wguo@tju.edu.cn (W. Guo), shaohongyu@tju.edu.cn (H. Shao), tjuwl@tju.edu.cn (L. Wang), Sean_key@tju.edu.cn (Z. Chang), zhang2052@tju.edu.cn (Y. Zhang), jx0011@gyu.edu.cn (Z. Liu).

https://doi.org/10.1016/j.aei.2024.102944
Received 23 June 2024; Received in revised form 1 November 2024; Accepted 8 November 2024
Available online 14 November 2024
1474-0346/© 2024 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

[7]. Enterprises anticipate technology convergence in advance and formulate R&D strategies to gain sustained competitive advantages.

Numerous studies have utilized link prediction for forecasting technology convergence. Specifically, link prediction based on patent data is increasingly employed, as patents serve as a robust representation of technological innovation, and patent data is regarded as a reliable data source [8]. Lee et al. [9] constructed a patent IPC co-occurrence network based on association rules and utilized the Adamic/Adar index to compute the similarity between IPC nodes for predicting technology convergence. Park and Yoon [10] established a Technological Knowledge Flow (TKF) network based on technological domain classification and patent citation information, employing link prediction similarity metrics to forecast potential technological links. Kim and Sohn [11] defined three metrics, including link prediction similarity metrics, bibliometrics, and patent semantic information. These metrics were concatenated into convergence vectors applied to machine learning models for predicting technology convergence, with new technology convergences represented by the initial combination of two IPCs. Lee et al. [12] applied Association Rule Mining (ARM) to construct a patent co-classification network, employing link prediction analysis based on a logistic regression model to predict multi-technology convergence patterns by identifying links that will be added to the network in the future.

Despite the extensive literature on predicting technology convergence, previous studies have limitations. The first limitation stems from the use of link prediction methods. Previous research has mostly utilized link prediction similarity methods to predict technology convergence [9,10], primarily relying on the network's topology information. However, whether two nodes will be linked depends not only on node similarity but also on the nodes' own attributes and the specific problem under study. Therefore, it is necessary to add attributes to nodes from the perspective of influencing technology convergence and to consider the combination of node attributes with network topology information. In addition, the technological network is a complex system with intricate connections and mutual influences among various technologies. Therefore, it is also necessary to consider the impact of relationships between technologies on the emergence of technology convergence. Second, in previous studies, predicted technology convergence was often defined as a combination of technology categories or patent classification codes [4,13]. However, such a representation of technology opportunities in the form of a combination of coarse-grained technological knowledge elements lacks a clear meaning. Therefore, current research does not provide detailed descriptions of technology opportunities, which are essential for generating research and development ideas and formulating R&D plans for engineers. Consequently, further research is needed to provide more detailed technology opportunity solutions.

Given the limitations of existing research, this study develops a machine learning-based link prediction method to forecast technology convergence and provides technological solutions for convergence opportunities. To address the first limitation, this paper analyzes the attributes of Cooperative Patent Classification (CPC) nodes and the relationships between node pairs from the perspective of technology convergence driving factors. CPC is a patent classification system developed in collaboration by the European Patent Office (EPO) and the United States Patent and Trademark Office (USPTO) in 2010. It is a system of codes used to classify and identify patents in different technological fields. By employing specific rules and symbol combinations, it categorizes numerous patents based on their technological themes, facilitating the organization, storage, retrieval, and management of patent literature. Furthermore, Graph Convolutional Networks (GCNs) are utilized for node embedding based on node attributes. As a graph representation learning method, Graph Neural Networks (GNNs) efficiently process non-Euclidean data by aggregating or diffusing messages from the neighborhood to generate an effective representation[14]. The problems, solutions, and innovations in the technological domain are encompassed within unstructured textual data such as patent abstracts and claims[15], and natural language processing techniques can extract technological information from large-scale patent text data[16,17], therefore, to address the second limitation, we utilize the entropy weight method to recommend high-value patents for CPC pairs based on five objectively quantified indicators of patent value. Additionally, we use Doc2Vec to analyze the similarity between high-value patents of two CPCs. The technological solutions for the convergence of two CPCs are obtained by extracting and integrating highly similar patent combinations.

The remainder of this paper is organized as follows. Section 2 reviews previous research on technology opportunity identification and technology convergence prediction. Section 3 provides a detailed description of the research methodology. Section 4 conducts empirical research, illustrating how our method can be applied in practice. Finally, Section 5 summarizes the contributions of the paper and discusses the limitations of the research.

2. Literature review

2.1. Technology opportunity discovery

Technology opportunities refer to the potential directions or possibilities of technological advancements [18]. Technology opportunity identification involves the process of mining potential technological advancements and innovations from knowledge databases through a series of technical means [19]. It is a branch of research within the field of technology forecasting, aiming to explore and assess the risks and opportunities of technological development [20]. Identifying promising technology opportunities can help companies understand the trends in technological development and reduce the risks associated with the technology innovation process. Effectively capturing technology opportunities in a fiercely competitive market environment has become a crucial issue affecting the survival of businesses [21].

Early methods for identifying technology opportunities include the Delphi method [22], Analytic Hierarchy Process [23], and others, which are mostly qualitative studies relying heavily on experts' knowledge and experience. Despite providing reliability support for the process of technology opportunity identification, experts' subjective biases remain a significant concern. Moreover, the exponential growth of data containing technological knowledge, such as journal articles and patents, makes it challenging for experts to identify technology opportunities solely based on their experience [24]. Therefore, quantitative analysis methods utilizing objective data have attracted significant attention from scholars. The emergence and development of technologies such as data mining, natural language processing, and machine learning have supported quantitative research. For instance, Yoon and Magee [3] utilized text mining to extract core keywords from patent data, developed a patent map for visualizing patents in a two-dimensional space, and employed support vector machines for link prediction to explore potential technology opportunities. Li et al. [25] proposed an approach integrating SAO semantic mining and outlier detection to identify technology opportunities in scientific papers and patents.

In recent years, some studies have focused on identifying potential opportunities based on technology convergence, as new technologies often emerge from the combination of multiple technologies. Kim et al. [26] employed clustering algorithms to categorize patents in citation networks into different technology groups and selected core patent combinations from these groups for convergence. Kim et al. [27] explored potential technology opportunities arising from technology convergence based on the presence of edge outliers in patent citation networks. Edge outliers were detected through centrality analysis, and the most promising convergence combinations were ultimately selected, along with their keywords, to propose directions for technology development. Lee and Sohn [28] designed a deep neural network to recommend convergence opportunities based on the technology portfolios and patent co-classification information of each company. Meanwhile,

literature on technology convergence has also increasingly utilized fusion prediction to discover technology opportunities [29].

2.2. Technology convergence forecasts for identifying technology opportunities

The concept of convergence was first introduced by N. Rosenberg, who proposed the notion of technology convergence as a social phenomenon in his study on the transformation of the U.S. machinery industry [30]. Technology convergence is typically regarded as the combination of two or more technologies and serves as an innovative mode for creating new technologies [31]. The emergence of new technology convergences may lead to the generation of fresh technology opportunities, thereby catalyzing technological change and innovation [32]. Companies can obtain a competitive edge in the market by predicting potential convergence opportunities, devising R&D strategies in advance, and preparing R&D investments.

Various methods have been developed to predict potential technology convergences [33], with most approaches utilizing link prediction techniques. In network analysis, link prediction is based on forecasting the association status of future networks based on an existing network structure [34]. Specifically, given a graph, link prediction anticipates potential links in the future based on the relationships among existing nodes. Existing methods for predicting technology convergences through link prediction can generally be classified into similarity-based approaches [9,35] and supervised learning-based approaches [10,12].

Patent-based link prediction has been widely employed in prior research to forecast technology convergences. Patents, as the primary output of technological research and development, can support activities aimed at exploring technological innovation and development trends [36]. This holds significant importance for enterprise managers in understanding relevant technological development directions and formulating policies [37]. Patent analysis is a process involving data collection and preprocessing, information mining, and knowledge discovery [38], which aids researchers in obtaining insights into technological development directions and assists enterprises in early research and development preparations [39,40].

Technology convergences in patent analysis are typically represented by the co-occurrence of patent classification codes (IPC, CPC) [9,41], or determined through citation relationships between patent classification codes based on patent citations [42,43]. In such cases, predicted technology convergences are presented in the form of combinations of classification codes. For instance, Wang et al. [36] constructed an IPC convergence network in the field of electric vehicles based on IPC co-occurrences and utilized the RA similarity index for link prediction analysis to identify potential convergences. Park and Geum [44] extracted node features, edge features, and link prediction similarity measures based on CPC co-occurrence information, employing machine learning models to forecast technology convergences. They also analyzed the prediction results from the perspective of enterprises to support decision-making at the organizational level. Cho et al. [45] built an IPC co-occurrence network using association rule mining and employed machine learning methods with various link prediction similarity indices to predict future technology convergences. For predicted IPC node pairs, they utilized an LDA topic model to identify key terms related to the technology represented by IPC and described technology convergences based on the convergence of keywords between IPCs. Wang et al. [13] proposed a machine learning-based link prediction method to identify convergence opportunities in the field of smart health. They constructed a Technical Knowledge Interaction (TKI) network based on CPC code co-occurrences and computed a set of 24 feature measures for each pair of nodes, including various link prediction similarity methods. PCA was used to extract representative features, which were then used to train machine learning models and predict potential links in future periods. Previous studies mostly relied on the topological information of networks. Additionally, most prior research represented predicted technology convergences in the form of IPC/CPC code combinations, where the meanings of classification codes are rather broad. Such simplistic combinations of knowledge elements may not generate specific R&D ideas for researchers.

Given the limitations of previous research, we propose a machine learning-based link prediction method to forecast technology convergences. This approach leverages both the network's topological information and node and node-pair information based on technology convergence issues. For identified technology opportunities, this paper further presents solutions to provide research and development guidance for researchers.

3. Methodology

The present study develops a machine learning-based link prediction method, which integrates various link prediction similarity indices, node attributes, and node pair relationships as features, to forecast potential technology convergence based on patent data. Furthermore, this research explores detailed technology opportunity solutions for use by technology R&D personnel. The research methodology consists of four stages, as depicted in Fig. 1. Firstly, patent data in the research field are collected and preprocessed. Secondly, appropriate features are selected for node pairs. These features for node pairs can form feature vectors for subsequent modeling. Thirdly, multiple machine learning models are employed to establish link prediction models based on the selected node pair features and node connectivity status, and by conducting experiments, the optimal prediction model is selected. Finally, the best prediction model is utilized to forecast convergences in future time periods. For newly predicted CPC pairs, high-value patents for each CPC are recommended, and Doc2Vec is utilized to analyze the similarity between two sets of high-value patents. By selecting patent combinations with high similarity, extracting and integrating technological solutions therein, support is provided for the generation of ideas for technology convergence. The overall process is divided into four stages: data collection and preprocessing, feature selection, establishment and evaluation of prediction models, and identification of technology opportunities and solutions, which will be elaborated on in subsequent sections.

3.1. Data collection and preprocessing

At this stage, we complet data collection and preprocessing. Patent data, being a significant source of technological innovation, are widely utilized in technology convergence prediction. Hence, we have chosen patent data as the primary data source for our study.

The data are collected from the Derwent Patent Database, including information such as patent titles, publication numbers, abstracts, CPC codes, patent assignees, and cited patents. Preprocessing is conducted on the CPC codes within the patents to analyze technological compositions, and a convergence network is constructed based on the co-occurrence information of CPC codes. CPC codes serve to indicate the technological fields involved in each patent [42], and if a patent is assigned multiple CPC codes, it implies that the patent is composed of the convergence of two or more technologies [46]. The co-occurrence network based on CPC codes is utilized to analyze the interaction relationships among technologies during the convergence process [47], where nodes are represented by CPC codes and the co-occurrence relationships of CPCs form the links in the network. The hierarchical structure of CPC codes can be divided into sections, classes, subclasses, groups, and subgroups from high to low levels. For instance, in a CPC such as G06T7/20, G represents the section, G06 represents the class, G06T represents the subclass, G06T7 represents the group, G06T7/20 represents the subgroup. Previous studies have indicated that CPC codes at the group level can precisely represent technological domains [13]. Therefore, in this study, a patent co-occurrence network was constructed using CPC codes at the group level, referred to as the convergence

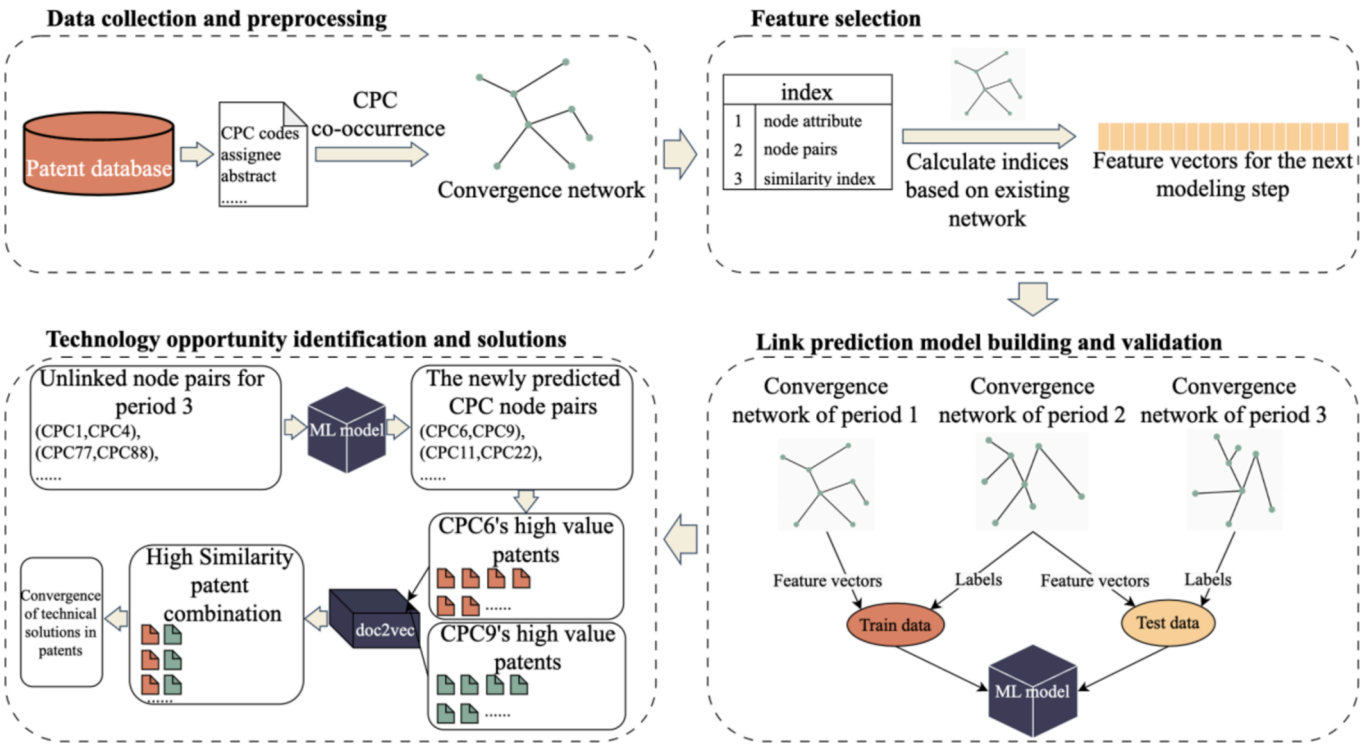


Fig. 1. Overview of the approach.

network.

3.2. Feature selection

Upon the completion of the construction of the convergence network, this phase involves selecting features for node pairs. Link prediction techniques based on network analysis are employed to forecast technology convergence, where the research question revolves around predicting whether links will be established between nodes based on various features of node pairs [13]. In this study, a machine learning-based link prediction approach is adopted. Learning-based methods allow for the construction of classification models using various link prediction similarity indices or other metrics as features [48]. The convergence of two technologies depends not only on the similarity of technology nodes but also on the attributes of nodes and the relationships between node pairs specific to the problem at hand. Therefore, in this phase, we integrated node attributes, inter-node relationships and link prediction similarity indices as node features. Node attributes and node pair relationships are extracted from indicators related to technology convergence driving factors. The collection of indicators for a pair of nodes can form a feature vector used for subsequent modeling.

3.2.1. Feature selection based on technology convergence drivers

Table 1 presents the node attributes and node pair relationship indicators related to technology convergence driving factors. Following Kim and Sohn [11], the rapid expansion of technological domains leads to convergence with other technological domains. Therefore, the growth rate of patents in a specific CPC node is measured using the growth rate (GR). Specifically, it is equal to the linear regression slope of the patent quantity change trend within a given time period. TN and NP measure the total number of patents and the number of patents in the most recent year, respectively, for each CPC node. Larger technological domains are more likely to experience new technology convergence [49]. Eilers et al. suggest that an increase in semantic similarity values between different technological domains indicates convergence between the domains [50]. Therefore, EV is defined as the embedding vector of the definition for each CPC node and is a 20-dimensional vector based on Doc2Vec embedding. For example, the definition of B60W60 is "Drive control systems specially adapted for autonomous road vehicles." We treat each CPC definition as a text segment and train Doc2Vec to obtain their vector representations.

Table 1
Indicators of driving factors affecting technology convergence.

Indicator	Definition	level
GR	Growth rate of CPC nodes in the network	node
TN	Total number of patents for CPC node	node
NP	Number of patents for CPC node in the past year	node
EV	Doc2Vec embedding vector for node definition	node
CS	Co-occurrence of CPC node pair at the subclass level	link
TS	Technical similarity of CPC node pair	link
NA	Number of same assignee of node pair	link

The node-pair relationship indicators measure the impact of the interaction relationship between nodes on technology convergence. CS represents the co-occurrence of CPC node pairs at the subclass level. The higher the degree of interaction between two technological domains, the more likely they are to experience technology convergence in the future [49]. The calculation formula is:

$$CS = (P(sub(CPC1) \cap P(sub(CPC2))/n \quad (1)$$

Where sub(CPC1) represents the subclass level of CPC1. For example, the subclass level of B60W60 is B60W. P(CPC1) represents the patent set containing CPC1, and n is the total number of patents included in this period.

According to the study by Kose and Sakata [51], technological similarity is a crucial measure for technology convergence; thus, Technological Similarity (TS) is employed to measure the technological similarity between pairs of CPC nodes. The CPC hierarchical structure itself is a logically rigorous tree structure. The CPC codes from sections to subgroups are obtained by gradually subdividing the field. Therefore, the closer the distance in the hierarchical structure of classification

codes, the more similar they are [52]. Therefore, technological similarity can be defined through the hierarchical structure of classification codes. The calculation formula is:

$$TS = 1 \times s + 1 \times s \times c + 1 \times s \times c \times sc + 1 \times s \times c \times sc \times g \quad (2)$$

Where s, c, sc, and g represent whether the remaining characters after removing the respective upper-level characters of CPC1 and CPC2 at the section, class, subclass, and group levels are the same. The value is 1 if they are the same and 0 if they are different. The value is 1 if they are the same and 0 if they are different. For example, if the sections of two CPCs are the same but the classes are different, then this indicator is recorded as 1. If the sections are the same and at the same time the classes are also the same but the subclasses are different, then this indicator is recorded as 2, and so on.

Lastly, Node Assignee (NA) quantifies the number of identical assignees at two nodes. The calculation formula is:

$$NA = ap(CPC1) \cap ap(CPC2) \quad (3)$$

Where ap(CPC1) represents the assignee set of patents containing CPC1.

Link prediction pertains to edges within a network, whereas CS, TS, and NA denote relationships between node pairs, which can be directly utilized as features for node pairs. However, GR, TN, NP, and EV are node attributes that need to be restructured at the node pair level. To simultaneously consider node attributes and network topology, we employ Graph Convolutional Network (GCN)-based graph representation learning for node embedding.

Graph representation learning is an efficient method for extracting graph structure information and capturing complicated interactions between nodes [53]. Given a graph G=(V,E),where V(|V|=n) and E are the set of nodes and edges respectively,let A be the adjacency matrix and X be the identity matrix,the layer-to-layer propagation of GCN is [54]:

$$H^{(l+1)} = \sigma(\widetilde{D}^{-\frac{1}{2}}\widetilde{A}\widetilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}) \quad (4)$$

where $\widetilde{A} = A + I_n$, I_n is the unit matrix. $\widetilde{D}_{ii} = \sum_j \widetilde{A}_{ij}$. $H^{(l)}$ is the activation matrix of layer l,and $H^{(0)} = X$. $W^{(l)}$ is the layer-specific trainable weight matrix. σ() denotes the activation function. The feature matrix X is made up of vectors stitched together by GR,TN,TP, and EV for each node. The graph's adjacency matrix A and feature matrix X are utilized as inputs, and the binary cross entropy is employed as a loss function to train the GCN model to generate the node embedding vectors, as shown in Fig. 2. To extract node pair-level characteristics, the Hadamard product operation is used to the two nodes' embedding vectors.

3.2.2. Feature selection based on link prediction similarity methods

Similarity-based methods for link prediction primarily focus on the topological structure of the graph. These approaches utilize structural properties of the graph to assign similarity scores to node pairs [55], which can be used to measure the likelihood of convergence between two unconnected CPC nodes. Similarity-based methods can be categorized into local, quasi-local, and global methods. Local similarity-based methods assume that nodes with similar neighborhood structures may form edges in the future [55]. Local similarity methods only use structural information related to the neighborhood to compute node similarity, making them faster than global similarity-based methods. Global similarity-based methods use the entire network's topology to measure similarity between nodes [56], resulting in higher algorithmic time complexity. Quasi-local similarity-based methods are introduced to strike a balance between local and global methods [56], exhibiting nearly the same computational efficiency as local methods. Some quasi-local methods utilize the entire network's topology, yet their algorithmic time complexity remains lower than that of global methods. Various similarity indices have been employed in the literature; however, the effectiveness of these indices depends on the specific problem and the structural characteristics of the network [48]. Hence, through the literature review of Mutlu et.al [56], we extensively select different categories of similarity indices as node pair features, with the selected similarity indices outlined in Table 2.

where v_x and v_y denote two distinct nodes; $\Gamma(v_x)$ denotes the set of neighboring nodes of node v_x; L + is the pseudo-inverse of the Laplace matrix of the network; A is the adjacency matrix, I is the unitary matrix of the appropriate dimensions, $\beta<1/\lambda_1$, where$\lambda_1$is the maximum eigenvalue of A; γ is a free parameter; consider q_{xy} to be the probability that a random walker who starts walking from vertex x and located at the vertex y in steady-state.

3.3. Link prediction model building and validation

After selecting the node pair features, this stage entails describing how to build and evaluate the link prediction model. Learning-based methods view the link prediction problem as a binary classification task. Four machine learning models, namely Random Forest (RF), Support Vector Machine (SVM), Extreme Gradient Boosting (XGB), and Light Gradient Boosting Machine (LGBM), are employed to learn from the node pair features computed from the existing convergence network.

To conduct link prediction, we partition the collected data into three time periods and construct a convergence network for each period accordingly. This study aims to predict technologies that are currently unlinked but may become associated in the future. Therefore, we focus

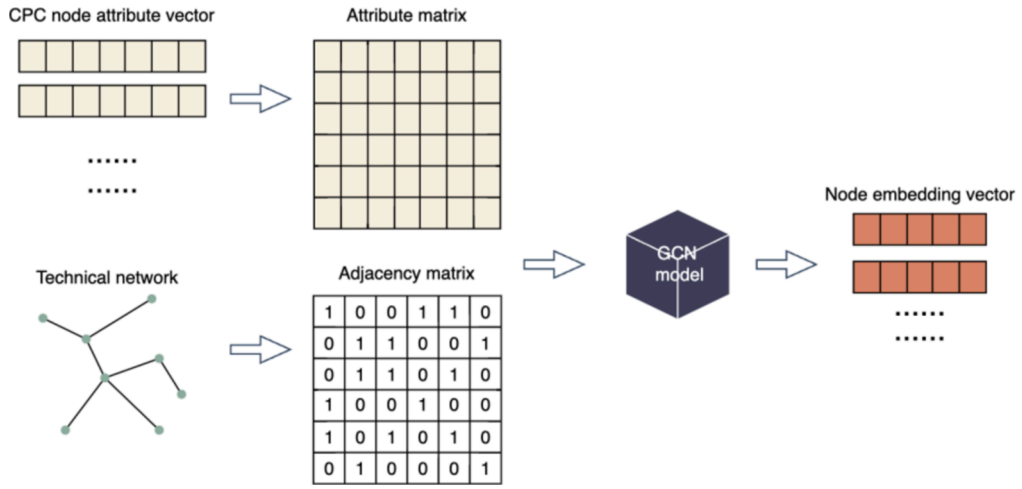


Fig. 2. GCN node embedding.

Table 2
Link prediction similarity indices.

Group	Index	Explanation
Local similarity	Adamic-Adar Index(AA)	$s^{AA}_{(v_x,v_y)} = \sum_{v_z \in (\Gamma(v_x) \cap \Gamma(v_y))} \frac{1}{\log\lvert\Gamma(v_z)\rvert}$
	Common Neighbors(CN)	$s^{CN}_{(v_x,v_y)} = \lvert\Gamma(v_x) \cap (v_y)\rvert$
	Hub Depressed Index (HD)	$s^{HD}_{(v_x,v_y)} = \frac{\lvert\Gamma(v_x) \cap \Gamma(v_y)\rvert}{\max(\lvert\Gamma(v_x)\rvert, \lvert\Gamma(v_y)\rvert)}$
	Hub Promoted Index (HP)	$s^{HP}_{(v_x,v_y)} = \frac{\lvert\Gamma(v_x) \cap \Gamma(v_y)\rvert}{\min(\lvert\Gamma(v_x)\rvert, \lvert\Gamma(v_y)\rvert)}$
	Jaccard Index(JC)	$s^{JC}_{(v_x,v_y)} = \frac{\lvert\Gamma(v_x) \cap \Gamma(v_y)\rvert}{\lvert\Gamma(v_x) \cup \Gamma(v_y)\rvert}$
	Leicht-Holme-Newman Index(LHN)	$s^{LHN}_{(v_x,v_y)} = \frac{\lvert\Gamma(v_x) \cap \Gamma(v_y)\rvert}{\lvert\Gamma(v_x)\rvert.\lvert\Gamma(v_y)\rvert}$
	Preferential Attachment (PA)	$s^{PA}_{(v_x,v_y)} = \lvert\Gamma(v_x)\rvert.\lvert\Gamma(v_y)\rvert$
	Resource Allocation Index (RA)	$s^{RA}_{(v_x,v_y)} = \sum_{v_z \in (\Gamma(v_x) \cap \Gamma(v_y))} \frac{1}{\lvert\Gamma(v_z)\rvert}$
	Sørensen Index (SI)	$s^{SI}_{(v_x,v_y)} = \frac{\lvert\Gamma(v_x) \cap \Gamma(v_y)\rvert}{\lvert\Gamma(v_x)\rvert + \lvert\Gamma(v_y)\rvert}$
	Salton index(SL)	$s^{SL}_{(v_x,v_y)} = \frac{\lvert\Gamma(v_x) \cap \Gamma(v_y)\rvert}{\sqrt{\lvert\Gamma(v_x)\rvert.\lvert\Gamma(v_y)\rvert}}$
Global similarity	Average Commute Time (ACT)	$s^{ACT}_{(v_x,v_y)} = \frac{1}{l^+_{xx} + l^+_{yy} - 2l^+_{xy}}$
	Cosine Similarity on L+(Cos +)	$s^{Cos+}_{(v_x,v_y)} = \frac{l^+_{xy}}{\sqrt{l^+_{xx} l^+_{yy}}}$
	Katz Index(KI)	$S = (I - \beta A)^{-1} - I$
	Random Walk with Restart (RWR)	$s^{RWR}_{(v_x,v_y)} = q_{xy} + q_{yx}$
Quasi-local similarity	Local Path Index	$S^{LP} = A^2 + \gamma A^3$

on node pairs that are unlinked in one time period's convergence network but form links in the subsequent period. Specifically, the set of node pairs unlinked in time period 1 but linked in time period 2 is chosen as the positive samples (labeled as 1) for the training set. The technology convergence prediction problem often involves an imbalanced dataset. In our study, the number of links (labeled as 1) in the convergence network is significantly lower than the number of unlinked pairs (labeled as 0). Hence, we employ negative sampling. The negative samples (label 0) of the training set consist of pairs of nodes in time period 1 that do not have links in time period 2. And the positive and negative samples are identical in number. The positive samples of the test set consist of node pairs that are unlinked in time period 2 but linked in time period 3, and the negative samples are node pairs in time period 2 that do not have links in time period 3. The negative samples of the test set are also randomly selected based on the number of positive samples.

To evaluate the model, five performance evaluation metrics are used to measure the model prediction effectiveness. Accuracy is the percentage of correctly predicted outcomes out of the total samples. Recall measures the probability that a sample that is actually positive is predicted positive. Precision measures the probability of correctly predicting an outcome out of the samples that are predicted to be positive. The F1 score is the reconciled mean of precision and recall.The Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve indicates the predictive performance, with a larger area corresponding to better prediction effectiveness.

3.4. Technology opportunity identification and solution analysis

At this stage, we apply the best predictive model to all unlinked pairs of CPC codes in period 3 to forecast potential future links. For the predicted CPC pairs, the technologies indicated by the two CPC codes may interact in the future, resulting in a new technology applicable to the research domain. Researchers can consider the predicted technology convergence as potential opportunities for future technological advancements.

However, representing a technology opportunity with a combination of CPC codes may not provide specific guidance for technology developers, as the meanings of CPC codes are broad and cannot offer concrete technology convergence solutions. Cho et al. [45] conducted LDA topic modeling on predicted IPC combinations, considering the convergnece of keywords derived from each IPC as a technology opportunity. The Latent Dirichlet Allocation (LDA) topic model is a probabilistic approach for generating topic models. It posits that each document is generated by mixing multiple topics, with each topic in turn being composed of a mixture of several words [57]. However, these fragmented keyword combinations still provide a coarse-grained description and do not specify how two IPCs can be converged into a new technology.

Therefore, this paper explores the generation of technology opportunity solutions in three steps. Firstly, We recommend high-value patents for each CPC. High-value patents often possess higher levels of technical solutions and greater technological innovation compared to other patents. We selected five objectively quantifiable indicators of patent value based on the literature review by Trappey et al. [58], as shown in Table 3. For CPCs, previous studies have demonstrated a positive correlation between the number of CPC subclasses assigned to each patent and its value [59]. The number of patent inventors indirectly represents the investment of the R&D company, thus positively influencing the patent's value [60]. The number of claims represents specific requirements for the scope of protection granted to the patent, and patents with a higher number of claims typically indicate more complex and innovative technologies, thus receiving broader legal protection. Such patents often have higher value. It is widely accepted in previous research that forward citations of patents are positively correlated with patent value, while backward citations are negatively correlated with patent value [61]. Subsequently, we use the entropy weight method to comprehensively evaluate the value of patents. Patents with high scores are recommended as high-value patents. Secondly, we use Doc2Vec to generate representation vectors for the abstracts of each patent. Doc2Vec is extended from Word2Vec [62]. It is an unsupervised text representation method based on the idea of Word2Vec. The Doc2Vec algorithm is an unsupervised text representation method based on the Word2Vec concept, aiming to represent documents of different lengths as fixed-length dense vectors. Doc2Vec has two training models, PV-DM, and PV-DBOW, and we use the PV-DM model for training. Thirdly, we calculate the textual similarity of patents between two CPCs based on cosine similarity. For patents with high similarity, we extract relevant technological solutions related to both CPCs and combine them to form convergence technology opportunity solutions. These convergence technology opportunity solutions provide R&D personnel with ideas for technology convergence, effectively guiding research and development.

4. Results and discussion

Autonomous driving, also known as driverless technology, refers to the capability of vehicles to perceive and navigate through their environment and reach destinations smoothly without driver intervention.

Table 3
Definition of patent value indicators.

Indicator	Definition
CPCs	Number of group level CPCs
Inventors	Number of patent inventors
Claims	Number of patent claims
Forward citations	Citations received from other patents under application
Backward citations	Number of other patents cited

Autonomous driving technology is not simply divided into two categories of "presence" and "absence", but rather can be classified into multiple levels based on the degree of intelligence. The international authoritative automotive standardization organization SAE categorizes autonomous driving technology into six levels, ranging from Level 0, which relies entirely on driver control, to Level 5, which represents fully autonomous driving, with each level representing different stages of technological development. Currently, autonomous driving technology has been diversifiedly applied in various fields such as public transportation, taxis, logistics and delivery, and urban infrastructure.According to McKinsey [63], by 2035, autonomous driving could generate revenue ranging from 300 billion to 400 billion. The most conservative estimate suggests that by 2030, the penetration rate of Level 3 and above autonomous driving vehicles globally will be around 4 %, increasing to 17 % by 2035; the most optimistic estimate indicates that by 2030, the penetration rate of Level 3 and above autonomous driving vehicles globally will reach 20 %, increasing to 57 % by 2035. Overall, autonomous driving technology has vast development prospects and will bring about safer, more efficient, and convenient travel options for people. Therefore, we chose the field of autonomous driving to demonstrate the effectiveness of the proposed method and identify potential technological opportunities in this field for the future.

4.1. Data collection and processing

This paper collected patent data in the field of autonomous driving based on the Derwent Patent Database. Patent classification code retrieval is suitable for retrieving patent data in well-defined specific technical areas [14]. However, for emerging technologies that have not yet been classified, this method is not feasible [64]. In such cases, keyword-based retrieval is more appropriate. In this study, we utilized the Cooperative Patent Classification (CPC) jointly developed by the European Patent Office (EPO) and the United States Patent and Trademark Office (USPTO), along with keywords related to autonomous driving, for patent retrieval. We referenced research on autonomous driving [65] and combined it with the search strategy outlined in the World Intellectual Property Report 2019 (Note: Readers can access the following website for details: https://www.wipo.int/publications/en/details.jsp?id = 4467) to determine the search strategy for this study (see Appendix A for details).

After removing duplicate patents, a total of 87,320 patents from the years 2014 to 2022 were collected. We extracted group-level Cooperative Patent Classification (CPC) codes and displayed the top 10 categories in Table 4. It is noteworthy that the top 10 CPC codes are mostly related to vehicle control and propulsion, indicating that the core of autonomous driving technology lies in achieving precise control over vehicle behavior. Therefore, technologies related to vehicle control systems (such as G05D1, B60W30, B60W50, B60W60, etc.) play a crucial role in the field of autonomous driving. Additionally, navigation and image processing-related technologies also hold significant importance, indicating that accurate perception and understanding of the driving environment are also critical. CPC H04W4 suggests that wireless communication networks play an essential role in autonomous driving technology. Through wireless communication technology, autonomous vehicles can communicate in real-time with other vehicles, infrastructure, and the cloud, thereby acquiring more information and making more accurate decisions.

Table 4
Technologies composition of the autonomous driving.

CPC code	CPC Definition	Percentage
G05D1	Control of position, course, altitude or attitude of land, water, air or space vehicles	9.4 %
G08G1	Traffic control systems for road vehicles	7.0 %
B60W30	Purposes of road vehicle drive control systems not related to the control of a particular sub-unit, e.g. of systems using conjoint control of vehicle subunits	5.2 %
H04W4	Services specially adapted for wireless communication networks; Facilities therefor	4.1 %
B60W60	Drive control systems specially adapted for autonomous road vehicles	2.8 %
G06T2207	Indexing scheme for image analysis or image enhancement	2.7 %
G01C21	Navigation; Navigational instruments not provided for in groups	2.6 %
B60W50	Details of control systems for road vehicle drive control not related to the control of a particular sub-unit	2.5 %
B60W2554	Input parameters relating to objects	2.5 %
B60W40	Estimation or calculation of driving parameters for road vehicle drive control systems not related to the control of a particular sub unit	2.3 %

Finally, by considering assigning similar network densities and numbers of nodes for each period, we divided the data into three periods: 2014–2018, 2019–2020, and 2021–2022, and constructed convergence networks for each period. During the period from 2014 to 2018, there were 30,322 patents; from 2019 to 2020, there were 27,924 patents; and from 2021 to 2022, there were 29,073 patents.

4.2. Feature selection

The selected node pair features are shown in Table 5. CS, TS, and NA are calculated respectively according to formulas (1), (2), and (3) in Section 3.2.1.. The 15 link prediction similarity indices can be directly computed based on the existing network. However, GCN (GR, TN, TP, EV) is computed in three steps. Firstly, GR, TN, TP, and EV are computed for each node based on the existing network. Then, the GR, TN, TP, and EV for each node are concatenated into a 23-dimensional vector, where GR, TN, and TP are each 1-dimensional, and EV is a 20-dimensional embedding vector based on Doc2Vec. We employed a two-layer GCN, wherein node cosine similarity is calculated, and binary cross-entropy loss is used to train the GCN model. Each node is represented as a 6-dimensional embedding vector. Finally, the Hadamard product operation is used to reconstruct the representation vectors of two nodes into a node-pair level.

The three types of features are concatenated into a feature vector. Ultimately, each node pair is represented as a 24-dimensional vector, which is used for subsequent modeling.

4.3. Link prediction model building and validation

In the learning-based link prediction method, the technology convergence prediction problem is defined as a binary classification problem, where label 1 indicates technology convergence and label 0 indicates non-technology convergence. Patent data from 2014 to 2018 is used to extract the node pair features required for the training set, while the node connection status from 2019 to 2020 serves as the labels (0 or 1) for the training set. Features in the test set are computed from the convergence network from 2019 to 2020, and the labels are derived from patent data from 2021 to 2022.

High precision implies a high proportion of true positive instances among the predicted positive samples, meaning fewer false alarms (false positives), but it may lead to missing some potential opportunities. High recall means the model can identify most of the true positive instances, resulting in fewer missed detections (false negatives). However, this can

Table 5
Final features.

	node attribute	node pair relationship	similarity index
Feature	GCN(GR, TN, TP, EV)	CS, TS, NA	AA, CN, HD, HP, JC, LHN, PA, RA, SI, SL, ACT, Cos+, KI, RWR, LP
Dimension	6	3	15

result in more incorrect predictions of technology convergence, thereby increasing the R&D burden for companies. Therefore, companies need to consider acceptable precision and recall rates based on their R&D strategies, R&D costs, and benefits.

We compared our proposed method with previous approaches, including AA [9], RA [36], as well as combinations of AA, PA, and RA [35]. To mitigate the impact of inappropriate threshold selection, machine learning classifiers were employed for all the comparative methods. As shown in Table 6, our proposed method exhibits significant improvements in terms of accuracy and AUC compared to the previously proposed methods. Furthermore, we also compared the performance of the model when using only network topology-based similarity indices (including AA, CN, HD, HP, JC, LHN, PA, RA, SI, SL, ACT, Cos+, KI, RWR, LP) as features, and when using the proposed comprehensive set of indicators (GCN node attributes, inter-node relationships, and similarity indices) as features, as shown in Table 7. The RF and SVM models exhibit superior performance across all five evaluation metrics when using all features. Although XGB and LGBM show higher recall when using only similarity indices, their precision, F1 score, and AUC are lower compared to using all features. This indicates that the latter achieves a better balance between prediction accuracy and recall overall, demonstrating superior performance and generalization capability. Furthermore, the RF model achieves a better balance between precision and recall, with optimal values for accuracy and F1 score, and also demonstrates decent performance in terms of AUC. Therefore, we opt for the RF model for subsequent analysis.

Fig. 3 illustrates the feature importance of the RF model. The random forest model determines the importance of features based on the improvement of model performance by features. As observed, the link prediction similarity method exhibits notably high feature importance. Specifically, among the top 6 features, RWR is a global similarity method, RA, PA, AA, and CN represent local similarity methods, and LP is a quasi-local similarity method. This suggests the crucial importance of the network's topology in predicting future associations within the convergence network. Furthermore, CS is also deemed a significant feature, indicating the degree of interaction between subclasses belonging to a pair of CPCs. The degree of association between two domains significantly impacts the convergence of technologies across domains. While the importance of node attributes based on GCN ranks relatively lower overall, it still surpasses similarity indices such as JC, SI and HD. Hence, in predicting technology convergence, it is necessary to select appropriate node attributes based on technology convergence influencing factors.

4.4. Technology convergence forecasting and solution analysis

The selected RF model is utilized to predict future technology convergence based on 2,029,595 unlinked CPC pairs from 2021 to 2022. A total of 275,485 new CPC pairs are predicted. As evident from the recall rate of the RF model in the previous section, these new CPC pairs encompass the majority (84.1 %) of new links in the future timeframe. Hence, the set composed of newly predicted CPC pairs can be interpreted as encapsulating most of the convergence opportunities in a significantly smaller set than the original one (2,029,595). The new CPC pairs that have been judged and filtered by experts will be used to support the research and development of new technologies in enterprises.

Fig. 4 depicts the top 10 newly added CPC links at the subclass level in the prediction results. Further analysis will focus on B60W, which represents the subclass with the highest predicted number of links.

There are 23 subgroups under CPC B60W with newly added links. We specifically focus on B60W60 (Drive control systems specially adapted for autonomous road vehicles). There are 1011 new links added for B60W60. Furthermore, we choose the New convergence B60W60 and H04W74(Wireless channel access) to illustrate how the technology opportunities solution can be accessed.

To obtain technology opportunity solutions, the following steps are undertaken. Firstly, patents between 2021 and 2022 that contain B60W60 and H04W74 are selected, for each patent, subclass-level CPCs, inventors, claims, and the number of forward and backward citations are extracted. Using the entropy method, scores are computed for each patent based on these five indicators to identify high-value patents. Table 8 shows the high-value patent recommendations for CPCs. Secondly, abstracts of high-value patents are preprocessed through tokenization and stop-word removal, then transformed into vector representations using Doc2Vec. The similarity of text vectors from different CPCs is calculated using cosine similarity, and patent combinations with high similarity values are used to obtain solutions for convergence opportunities. Table 9 outlines the top 10 patent combinations.

We take the combination of CN113753077A and CN112969141A to illustrate the process of generating convergence technology opportunity solutions from highly similar patent combinations, as shown in Fig. 5. Patent CN113753077A is an invention for predicting the trajectory of an obstacle by means of electronic devices. By acquiring environmental information in the target scenario (including road data, traffic signal data, obstacle data, etc.), historical status information of target obstacles, and planned trajectory information of target vehicles, a neural network model is used to predict the motion trajectory of target obstacles. This enhances the prediction accuracy of obstacle motion trajectories for autonomous driving vehicles in interactive scenarios, thereby improving driving safety and smoothness.Patent CN112969141A introduces a method for implementing vehicle radar integrated non-orthogonal multiple access random communication. Intelligent vehicles determine communication resource occupancy information and real-time road conditions based on onboard radar perception results. This information is input into a Markov decision model to determine action decisions, including resource allocation. Real-time road condition information is sent to the target intelligent vehicle using non-orthogonal multiple access random access, and data validity is evaluated to determine transmission priority.

We can integrate the technical solutions related to autonomous driving control (B60W60) and wireless channel access (H04W74) from these two patents. Patent CN113753077A utilizes electronic maps to obtain environmental information (such as road data and traffic light data), historical information of obstacles (such as historical speed and historical position), and is used to predict the movement trajectory of obstacles. However, the real-time information provided by electronic maps is often more generalized and lacks specific dynamic interaction information among vehicles within particular areas. Moreover, electronic maps may not promptly reflect sudden vehicle breakdowns or unusual road conditions (such as water accumulation or landslides) encountered by vehicles ahead on their route. The wireless

Table 6
Model performance comparison.

Model	Accuracy				AUC			
	All features	AA only	RA only	AA + RA + PA	All features	AA only	RA only	AA + RA + PA
RF	**0.854**	0.762	0.783	0.822	0.921	0.826	0.843	0.885
SVM	0.793	0.810	0.831	0.807	0.861	0.871	0.881	0.867
XGB	0.845	0.814	0.827	0.832	0.917	0.887	0.896	0.900
LGBM	0.851	0.813	0.828	0.834	**0.924**	0.887	0.897	0.903

Table 7
Model performance.

Model	Accuracy		Precision		Recall		F1 score		AUC	
	Similarity only	All features	Similarity only	All features	Similarity only	All features	Similarity only	All features	Similarity only	All features
RF	0.843	**0.854**	0.841	**0.858**	0.845	**0.848**	0.843	**0.853**	0.911	0.921
SVM	0.782	0.793	0.832	0.834	0.728	0.732	0.777	0.780	0.857	0.861
XGB	0.841	0.845	0.840	0.857	0.842	0.825	0.841	0.841	0.914	0.917
LGBM	0.847	0.851	0.844	0.858	0.851	0.841	0.847	0.849	0.918	**0.924**

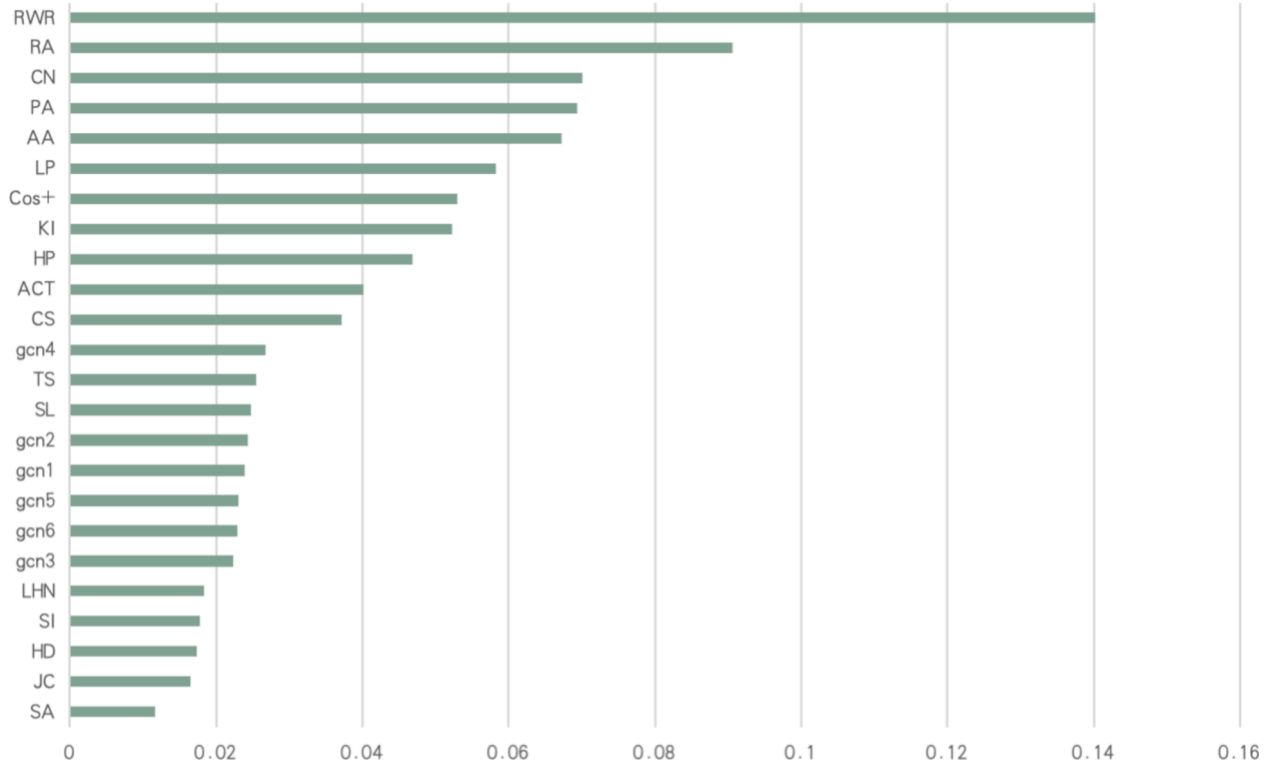


Fig. 3. The importance of each feature in random forest.

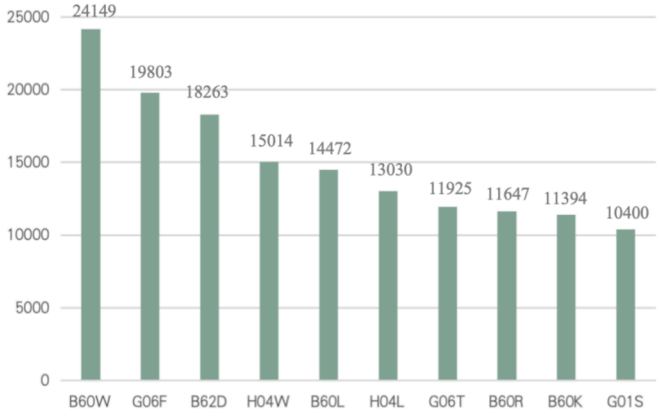


Fig. 4. The top-10 number of CPC new link added at the class level.

communication access method invented by patent CN112969141A facilitates the sharing of motion state information and real-time traffic information between intelligent vehicles, enabling information exchange and cooperative driving among them. For instance, at complex traffic intersections, vehicle-to-vehicle communication allows mutual understanding of each other's driving intentions and dynamics. Therefore, for localized areas, the wireless communication access method proposed in patent CN112969141A offers more precise information compared to electronic maps. Hence, we propose to complement the information retrieval from electronic maps with the wireless

Table 8
High-value patents of B60W60 and H04W74.

CPC	High-value patents
B60W60	US20210253128A1, US20210223051A1 CN113386795A, US11249506B1 US11119219B1, CN112373472A US20220084340A1, US11225226B1 CN113753077A, US20210182604A1 US20220297635A1, US11327520B1 US20210261159A1, US20220180419A1 CN112849055A, CN113246985A US20220055661A1, EP3943970A1 US20210261163A1, US20220092673A1
H04W74	US20210385865A1, WO2021234164A2 US20230269766A1, CN114189824A CN113783599A, WO2022093499A2 CN116233762A, WO2022080914A1 CN115268974A, US20210266919A1 US20240035831A1, WO2022003031A1 CN114615645A, WO2022149629A1 KR2023087399A, US20210329501A1 CN112969141A, WO2023089227A1 WO2022086182A1, GB2602812A

Table 9
Top 10 patent combinations ranked by similarity.

Patent combination	Similarity score
(US20220180419A1, WO2021234164A2)	0.323
(CN113753077A, CN112969141A)	0.27
(US20220297635A1, WO2022149629A1)	0.267
(CN113753077A, WO2023089227A1)	0.251
(CN113246985A, WO2022149629A1)	0.25
(US20220084340A1, CN112969141A)	0.248
(EP3943970A1, US20230269766A1)	0.236
(US20210253128A1, US20240035831A1)	0.23
(CN113386795A, US20230269766A1)	0.229
(CN112849055AGB2602812A)	0.227

communication access method from patent CN112969141A, enabling intelligent vehicles to acquire more comprehensive and accurate information to support autonomous driving control. In the convergence technology opportunity solution, the electronic map provides long-distance planning and navigation for autonomous vehicles and plans the optimal path through statistical information, such as vehicle flow and traffic accident occurrence rate at different times. At the same time, the vehicle receives real-time movement information and road condition information sent by vehicles within a certain range through wireless channel access and transmits its own status information and surrounding information to other intelligent vehicles. Integrate these pieces of information to provide a comprehensive and accurate data basis for autonomous driving control. This convergence technology opportunity provides researchers with a concrete demonstration of how to integrate these two technologies, allowing for technological innovation and inventive creations.

5. Conclusions and limitations

5.1. Conclusions

This study presents a novel systemic approach, showcasing the process from identifying technology opportunities to generating technology opportunity solutions through the integration of machine learning-based link prediction and analysis of technological solutions in patents. For this method, we initially collect patents in the studied field based on a predefined search strategy, analyze CPC compositions and co-occurrence information, and construct a convergence network. Subsequently, we select a comprehensive set of indicators for building machine learning models based on technology convergence driving factors and link prediction similarity metrics. Next, we focus on node pairs that unlink in the current period but link in the subsequent period, and based on this data, various machine learning models are trained and tested. Finally, we utilize the optimal predictive model to forecast technology convergnece, employing high-value patent analysis and patent text similarity calculations to select suitable patent combinations. We extract technological solutions related to the predicted CPCs from these combinations as convergnec technology opportunity solutions, aiding in the generation of convergence ideas and the development of convergnec technologies.

We believe our research has a positive impact on both academia and industry. From a theoretical perspective, this study enhances the information within the network by incorporating more comprehensive details such as node attributes and relationships between nodes based on the research question of technology convergence prediction. This research addresses the limitations of previous methods, which primarily focused on network topology and lacked the selection of appropriate features to effectively predict future links. Moreover, previous approaches only utilized CPC pairs or keywords as scattered features to define technology opportunities. The proposed method, by obtaining convergence ideas through the convergence of technological solutions in patents, addresses the ambiguity in the meaning of technology opportunities in existing research.

Additionally, from a practical standpoint, identifying potential technology opportunities can serve as a reference for R&D engineers, as these opportunities are likely to be proposed as inventions in the future, thus providing researchers with more innovation prospects. The method proposed in this study can effectively identify potential convergence

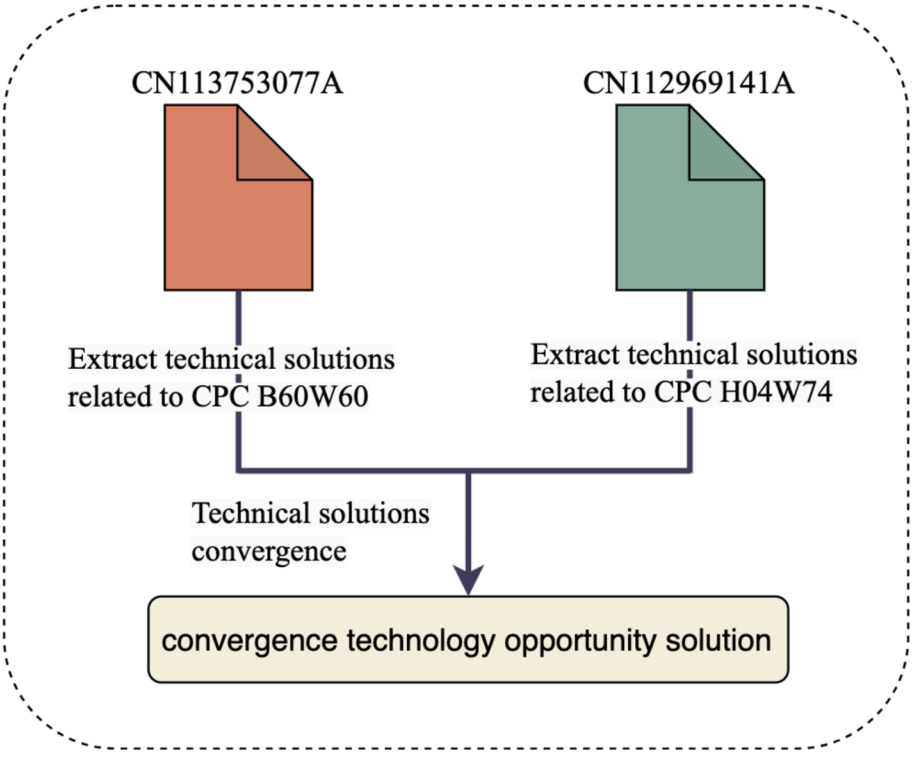


Fig. 5. Generating convergence technology opportunity solution from highly similar patent combination.

opportunities and offer more detailed technology opportunity solutions, providing clearer guidance for R&D engineers. Furthermore, the method proposed in this study can be systematically implemented. It consists of a series of automated processes, making it highly applicable.

5.2. Limitations and further research directions

Despite its contributions, this study has several limitations. Firstly, the patent data is limited to the Deventer patent database. Relying solely on a single database may not capture a comprehensive range of information regarding the technology convergence in the field of autonomous driving. Therefore, future research could explore the utilization of data from multiple databases. Secondly, it is based on group-level CPC predictions for technology convergence, thus overlooking the deeper linkage information of sub-group level CPCs. Therefore, future considerations may involve analyzing the prediction of technology convergence at different levels using sub-group level CPCs. Thirdly, this study aims to identify the convergence between pairs of technologies and cannot predict the convergence of three or more technologies. However, new technologies may be obtained from the convergence of more than two technologies. Therefore, future research can explore identifying technology opportunities from the perspective of the convergence of three or more technologies. Fourthly, the CPC classification code is a logically rigorous tree structure. The codes from sections to subgroups are obtained by gradually subdividing the field. Therefore, the distance in the CPC classification structure can be used to represent technological similarity. However, the similarity calculated by this method may be incomplete and simplified, and more detailed considerations are needed. For example, there may be two codes on different branches of the classification tree. They may still be similar in technology, but their technical similarity scores will be very low. Therefore, further exploration is needed for the characterization of technological similarity. Finally, although our model can identify most of the links in the future time period, the proportion of false positives in the positive samples predicted due to the use of all unlinked node pairs in time period 3 is relatively high. This necessitates additional validation processes, resulting in the wastage of R&D resources. Future research could explore methods to reduce false positives, such as removing CPC nodes with low frequency occurrences in the convergence network.

CRediT authorship contribution statement

Ziliang Wang: Writing – original draft, Methodology, Formal analysis, Conceptualization. **Wei Guo:** Supervision, Project administration, Funding acquisition, Formal analysis. **Hongyu Shao:** Writing – review & editing, Methodology, Formal analysis. **Lei Wang:** Writing – review & editing, Resources, Conceptualization. **Zhixing Chang:** Investigation, Data curation, Conceptualization. **Yuanrong Zhang:** Visualization, Validation, Methodology, Investigation. **Zhenghong Liu:** Validation, Resources.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Grant No. 72371185 and 52105248).

Appendix A. . Query strategies for autonomous driving patents

NO.	Search terms
#1 keyword search	Title Abstract Claims = (driverless car* OR unmanned car* OR self-driving car* OR pilotless car* OR automatic driving car* OR autonomous car* OR unpiloted car* OR autonomous driving car* OR driverless motor vehicle* OR unmanned motor vehicle* OR self-driving motor vehicle* OR pilotless motor vehicle* OR autonomous motor vehicle* OR unpiloted motor vehicle* OR automatic driving motor vehicle* OR autonomous driving motor vehicle* OR driverless automobile* OR unmanned automobile* OR self-driving automobile* OR pilotless automobile* OR autonomous automobile* OR unpiloted automobile* OR automatic driving automobile* OR autonomous driving automobile* OR driverless vehicle* OR unmanned vehicle* OR self-driving vehicle* OR pilotless vehicle* OR autonomous vehicle* OR unpiloted vehicle* OR automatic driving vehicle* OR autonomous driving vehicle*) NOT ALLD=(aerial vehicle* OR arial vehicle* OR underwater vehicle* OR air vehicle* OR flight vehicle* OR airplane OR aircraft OR flying machine* OR train OR ship OR boat OR plane OR aviation OR aeronautical OR aerobat OR aerocraft OR bicycle* OR AGV OR UAV OR AUV OR ROV)
#2 CPC search	CPC = (G05D00010088 OR G08G000116 OR G08G000122 OR G08G0001096791 OR B60K00310008 OR B60K00310058 OR B60K00310066 OR B60K20310091 OR B60K003100 OR G06T220730252 OR H04W000444 OR H04W000446 OR B60W003006 OR B60W003008 OR B60W003014 OR B60W0030085 OR B60W0030095 OR B60W0030165 OR B62D000600 OR B60W0060 OR B60W003012 OR B60W003016 OR B60W0030162 OR B60W003017 OR G05D0001021 OR G05D00010221 OR G05D00010223)
Full search terms	#1 OR #2

Data availability

Data will be made available on request.

References

[1] D. Zhu, A.L. Porter, Automated extraction and visualization of information for technological intelligence and forecasting, Technol. Forecast. Soc. Chang. 69 (2002) 495–506, https://doi.org/10.1016/S0040-1625(01)00157-3.
[2] S. Cozzens, S. Gatchair, J. Kang, K.S. Kim, H.J. Lee, G. Ordóñez, A. Porter, Emerging technologies: quantitative identification and measurement, Tech. Anal. Strat. Manag. 22 (2010) 361–376, https://doi.org/10.1080/09537321003647396.
[3] B. Yoon, C.L. Magee, Exploring technology opportunities by visualizing patent information based on generative topographic mapping and link prediction, Technol. Forecast. Soc. Chang. 132 (2018) 105–117, https://doi.org/10.1016/j.techfore. 2018.01.019.
[4] D. Kwon, S.Y. Sohn, Convergence Technology Opportunity Discovery for Firms Based on Technology Portfolio Using the Stacked Denoising AutoEncoder (SDAE), IEEE Trans. Eng. Manag. 71 (2024) 1804–1818, https://doi.org/10.1109/tem.2022. 3208871.
[5] C.S. Curran, J. Leker, Patent indicators for monitoring convergence–examples from NFF and ICT, Technol. Forecast. Soc. Chang. 78 (2011) 256–273, https://doi.org/10.1016/j.techfore.2010.06.021.
[6] P. Sharma, R.C. Tripathi, Patent citation: A technique for measuring the knowledge flow of information and innovation, World Pat. Inf. 51 (2017) 31–42, https://doi.org/10.1016/j.wpi.2017.11.002.
[7] S. Choi, M. Afifuddin, W. Seo, A Supervised Learning-Based Approach to Anticipating Potential Technology Convergence, IEEE Access. 10 (2022) 19284–19300, https://doi.org/10.1109/access.2022.3151870.
[8] C.S. Curran, S. Bröring, J. Leker, Anticipating converging industries using publicly available data, Technol. Forecast. Soc. Chang. 77 (2010) 385–395, https://doi.org/10.1016/j.techfore.2009.10.002.
[9] W.S. Lee, E.J. Han, S.Y. Sohn, Predicting the pattern of technology convergence using big-data technology on large-scale triadic patents, Technol. Forecast. Soc. Chang. 100 (2015) 317–329, https://doi.org/10.1016/j.techfore.2015.07.022.

[10] I. Park, B. Yoon, Technological opportunity discovery for technological convergence based on the prediction of technology knowledge flow in a citation network, J. Informetr. 12 (2018) 1199–1222, https://doi.org/10.1016/j.joi.2018.09.007.
[11] T.S. Kim, S.Y. Sohn, Machine-learning-based deep semantic analysis approach for forecasting new technology convergence, Technol. Forecast. Soc. Chang. 157 (2021) 120095, https://doi.org/10.1016/j.techfore. 2020.120095.
[12] C. Lee, S. Hong, J. Kim, Anticipating multi-technology convergence: A machine learning approach using patent information, Scientometrics. 126 (2021) 1867–1896, https://doi.org/10.1007/s11192-020-03842-6.
[13] J. Wang, J.-J. Lee, Predicting and analyzing technology convergence for exploring technological opportunities in the smart health industry, Comput. Ind. Eng. 182 (2023) 109352, https://doi.org/10.1016/j.cie.2023. 109352.
[14] M. Dong, Y. Sun, Y. Jin, C. Song, X. Zhang, X. Luo, Uncertainty graph convolution recurrent neural network for air quality forecasting, Adv. Eng. Inf. 62 (2024) 102651, https://doi.org/10.1016/j.aei.2024.102651.
[15] J. Wang, Y.-J. Chen, A novelty detection patent mining approach for analyzing technological opportunities, Adv. Eng. Inf. 42 (2019) 100941, https://doi.org/10.1016/j.aei.2019.100941.
[16] A.J.C. Trappey, C.V. Trappey, C.-Y. Fan, I.J.Y. Lee, Consumer driven product technology function deployment using social media and patent mining, Adv. Eng. Inf. 36 (2018) 120–129, https://doi.org/10.1016/j.aei.2018.03.004.
[17] A.J.C. Trappey, C.V. Trappey, J.-L. Wu, J.W.C. Wang, Intelligent compilation of patent summaries using machine learning and natural language processing techniques, Adv. Eng. Inf. 43 (2020) 101027, https://doi.org/10.1016/j.aei.2019.101027.
[18] A.K. Klevorick, R.C. Levin, R.R. Nelson, S.G. Winter, On the sources and significance of interindustry differences in technological opportunities, Res. Policy. 24 (1995) 185–205, https://doi.org/10.1016/0048-7333(93)00762-i.
[19] O. Olsson, Technological opportunity and growth, J. Econ. Growth. 10 (2005) 31–53, https://doi.org/10.1007/s10887-005-1112-4.
[20] F. Teng, Y. Sun, F. Chen, A. Qin, Q. Zhang, Technology opportunity discovery of proton exchange membrane fuel cells based on generative topographic mapping, Technol. Forecast. Soc. Chang. 169 (2021) 120859, https://doi.org/10.1016/j.techfore.2021.120859.
[21] B. Yoon, Y. Park, A systematic approach for identifying technology opportunities: Keyword-based morphology analysis, Technol. Forecast. Soc. Chang. 72 (2005) 145–160, https://doi.org/10.1016/j.techfore.2004.08.011.
[22] J. Cho, J. Lee, Development of a new technology product evaluation model for assessing commercialization opportunities using Delphi method and fuzzy AHP approach, Expert Syst. Appl. 40 (2013) 5314–5330, https://doi.org/10.1016/j.eswa.2013.03.038.
[23] P. Yu, J.H. Lee, A hybrid approach using two-level SOM and combined AHP rating and AHP/DEA-AR method for selecting optimal promising emerging technology, Expert Syst. Appl. 40 (2013) 300–314, https://doi.org/10.1016/j.eswa.2012.07.043.
[24] X. Li, Q. Xie, T. Daim, L. Huang, Forecasting technology trends using text mining of the gaps between science and technology: the case of perovskite solar cell technology, Technol. Forecast. Soc. Chang. 146 (2019) 432–449, https://doi.org/10.1016/j.techfore.2019.01.012.
[25] X. Li, Y. Wu, H. Cheng, Q. Xie, T. Daim, Identifying technology opportunity using SAO semantic mining and outlier detection method: A case of triboelectric nanogenerator technology, Technol. Forecast. Soc. Chang. 189 (2023) 122353, https://doi.org/10.1016/j.techfore.2023.122353.
[26] B. Kim, G. Gazzola, J.-M. Lee, D. Kim, K. Kim, M. Jeong, Inter-cluster connectivity analysis for technology opportunity discovery, Scientometrics. 98 (2014) 1811–1825, https://doi.org/10.1007/s11192-013-1097-2.
[27] B. Kim, G. Gazzola, J. Yang, J.M. Lee, B.Y. Coh, M. Jeong, Y.S. Jeong, Two-phase edge outlier detection method for technology opportunity discovery, Scientometrics. 113 (2017) 1–16, https://doi.org/10.1007/s11192-017-2472-1.
[28] J. Lee, S.Y. Sohn, Recommendation system for technology convergence opportunities based on self-supervised representation learning, Scientometrics. 126 (2020) 1–25, https://doi.org/10.1007/s11192-020-03731-y.
[29] J. Wang, T.-Y. Hsu, Early discovery of emerging multi-technology convergence for analyzing technology opportunities from patent data: the case of smart health, Scientometrics. 128 (2023) 4167–4196, https://doi.org/10.1007/s11192- 023-04760-z.
[30] N. Rosenberg, Technological change in the machine tool industry, 1840–1910, J. Econ. Hist. 23 (1963) 414–443, https://doi.org/10.1017/S0022050700109155.
[31] N. Sick, S. Bröring, Exploring the research landscape of convergence from a TIM perspective: a review and research agenda, Technol. Forecast. Soc. Chang. 175 (2021) (2021) 121321, https://doi.org/10.1016/j.techfore.2021.121321.
[32] M. Karvonen, T. Kassi, R. Kapoor, Technological innovation strategies in converging industries, Int. J. Bus. Innovat. Res. 4 (2010) 391–410, https://doi.org/10.1504/IJBIR.2010.034378.
[33] D. Rotolo, D. Hicks, B.R. Martin, What is an emerging technology? Research Policy. 44 (2015) 1827–1843, https://doi.org/10.1016/j.respol.2015.06.006.
[34] X. Chen, S. Jia, Y. Xiang, A review: knowledge reasoning over knowledge graph, Expert Syst. Appl. 141 (2020) 112948, https://doi.org/10.1016/j.eswa.2019.112948.
[35] J. Kim, S. Kim, C. Lee, Anticipating technological convergence: link prediction using wikipedia hyperlinks, Technovation. 79 (2019) 25–34, https://doi.org/10.1016/j.technovation.2018.06.008.
[36] S. Feng, H. An, H. Li, Y. Qi, Z. Wang, Q. Guan, Y. Li, Y. Qi, The technology convergence of electric vehicles: Exploring promising and potential technology convergence relationships and topics, J. Clean. Prod. 260 (2020) 120992, https://doi.org/10.1016/j.jclepro.2020.120992.
[37] S. Chang, M.-F. Francis Siu, H. Li, X. Luo, Evolution pathways of robotic technologies and applications in construction, Adv. Eng. Inf. 51 (2022), 101529, Doi: 10.1016/j.aei.2022.101529.
[38] U.H. Govindarajan, A.J.C. Trappey, C.V. Trappey, Intelligent collaborative patent mining using excessive topic generation, Adv. Eng. Inf. 42 (2019) 100955, https://doi.org/10.1016/j.aei.2019.100955.
[39] M.E. Leusin, J. Günther, B. Jindra, M.G. Moehrle, Patenting patterns in Artificial Intelligence: Identifying national and international breeding grounds, World Pat. Inf. 62 (2020) 101988, https://doi.org/10.1016/j.wpi.2020.101988.
[40] S. Chang, M.-F. Francis Siu, H. Li, X. Luo, Evolution pathways of robotic technologies and applications in construction, Adv. Eng. Inf. 51 (2022), 101529, Doi: 10.1016/j.aei.2022.101529.
[41] C.H. Song, D. Elvers, J. Leker, Anticipation of converging technology areas—A refined approach for the identification of attractive fields of innovation, Technol. Forecast. Soc. Chang. 116 (2017) 98–115, https://doi.org/10.1016/j. techfore.2016.11.001.
[42] E. Kim, Y. Cho, W. Kim, Dynamic patterns of technological convergence in printed electronics technologies: patent citation network, Scientometrics. 98 (2014) 975–998, https://doi.org/10.1007/s11192-013-1104-7.
[43] J. Kim, S. Lee, Forecasting and identifying multi-technology convergence based on patent data: the case of IT and BT industries in 2020, Scientometrics. 111 (2017) 47–65, https://doi.org/10.1007/s11192-017-2275-4.
[44] M. Park, Y. Geum, Two-stage technology opportunity discovery for firm-level decision making: GCN-based link-prediction approach, Technol. Forecast. Soc. Chang. 183 (2022) 121934, https://doi.org/10.1016/j.techfore. 2022.121934.
[45] J.H. Cho, J. Lee, S.Y. Sohn, Predicting future technological convergence patterns based on machine learning using link prediction, Scientometrics. 126 (2021) 5413–5429, https://doi.org/10.1007/s11192-021-03999-8.
[46] J. Wang, L. Cheng, L. Feng, K.Y. Lin, L. Zhang, W. Zhao, Tracking and predicting technological knowledge interactions between artificial intelligence and wind power: Multimethod patent analysis, Adv. Eng. Inf. 58 (2023) 102177, https://doi.org/10.1016/j.aei.2023.102177.
[47] H. Sasaki, I. Sakata, Identifying potential technological spin-offs usinghierarchical information in international patent classification, Technovation. 100 (2021) 102192, https://doi.org/10.1016/j.technovation.2020.102192.
[48] V. Martínez, F. Berzal, J.-C. Cubero, A survey of link prediction in complex networks, ACM Comput. Surv. 49 (2016) 1–33, https://doi.org/10.1145/3012704.
[49] F. Caviggioli, Technology fusion: Identification and analysis of the drivers oftechnology convergence using patent data, Technovation. 55 (2016) 22–32, https://doi.org/10.1016/j.technovation.2016.04.003.
[50] K. Eilers, J. Frischkorn, E. Eppinger, L. Waltera, M.G. Moehrlea, Patent-based semantic measurement of one-way and two-way technology convergence: the case of ultraviolet light emitting diodes (UV-LEDs), Technol. Forecast. Soc. Chang. 140 (2019) 341–353, https://doi.org/10.1016/j.techfore.2018.12.024.
[51] T. Kose, I. Sakata, Identifying technology convergence in the field of robotics research, Technol. Forecast. Soc. Chang. 146 (2019) 751–766, https://doi.org/10.1016/j.techfore.2018.09.005.
[52] Z. Chang, W. Guo, L. Wang, Z. Fu, J. Ma, G. Zhang, Z. Wang, A novel patent technology characterization method based on heterogeneous network message passing algorithm and patent classification system, Expert Syst. Appl. 256 (2024) 124895, https://doi.org/10.1016/j.eswa.2024.124895.
[53] Sun, K.; Liu, J.; Yu, S.; Xu, B.; Xia, F. Graph force learning. In Proceedings of the 2020 IEEE International Conference on Big Data(Big Data), Atlanta, GA, USA, 2020, pp. 2987–2994, Doi: 10.1109/BigData50022.2020.9378120.
[54] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, Proceedings of ICLR 2017 (2017) 1–14, https://doi.org/10.48550/arXiv.1609.02907.
[55] E.C. Mutlu, T. Oghaz, A. Rajabi, I. Garibay, Review on Learning and Extracting Graph Features for Link Prediction, Mach. Learn. Knowl. Extr. 2 (2020) 672–704, https://doi.org/10.3390/make2040036.
[56] A. Kumar, S.S. Singh, K. Singh, B. Biswas, Link prediction techniques, applications, and performance: A survey, Phys. A Stat. Mech. Its Appl. 553 (2020) 124289, https://doi.org/10.1016/j.physa.2020.124289.
[57] C.-H. Lee, C.-L. Liu, A.J.C. Trappey, J.P.T. Mo, K.C. Desouza, Understanding digital transformation in advanced manufacturing and engineering: A bibliometric analysis, topic modeling and research trend discovery, Adv. Eng. Inf. 50 (2021) 101428, https://doi.org/10.1016/j.aei.2021.101428.
[58] A.J.C. Trappey, C.V. Trappey, U.H. Govindarajan, J.J.H. Sun, Patent Value Analysis Using Deep Learning Models—The Case of IoT Technology Mining for the Manufacturing Industry, IEEE Trans. Eng. Manag. 68 (2021) 1334–1346, https://doi.org/10.1109/tem.2019.2957842.
[59] M. Reitzig, Improving patent valuations for management purposes-Validating new indicators by analyzing application rationales, Res. Policy. 33 (2004) 939–957, https://doi.org/10.1016/j.respol.2004.02.004.
[60] A.M. Ruiz, T.A. Banet, Toward the definition of a structural equation model of patent value: PLS path modelling with formative constructs, Revstat-Statistical J. 7 (2009) 265–290.
[61] A. Rodriguez, et al., Patent clustering and outlier ranking methodologies for attributed patent citation networks for technology opportunity discovery, IEEE Trans. Eng. Manage. 63 (2016) 426–437, https://doi.org/10.1109/TEM.2016.2580619.
[62] Trappey C.V., Trappey A. J. C., Liu, B, Identify trademark legal case precedents - Using machine learning to enable semantic analysis of judgments, world Pat. inf. 62 (2020) 101980, Doi: 10.1016/j.wpi.2020.101980.

[63] J. Deichmann, E. Ebel, K. Heineke, R. Heuss, M. Kellner, F. Steiner, Autonomous driving's future: Convenient and connected, McKinsey & Company, 2023.
[64] Z. Xie, K. Miyazaki, Evaluating the effectiveness of keyword search strategy for patent identification, World Pat. Inf. 35 (2013) 20–30, https://doi.org/10.1016/j.wpi.2012.10.005.
[65] Y. Ji, X. Zhu, T. Zhao, L. Wu, M. Sun, Revealing technology innovation, competition and cooperation of self-driving vehicles from patent perspective, IEEE Access 8 (2020) 221191–221202, https://doi.org/10.1109/ACCESS.2020.3042019.