IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 26 September 2023, accepted 16 November 2023, date of publication 22 November 2023, date of current version 29 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3335835

RESEARCH ARTICLE

Using Machine Learning With Technological Innovation Factors to Predict the Transferability of University Patents

DISHA DENG[1] **AND TAO CHEN**[1,2]
[1]School of Management, Wuhan University of Science and Technology, Wuhan 430081, China
[2]Industrial Policy and Management Research Center of Hubei, Wuhan 430081, China

Corresponding author: Tao Chen (chentao@wust.edu.cn)

This work was supported in part by the Hubei Provincial University Teaching Reform Research Project: Culture Construction and Practice of Science and Engineering Universities from the Perspective of Higher Education Quality Assurance under Grant 2018243, and in part by the Key Project of Hubei Industrial Policy and Management Research Center: Research on the Service System of Industrial Technology Transfer in Hubei Local Universities under Grant 17CYZ01.

ABSTRACT We quantify the impact of technological innovation factors on university patent transferability, accurately identify transferable patents, and address the lack of interpretability in existing patent transferability models. Firsly, we apply the latent Dirichlet allocation (LDA) model to conduct text mining and feature extraction on abstracts of university patents in the field of artificial intelligence to obtain the technological innovation features of university patents. We then construct a patent transferability fusion index system that includes technological innovation features and quality features. Four typical machine learning algorithms, namely support vector machine (SVM), random forest (RF), artificial neural network (ANN), and extreme gradient boosting (XGBoost) are used to predict university patent transferability. We use SHapley Additive exPlanations (SHAP) to explore feature importance and interactions based on the model with the strongest performance. Our results show that (1) XGBoost outperforms the other algorithms in predicting university patent transferability; (2) fusion indicators can effectively improve prediction performance with respect to university patent transferability; (3) the importance of technological innovation features generated with XGBoost is generally high; and (4) the impact of both technology innovation and patent quality features on university patent transferability is nonlinear and there are significant positive interaction effects between them.

INDEX TERMS Artificial intelligence, LDA, machine learning, patent transferability, SHAP, technological innovation, university patents.

I. INTRODUCTION

In today's knowledge-based economy, technological innovation is not only essential to an enterprise's survival and sustainable development, but it also embodies the competitiveness that is at the core of industries, regions, and countries. Patents are an institutional by-product of technological innovation that provide both protective and incentive benefits. However, not all patents spread technological innovation or involve innovation effects. From an economic perspective, new products, processes, and systems that become a part of commercial activities for the first time are considered innovations [1]. Therefore, only through industrialization can patents achieve market value and promote social and economic development. Universities play an important role in a country's innovation system by producing many patents, but many of these patents do not support economic activity, resulting in a waste of scientific and technological resources. Identifying transferable patents can support universities to recover the investment costs of patent research and development as well as the management costs required to maintain patents by transferring patents to generate profits. In addtion,

The associate editor coordinating the review of this manuscript and approving it for publication was Bo Pu.

identifying transferable patents can help enterprises purchase patents with economic potential to enhance their market competitiveness. Therefore, selecting suitable patents for trading is conducive to the process of patent transfer between universities and enterprises, thereby promoting the efficiency of university patent transfer. In this context, constructing a research to evaluate the university patent transferability is of great significance.

A few studies show the impact of patent quality and intrinsic value on patent transferability, mainly by relying on features of the patent literature and patent inventor characteristics to construct patent transferability prediction models [2]. However, patents with transferability potential usually have both technical quality and market value [3]. Therefore, focusing on the effects of patent quality and intrinsic value on patent transferability is not sufficient, as it neglects the fact that patents are tradable commodities in the technology market; their value must be recognized, and an investment is required to realize that economic value. Many studies propose that the process of patenting technological innovation should be improved to facilitate the transformation of scientific and technological achievements to help promote economic growth [4], [5].

The ability to engage in technological innovation contributes to factor to patent transferability, and including features associated with technological innovation involving patents into the study of how to predict patent transferability is of great significance, as is exploring the interaction effect between patent technological innovation features and patent quality features. Most studies predict patent transferability using black-box machine learning models that perform well but lack interpretability, making it difficult to provide practical guidance for universities and enterprises seeking to improve patent transferability.

As the most effective repository of technical information, patent texts typically contain the latest technological intelligence in each field and can reflect the current level of technological innovation [6]. As the core of the 4$^{\text{th}}$ industrial revolution, artificial intelligence has developed into an extensive technology that integrates multiple disciplines and fields and is now considered one of the key technologies driving economic and social development [7]. Based on the above, we use patent data in the field of artificial intelligence from Chinese universities over the past five years and obtain patent technology innovation indicators by applying the latent Dirichlet allocation (LDA) model, establishing a patent transferability indicator system that includes features of patents that indicate both technological innovation and quality, and exploring the importance of, as well as the interaction effect between these features in predicting university patent transferability based on machine learning models and the SHapley Additive exPlanations (SHAP) framework. We seek to improve prediction accuracy regarding the transferability of university-based patents and provide a decision-making tool for universities and enterprises to promote patent transfer.

II. RELATED WORK

A. PATENT TRANSFERABILITY

As the term suggests, ''patent transferability'' refers to the ease with which a patent can be transferred. Existing studies evaluate patent transferability from a patent quality perspective by constructing evaluation index systems that mostly involve technical, legal, and economic dimensions based on the characteristics of the patent literature [8], [9], [10]. Recent studies have observed that the content of patent texts affects patent transferability. For example, Lee et al. [11] construct a patent transferability model using the LDA model and Adaboost algorithm and find that theme-related factors improve the model's prediction accuracy. Ran et al. [12] combine the LDA model with the K-means algorithm to determine a patent's technical theme and integrated this with patent evaluation indicators to identify convertible patents from universities. Biao et al. [13] apply the Bidirectional Enoceder Representations from Transformers model to represent the semantic features of patent texts using d-dimensional feature vectors, construct a university patent value evaluation model using machine learning algorithms, and predict the probability of patent transfer. Although these studies consider the impact of patent text content on patent transferability, they are limited in that they cluster and partition patent texts as feature variables for use in predicting patent transferability, rather than extracting, defining and measuring innovative elements in the patent text; thus, such studies cannot explore the impact of technological innovation factors on patent transferability. Furthermore, these studies mainly focus on the predictive performance of patent transferability models and pay less attention to the model's interpretability, making it difficult to provide practical guidance to universities and other enterprises seeking to improve patent transferability.

To address the shortcomings of the existing research in this area, we construct a patent transferability index system that includes both patent technological innovation features and quality features, and then establishes a patent transferability prediction model using different machine learning algorithms. To enhance model interpretability, the SHAP framework is applied based on the model with the best performance in terms of identifying the importance of and interaction effects between features that affect patent transferability.

B. IMPACT OF TECHNOLOGICAL INNOVATION CAPABILITY ON PATENT TRANSFERABILITY

Many scholars have pointed out that theoretically, an insufficient level of technological innovation leads to a low patent conversion rate. Specifically, Zong et al. [14] show that many experimental studies conducted in Chinese universities are solely aimed at applying for patents and resulted in fewer original achievements, and that their patents had less value in terms of industrial application, impeding the growth in patent conversion rates for those universities. Yang et al. [15] show that the innovation efficiency of Chinese key core patents is low, that problems of ''patent foam'' and ''innovation illusion'' are still prominent, and that key core patents are severely disconnected from the market, with the result that most key core patents do not transform into technological competitive advantages for commercial enterprises.

In terms of empirical analyses, most of the relevant studies [16], [17] measure the impact of technological innovation on patent industrialization from a mesoscopic perspective, which uses the amount of technology research and funding in each university, enterprise, or region as the input indicator and the revenue generated from new products as the output indicator for technological industrialization. The impact is then estimated in terms of technological innovation efficiency using the data envelopment analysis (DEA) model. This type of research confirms the importance of technological innovation capabilities in patent transformation but still has certain limitations. First, in the DEA method the process of technological innovation is a black box, considering only the efficiency of technological innovation as the output, and lacks the direct measurement of patent technological innovation. Second, studying the technological innovation capability of a certain unit or region from a mesoscopic perspective does not capture the relationship between the technological innovation capability of a given patent technology and its transferability. Therefore, it is important to identify the technological innovation ability reflected in patent content from a micro perspective to empirically analyze the impact of technological innovation ability on university patent transferability.

C. TEXT-BASED PATENT TECHNOLOGY INNOVATION EVALUATION METHOD

As important repositories of technological innovation, patents contain a large amount of leading-edge technological information, and the semantic information included in patent texts can reflect content about technological innovation. Previous research mainly analyses patent documents based on probabilistic topic models, including the widely used probabilistic latent semantic analysis (pLSA) and LDA models. For example, Bao et al. [18] use the pLSA algorithm to construct the technical and efficacy dimensions of a patent technology mining model for patents involving titanium; Han et al. [19] identify breakthrough innovative technology topics in the solar photovoltaic field based on the LDA model, and Kim et al. [20] use the LDA model to analyze patents in the United States Patent and Trademark Office database to identify emerging and nascent technology areas for wireless power transmission. Based on the available research, the LDA model has proven to be superior to the pLSA model in some respects. First, the LDA approach can eliminate overfitting problems and compute scalable fine-grained, low dimensional semantic representations [21]. Second, LDA can enhance the capture of interchangeability between mixed model words and documents [22]. When dealing with big data, computational complexity can be reduced [23] and overfitting problems can also be avoided through the Dirichlet distribution used in LDA [24]. Owing to the large amount of university patent text data involved, we apply the LDA method in this study.

III. MATERIALS AND METHODS

A. LATENT DIRICHLET ALLOCATION

The LDA topic model is an unsupervised machine learning algorithm based on the word bag model, which can be used to identify hidden topic information in large-scale documents or corpora. As a three-layer Bayesian probability model, the LDA model consists of a three-layer structure of documents, topics, and words [25]. Each document is composed of a certain probability of topics, and each topic is composed of a certain probability of words. Polynomial distributions of the document and topic words can be obtained through the LDA topic model. The probability of words reflects the strength of the correlation between the words and the topics. In studying topic discovery and evolution, the LDA model can accurately extract topics from texts and discover popular topics. Perplexity is an evaluation criterion for the quality of a probability distribution in the LDA model, which represents the uncertainty of the topic to which the document belongs, the value of perplexity basically shows a decreasing trend as the number of potential topics increases. Models with better performance generally have lower perplexity. The optimal number of topics can be obtained according to the perplexity with the lowest value or at the inflection point of the perplexity curve. The perplexity can be calculated using euqation (1):

$$Perplexity(D) = \exp \frac{\sum_{d=1}^{M} \log P(W_d)}{\sum_{d=1}^{M} N_d} \tag{1}$$

where D represents the set of all words in the document; M represents the number of documents; W_d represents the word in document d; N_d represents the number of words for d in each document; $P\ (W_d)$ represents the probability of words appearing in the document.

B. PATENT TECHNOLOGY INNOVATION INDICATORS

Given that patent content is related to the core elements of technological innovation in each research field and illustrates the technological evolution process of each industry [26], the consistency between a patent and the industry's development direction can be reflected by the number of core keywords contained in the patent. Furthermore, due to differences in the importance of various innovative elements there will also be differences in the popularity of core keywords. The greater the popularity of a keyword, the more attention one should pay to the content associated with the keyword for a given industry. Therefore, innovation core value can be regarded as a numerical expression of the degree to which core keywords are valued in a patent's content, which also indicates the patent applicant's understanding of the core technical elements in each field [27]. Accordingly, the number of core keyword and innovation core value are used as patent technology innovation indicators in this research.

Core keywords are commonly used terms related to technological innovation achievements in a specific field. To obtain the core keywords, we apply the term frequency–inverse document frequency (TF-IDF) value as the word frequency. We then perform LDA modeling on all university patents texts in the field of artificial intelligence, retaining the h topic words with the highest weights under each topic, and selecting nouns as the core keywords to obtain the set of core keywords $\{T_1 \cdots T_j \cdots T_n\}$. We then count the number of core keywords in patent t as the core keyword indicator for that patent.

After obtaining the core keyword set $\{T_1 \cdots T_j \cdots T_n\}$ for the field of artificial intelligence, the importance value $I(T_j)$ of keyword T_j can be calculated using equation (2) [27]:

$$I(T_j) = \sum_{i=1}^{K} P_i(T_j) * N_i \tag{2}$$

where K represents the number of topics for LDA modeling of all university patents, $P_i T_j$ represents the probability of keyword T_j in topic i, and N_i represents the number of core keywords included in topic i. Subsequently, the innovation core value C_t of a patent in each field can be calculated according to equation (3) as follows:

$$C_t = \sum_{j=1}^{n} I(T_j) * W(T_j) \tag{3}$$

where $I(T_j)$ is the important value of keyword i, and $W(T_j)$ represents the frequency of core keyword T_j occurring in a single patent abstract.

C. EXTREME GRADIENT BOOSTING

The Extreme Gradient Boosting (XGBoost) is an ensemble algorithm of machine learning, which evolved from the Gradient Augmented Regression Tree algorithm, has been widely used due to its outstanding efficiency and accuracy [28]. The XGBoost algorithm can accelerate computation speed through parallel learning and add modifications to the objective function to improve prediction accuracy, as well as utilize regularization enhancement techniques to reduce overfitting, thereby ensuring the robustness of the model and achieving better prediction performance under limited training samples and time. The distinctive properties of XGBoost make the algorithm more accurate and faster than other existing algorithms [29].

D. RANDOM FOREST

The Random Forest model developed by Breiman is an integrated algorithm that includes multiple decision trees, and has a high classification accuracy and generalization ability [30]. The basic idea of the RF algorithm is to combine multiple weak classifiers into a strong classifier, which makes it can effectively handle multi label problems in classification problems. The RF algorithm first extracts multiple samples from the original training set using bootstrap resampling, then constructs and combined decision trees for each bootstrap sample; the final prediction result is maintained by voting.

E. SUPPORT VECTOR MACHINE

The Support Vector Machine (SVM) algorithm developed by Vapnik for binary classification is a structural risk minimization algorithm theory based on machine learning, proposed to solve nonlinear regression and classification problems. The SVM searches for the optimal hyperplane in n-dimensional classification space that has the highest margin between classes [31]. The basic idea of the SVM algorithm is to use appropriate kernel functions to map the data to be classified into a higher dimensional feature space with certain fault-tolerant conditions. The SVM algorithm builds two parallel hyperplanes on both sides of the separated dataset, and maximizes the distance between the two parallel hyperplanes to achieve data classification.

F. ARTIFICIAL NEURAL NETWORK

The Artificial Neural Network (ANN) algorithm, which was derived from biological neural networks [32], is commonly used for fitting nonlinear functions and has an outstanding ability to determine the meaning and rules of complicated data. The ANN algorithm establishes a nonlinear relationship between input and output variables by setting input layers, hidden layers, and output layers combined with excitation functions. The ANN algorithm optimizes the model based on the error between calculated and true values, and then changes the weight of the model equation through back propagation. The high-precision nonlinear approximate mathematical model can be obtained through repeatedly training..

G. SHAPLEY ADDITIVE EXPLANATIONS

Although machine learning models based on ensemble algorithms perform extremely well, increased complexity reduces the interpretability of these model. The SHAP framework was introduced to improve the interpretability of the results produced by machine learning models. SHAP is an explanatory framework for black-box models [33] based on Shapley value-based calculations derived from alliance game theory that measures how features and their interactions affect the dependent variable [34]. SHAP considers each feature as a contributor and calculates and summarizes each feature's contribution value to obtain the model's final prediction. A SHAP value greater than zero indicates a feature has a positive effect on the result; a SHAP value below zero suggests the feature has a negative effect on the result. SHAP indicates a feature's importance and show the relationship between features, thereby explaining how variables affect the prediction results. SHAP can improve model interpretability by providing feature importance, feature dependence, local explanations and summary plots [35].

TABLE 1. Variables in predictive models.

Dimension	Indicator	Symbol	Explanation	Type
Technical Dimension	Classification number	nIPC	Reflects the technological width of patents	Numerical
	Backward Citation	nBWD_citing	Reflects the impact of patents on subsequent innovation	Numerical
	Forward Citation	nFWD_citing	Reflects the reference and inheritance of the patents to other patents	Numerical
	Non-patent citation	nNPL	Reflects the integration between a patented technology and scientific research	Numerical
	Inventor	nInventor	Reflects the size of the core research team	Numerical
Legal Dimension	Claim	nClaim	Reflects the stability of patent technology rights	Numerical
	Litigation	Litigation	Indicates whether the patent has been involved in a lawsuit, reflecting its economic value	Nominal
Market Dimension	PCT application	PCT	Indicates whether the patent was submitted through the Patent Cooperation Treaty, reflecting the breadth of the international patent market layout	Nominal
	Family country	nFamily_country	Reflects the scope of patent inflows into other countries' markets	Numerical
	Family size	Family_size	Reflects the scope of legitimate protection for patents	Numerical
Technological Innovation Dimension	Core keyword	nKeyword	Reflects the consistency between patent research and industry development	Numerical
	Innovation core value	Innovation_value	Reflects the extent to which the patent reflects the core elements of innovation	Numerical

H. MODEL EVALUATION INDICATORS

We evaluate our model's accuracy and generalization performance using the area under the ROC curve (AUC), along with accuracy, specificity, and sensitivity [36]. Accuracy is measured as the ratio of the number of correctly classified samples to the total number for a given test dataset, while recall rate represents the ratio of the predicted positive sample size to the actual positive sample size. Precision represents the ratio of actual positive examples to the predicted positive examples. The F-score represents the harmonic average between accuracy and the recall rate. The ROC curve was drawn with the true positive rate as the vertical axis and false positive rate as the horizontal axis, which can reflect the performance of different classification algorithms. The closer the ROC curve is to the vertical axis, the larger the AUC and the better the algorithm's performance.

IV. DATA PREPROCESSING

A. DATA SOURCE AND INDICATOR SYSTEM CONSTRUCTION

We selected 45,756 invention patents in the field of artificial intelligence authorized by Chinese universities from 2018 to 2022 as our research sample. Patent data was obtained from the IncoPat scientific and technological database, a platform for innovation information. We construct an evaluation index system of university patent transferability using patent quality and technological innovation dimensions. Specifically, we select ten specific patent quality indicators from three aspects of patents that refer to previous research, namely the technology, legal, and market dimensions of that previous research [37], [38], [39], [40]. In addition, we use the number of core keywords and innovation core values as patent technological innovation indicators, as shown in Table 1.

B. ACQUISITION OF PATENT TECHNOLOGY INNOVATION INDICATORS

In contrast to the quality indicators, technology innovation indicators cannot be obtained directly from patent data statistics. We first need to include keywords extracted from the text of university patent abstracts using the LDA model.

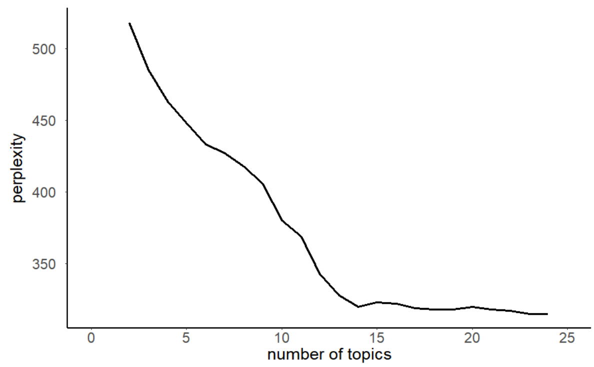


FIGURE 1. Perplexity of LDA trained on the university patents abstracts.

In preprocessing the text before modeling, each patent abstract was regarded as a document; those that were shorter than 100 characters in Chinese were excluded. After word segmentation, stemming, stop word removal, and other natural language processing processes were completed, keywords were extracted to establish the user dictionary and form an experimental corpus. The graph of topics versus perplexity using the LDA model is shown in Figure 1. We select 14 as the optimal number of topics as that is the number at the inflection point of the perplexity curve.

After LDA modeling and TF-IDF calculations were completed, the top ten subject terms with the highest weight for each topic were retained, and the patent core keywords in the field of artificial intelligence were selected from these subject terms, which are presented in Table 2.

After obtaining the core keywords in the field of artificial intelligence and counting the number of core keyword contained in each patent, we can calculate the innovation core value of each patent using equation (3) and obtain the input dataset as listed in Table 3, which establishes a data foundation for subsequent model training.

TABLE 2. Top ten terms associated with fourteen topics learned on LDA model.

Topic	Keyword
Digital Signal Processing	signal (0.062), sparse matrix (0.027), noise (0.019), radar (0.016), vibration (0.010), imaging (0.009), filtering (0.008), phase (0.007), frequency (0.005), spectrum (0.002)
Bayesian Probability Model	classifier (0.054), clustering (0.033), matrix (0.023), label (0.018), eigenvector (0.016), node (0.015), Bayesian (0.011), hash (0.008), Gauss (0.006), network topology (0.005)
Unmanned Aerial Vehicle	unmanned aerial vehicle (0.104), path (0.026), aircraft (0.021), cluster (0.015), track (0.015), rotor (0.014), formation (0.011), base station (0.009), obstacle (0.012), distributed (0.007)
Positioning and Tracking	positioning (0.033), matching (0.032), 3D Point Cloud (0.030), region (0.021), coordinate system (0.019), fingerprint (0.018), pixel (0.012), projection (0.011), lidar (0.005), contour (0.002)
Medical Devices	patients (0.078), diseases (0.063), medical treatment (0.062), Doppler (0.055), Fourier (0.051), radiation source (0.042), image denoising (0.036), heart rate (0.022), ultrasound (0.016), magnetic resonance (0.015)
Autonomous Driving Technology	vehicle (0.062), driverless (0.023), trajectory (0.022), road (0.016), dynamics (0.014), controller (0.013), acceleration (0.010), angular velocity (0.009), nonlinearity (0.007), obstacle (0.006)
Natural Language Processing	text (0.048), knowledge (0.022), semantics (0.019), embedding (0.018), natural language (0.014), keywords (0.009), corpus (0.008), document (0.006), context (0.005), topic extraction (0.002）
Face Recognition	video (0.071), face recognition (0.040), tracking (0.027), action (0.021), posture (0.016), occlusion (0.016), monitoring (0.015), expression (0.009), gesture (0.004), light (0.003)
Image Processing	region (0.043), spectrum (0.024), pixel (0.022), remote sensing (0.021), infrared (0.016), resolution (0.013), edge (0.012), gray scale (0.010), texture (0.009), pixel (0.009)
Automatic Control System	module (0.054), device (0.031), controller (0.023), sensor (0.020), flexibility (0.011), spring (0.009), drive (0.008), monitoring (0.005), support (0.003), command (0.002)
Swarm Intelligence Optimization Algorithm	optimization (0.060), iteration (0.047), particle swarm optimization (0.033), solving (0.031), fault diagnosis (0.029), simulation (0.028), convergence (0.023), initialization (0.012), dynamic constraints (0.007), local (0.003)
Decision Support System	decision (0.023), terminal (0.021), auxiliary (0.019), storage (0.018), test case (0.016), threshold (0.015), cognition (0.013), global (0.009), simulation (0.005), decision (0.003)
Neural Network	convolution (0.072), depth (0.032), module (0.031), encoding (0.025), feature extraction (0.019), neuron (0.015), decoding (0.015), circulation (0.014), pooling (0.012), synaptic (0.009)
Intelligent Fault Localization System	prediction (0.106), time series (0.038), status (0.011), load (0.01), indicator monitoring (0.008), dynamic warning (0.007), real time collection (0.007), historical data (0.006), risk (0.004), fault (0.003)

TABLE 3. Input values for university patent transferability model (partial data).

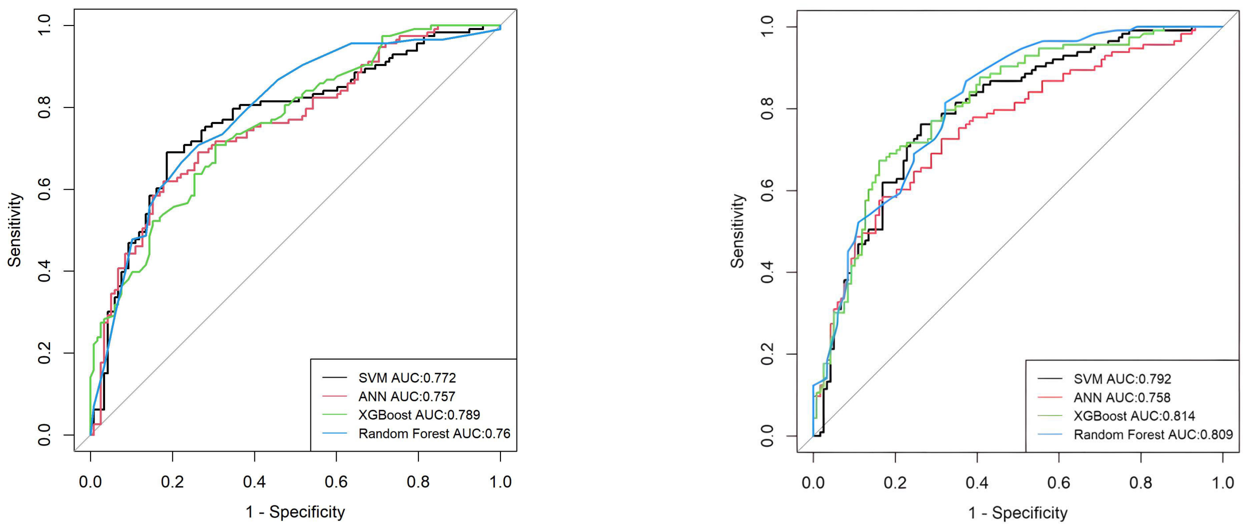
Patent Number	CN110310277B	CN110475205B	CN108320017B	CN109871838B	CN110536299B	CN109146944B
nIPC	8	6	2	3	3	3
nBWD_citing	7	7	3	11	5	8
nFWD_citing	0	0	1	0	0	1
nNPL	2	3	3	0	0	2
nInventor	6	6	7	4	6	6
nClaim	8	4	10	5	5	3
Litigation	0	0	0	0	0	0
PCT	0	0	0	0	0	0
nFamily_country	1	1	1	1	2	1
Family_size	2	2	2	2	3	2
nKeyword	3	11	12	7	4	15
Innovation_value	0.073	0.693	0.418	0.294	0.288	0.760
Transfered or not	0	1	1	0	0	1

C. SOFTWARE IMPLEMENTATION

The data shows that few of the university patents in our sample have been transferred. Since an imbalance in the binary classification outcome variables in the training set reduces model efficiency and prediction accuracy [41], we applied the SMOTE algorithm, which can effectively balance the minority sample size with the majority sample size, to address the problem of unbalanced outcome variables, thereby reducing excessive skewing in our datasets [42]. The SMOTE algorithm was implemented using the smotefamily package of R. SVM, XGBoost, RF, and artificial neural network models were implemented using the

TABLE 4. **Prediction performance of the four machine learning models.**

Model	Random Forest		XGBoost		SVM		ANN	
Technical innovation characteristics included	Yes	No	Yes	No	Yes	No	Yes	No
Accuracy	0.709	0.675	0.721	0.697	0.703	0.682	0.696	0.685
Recall	0.735	0.717	0.739	0.752	0.728	0.731	0.711	0.728
F-score	0.722	0.695	0.730	0.723	0.715	0.706	0.703	0.701



(a) Technological features excluded

(b) Technological features included

FIGURE 2. **ROC curves of the four machine learning models when technological innovation features were excluded (a) and included (b).**

e1071, xgboost, randomforest, and net packages, respectively. We used the grid search method to adjust the parameters, and a 5-fold cross-validation method to find the optimal parameter combination with the highest accuracy and the lowest cross-validation error. We used the shapviz package to build the SHAP model and produce visualizations of the results.

V. RESULTS

A. MODEL EVALUATION

First, we compare the prediction performance for each of the four models twice; the technological innovation features are excluded from the first analysis and included in the second analysis. We measure the results in terms of accuracy, recall, and F-score, as shown in Table 4. Figure 2 shows the ROC curve for each model. The analysis shows that when technological innovation indicators are included, the prediction performance for each of the four models improves markedly compared with the results when the technological innovation indicators are excluded. The XGBoost algorithm produces the best prediction performance among the four models, regardless of whether technological innovation indicators are considered.

B. MODEL INTERPRETATION AND ANALYSIS BASED ON SHAP

1) MODEL GLOBAL INTERPRETATION

Taking XGBoost as the optimal model, feature importance is denoted using the SHAP value, as shown in Figure 3.

Figure 3 shows that the SHAP values for most of the indicators are either ''low on the left'' or ''high on the right.'' The indicators that have a positive impact on patent transfer are innovation core value, non-patent citations, claims, core keywords, classification number, family country, backward citation, and inventor. Among them, the innovation core value, non-patent citation, and claims make relatively significant contributions to patent transferability. In contrast, forward citation has a negative impact on patent transferability. Whether or not the patent was submitted through the Patent Cooperation Treaty or has been the subject of litigation has no significant impact on patent transferability.

2) VARIABLE INTERACTIVE INTERPRETATION

The influence of these factors on patent transfers includes the main influence of each factor and the interaction between factors. An interaction value from the SHAP algorithm greater

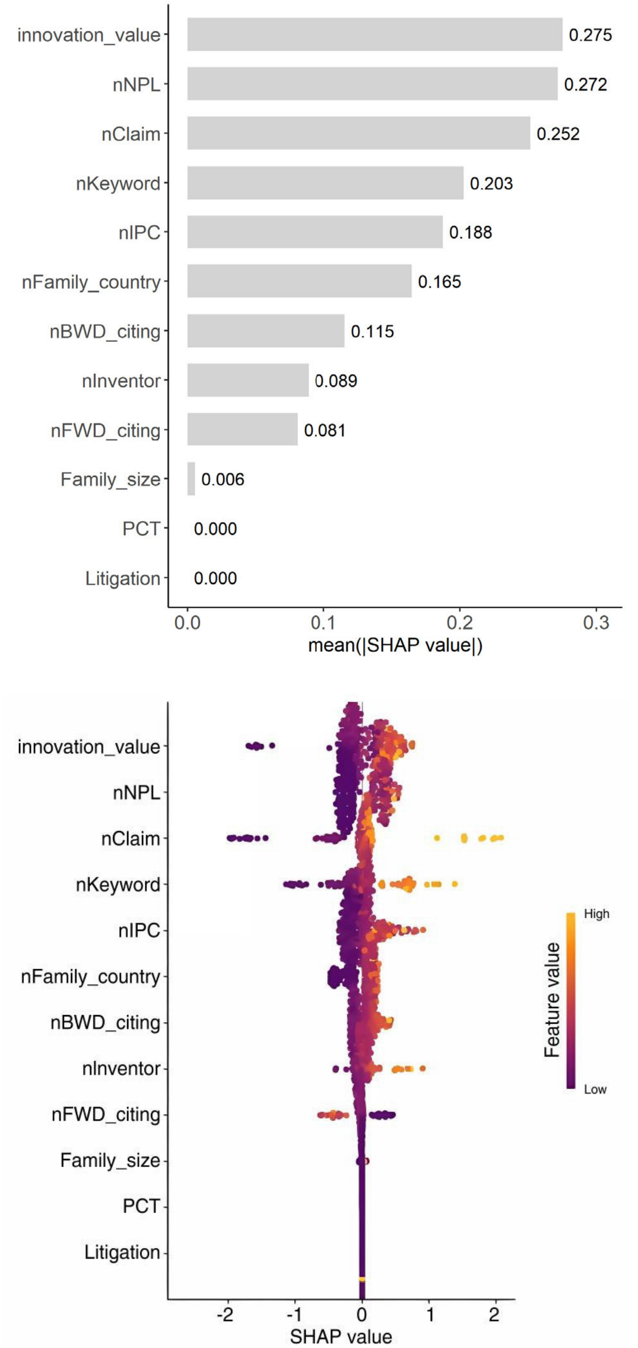


FIGURE 3. **Feature importance generated with the XGBoost model based on SHAP value. Features are ranked in descending order of importance.**

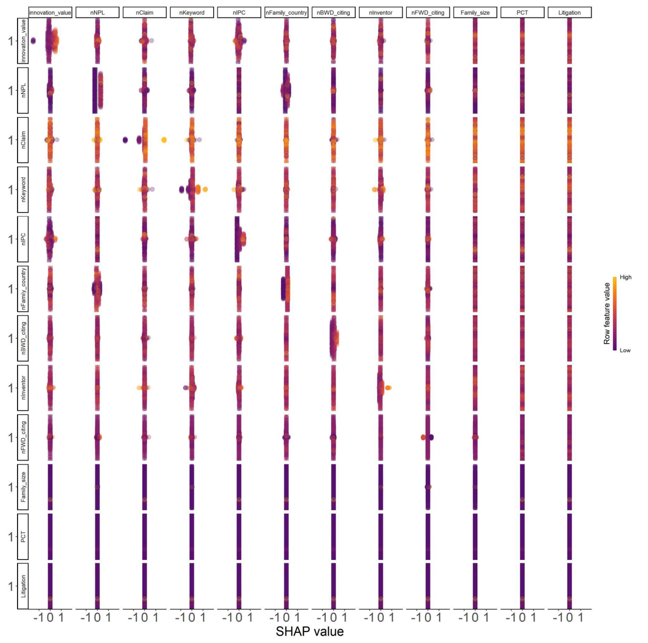


FIGURE 4. **SHAP dependence overview.**

than 0 indicates a synergistic effect between the two factors, i.e., the impact of a given factor on patent transfer is increased by the simultaneous action of another factor. Obtaining the interval at which the interaction value between two indicators is greater than zero can reveal the promotion space of the model by indicators of joint actions, thereby improving model interpretation. The SHAP dependence overview is presented in Figure 4.

Figure 4 shows there are interaction effects among the factors affecting university patent transferability, although the interaction effects reflected in the non-diagonal elements are lower than the main effects reflected in the diagonal elements. Among them, the innovation core value, claims, and core keywords have the strongest interactions with other factors, indicating that these three factors can effectively enhance patent transferability that is impacted by other factors under the interaction effect.

3) MODEL LOCAL INTERPRETATION

By studying the impact mechanism of a single indicator on university patent transfers we can further understand the conditions under which university patent transferability can be promoted. Nine indicators with relatively significant impacts on patent transfers are selected for model local interpretation, as visualized in Figure 5.

Figure 5 shows that an increase in the innovation core value, non-patent literature citations, and family countries causes the SHAP values to increase rapidly, rise above zero and maintain a strong positive influence on patent transfers. As for the number of claims and core keywords, the SHAP values increase slightly as the independent variable starts to increase, then maintains a relatively stable level near zero. When the number of claims exceeds eight and the number of keywords exceeds nineteen, the SHAP value shows an obvious upward trend indicating the independent variables have a significant positive impact on the outcome variables. An increase in the classification numbers leads to a fluctuating upward trend of the SHAP value, and an increase in the number of inventors and backward citations also leads to an increase in the SHAP value but the SHAP value has a slight downward trend after it reaches a certain level. It is worth noting that, in contrast to other indicators, the SHAP value

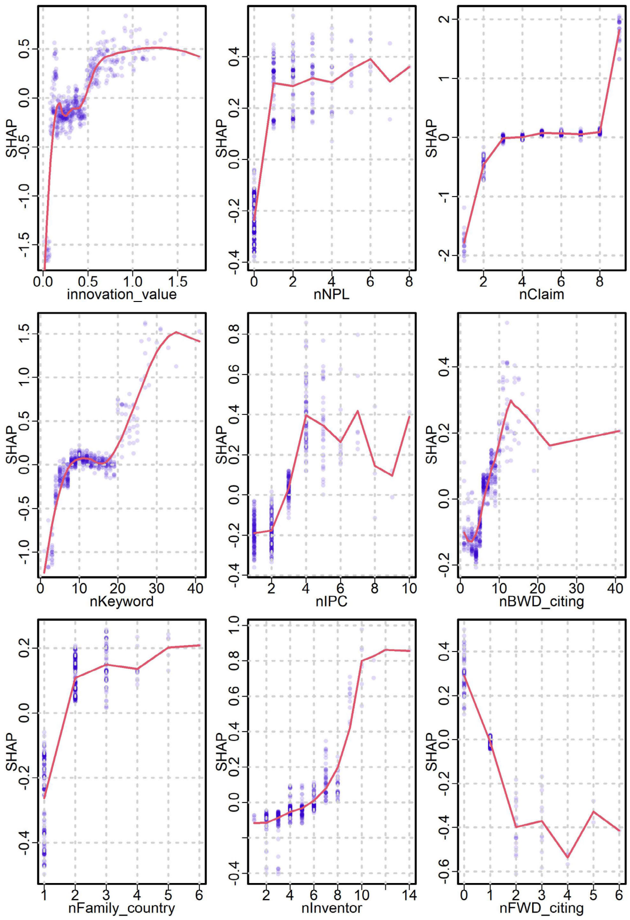


FIGURE 5. Local interpretations for SHAP.

shows an overall downward trend with an increase in the number of forward citations, which illustrates that forward citations have a certain negative effect on patent transfer. In general, the impact of indicators on patent transfers is nonlinear.

VI. DISCUSSION

Previous studies indicate that patent quality is considered the main factor affecting patent transfer [43], [44]. However, some high-quality patents are not necessarily easy to transfer [45], [46], and high-quality patents do not necessarily mean high invention quality [47], suggesting there are other factors that affect patent transferability. Technological innovation factor has been shown to have a certain impact on the patent conversion rate at a theoretical level, but no studies have quantified the impact of technological innovation factors on patent transferability empirically. To improve the body of research on patent transferability, this study identifies patent technology innovation indicators by applying the LDA model, integrates these technology innovation factor in researching patent transferability predictions, and considers the interaction effect between the technology innovation indicators and patent quality features.

We use the XGBoost, SVM, RF, and ANN machine learning algorithms to establish a patent transferability model before and after including the technological innovation factors, and the influence of each feature on patent transferability as well as the interactions between features based on the SHAP framework are explored. In terms of the model choice, the XGBoost model demonstrates the best prediction performance among the four models, and including the technological innovation factors improves the prediction performance for university patent transferability. In terms of variable importance, the main factors affecting university patents transferability are the innovation core value, non-patent literature citations, and claims. Technological innovation characteristics have a generally high and positive impact on university patent transferability. Although the number of core keywords has a moderate positive impact on patent transfer, its effect is relatively low compared to the innovation core value, indicating that there are more detailed technical solutions and key points in the research and development hotspots of each technology field, and patents that apply in a mainstream forward research direction are more likely to be transferred. From the perspective of variable interactions and local interpretations of the model, the impact of patent technology innovation features and quality features on university patent transferability is nonlinear and there is a positive interaction between them. This suggests university patent transferability can be promoted through the influence of this synergy.

Notably, the number of forward citations is the only factor that has a negative impact on patent transferability among all the influencing factors, whereas in other studies it has been shown to be independent of, or positively correlated with patent transfer [48], [49]. It is possible that although forward citations suggest a patent's technological impact and quality [50], [51], as technology is constantly updated the cited patent technology will gradually become obsolete relative to new technologies, and all technologies are at risk of being replaced by more advanced patented technologies during their lifespan. Moreover, forward citations reflect the social competition mechanism between private knowledge, which is mostly used for technology comparisons [52]. Patents may lose market share when being cited by more competitors [53]. Therefore, the more frequently a patent is cited, the more new technical challenges it is likely to face, which may reduce the likelihood of transfer.

VII. CONCLUSION

In this study we apply content mining and feature extraction of abstracts of patents granted in universities in the field of artificial intelligence. Using the LDA model we construct a patent transferability index system that combines of patent technological innovation features and quality features. We consider the SVM, RF, ANN and XGBoost algorithms for data training and prediction. By exploring the importance of features in patent transferability prediction as well as the interactions between innovation and quality features using the XGBoost model and SHAP framework, we can draw the following conclusions:

(1) A patent transferability model that includes factors related to technological innovation improves the prediction performance regarding university patent transferability, validated the ideas and methods proposed in this paper.

(2) Compared with the other machine learning models we evaluated, the XGBoost model has the best performance in predicting university patent transferability. The main factors affecting university patent transferability are the innovation core value, non-patent citations, and patent claims. The technological innovation features generally have a significant impact on university patent transferability.

(3) Patent technological innovation features and quality features have a nonlinear effect on the transferability of university patents, and there is a significant positive interaction between the two—in other words, improving either the level of technology innovation in patents or patent quality will enhance the effect of the other variable on patent transferability. University-based patents with both strong technological innovation features and high quality are more likely to be transferred.

This study enriches the indicator system for patent transferability research and provides guidance for universities seeking to apply for patents with higher transferability. It also provides a basis for enterprises to identify patents with a relatively high market value, which is conducive to promoting the transfer rate of university patents in practice. This research also has certain shortcomings. Limited by the scale of the study, the models constructed in this study were applied only to patent data in the field of artificial intelligence, so the conclusion has domain limitations. Besides, to ensure the accessibility of the index, only twelve patent evaluation indexes were selected based on related research. According to these deficiencies, a follow-up study could expand the patent transferability evaluation index based on the theoretical analysis and practical experience. In addition, the patent transferability prediction model that includes technological innovation factors could be extended to other technology-related fields to explore the generalizability of the patent transferability prediction model developed here, more universal conclusions are supposed to be drawn by comparing domain differences based on patent data in all fields.

REFERENCES

[1] C. Freeman and L. Soete, *The Economics of Industrial Innovation*, vol. 1. Cambridge, U.K.: MIT Press, 1982.

[2] Z. Dou, B. Ma, and H. Xu, "A study on university patent transformability based on the characteristics of inventors," *J. Intell.*, vol. 42, no. 5, pp. 169–174, May 2023, doi: 10.3969/j.issn.1002-1965.2023.05.023.

[3] N. Ko, B. Jeong, W. Seo, and J. Yoon, "A transferability evaluation model for intellectual property," *Comput. Ind. Eng.*, vol. 131, pp. 344–355, May 2019, doi: 10.1016/j.cie.2019.04.011.

[4] Y. Liu, "Path of patent quality and value improvement for the transformation of scientific and technological achievements," *Sci. Technol. Manag. Res.*, vol. 42, no. 8, pp. 158–164, Jun. 2023, doi: 10.3969/j.issn.1000-7695.2023.8.019.

[5] H. Sun and Z. Yang, "The strategic theoretical framework of patent technology innovation from the perspective of high-quality growth," *Sci. Manag. Res.*, vol. 39, no. 2, pp. 9–16, May 2021, doi: 10.19445/j.cnki.15-1103/g3.2021.02.002.

[6] B. Ribeiro and P. Shapira, "Private and public values of innovation: A patent analysis of synthetic biology," *Res. Policy*, vol. 49, no. 1, Feb. 2020, Art. no. 103875, doi: 10.1016/j.respol.2019.103875.

[7] A. B. Bonab, I. Rudko, and F. Bellin, "A review and a proposal about socio-economic impacts of artificial intelligence," in *Business Revolution in a Digital Era*, A. M. Dima and F. D'Ascenzo, Eds. Jan. 2021, pp. 251–270, doi: 10.1007/978-3-030-59972-0_18.

[8] M. Li, Z. Wu, L. Wang, and K. Zhou, "Does firm's value matter with firm's patent quality in technology-intensive industries?" *IEEE Trans. Eng. Manag.*, vol. 70, no. 4, pp. 1587–1604, Apr. 2023, doi: 10.1109/TEM.2020.3014228.

[9] K. Higham, G. de Rassenfosse, and A. B. Jaffe, "Patent quality: Towards a systematic framework for analysis and measurement," *Res. Policy*, vol. 50, no. 4, May 2021, Art. no. 104215, doi: 10.1016/j.respol.2021.104215.

[10] Y. Wu, Y. Ji, F. Gu, and J. Guo, "A collaborative evaluation method of the quality of patent scientific and technological resources," *World Pat. Inf.*, vol. 67, Dec. 2021, Art. no. 102074, doi: 10.1016/j.wpi.2021.102074.

[11] J. Lee, J.-H. Kang, S. Jun, H. Lim, D. Jang, and S. Park, "Ensemble modeling for sustainable technology transfer," *Sustainability*, vol. 10, no. 7, p. 2278, Jul. 2018, doi: 10.3390/su10072278.

[12] C. Ran and K. Song, "The construction of the university transferability patent recognition model: A case analysis of artificial intelligence," *Inf. Stud., Theory Appl.*, vol. 43, no. 11, pp. 79–85, Dec. 2020, doi: 10.16353/j.cnki.1000-7490.2020.11.013.

[13] B. Zhang, W. Hong, D. Gao, J. Li, and Z. Cui, "On university transferable patent recognition based on feature fusion," *J. Intell.*, vol. 41, no. 9, pp. 159–165, Sep. 2022, doi: 10.3969/j.issn.1002-1965.2022.09.023.

[14] Q. Zong, "The realistic obstacles of the transformation of scientific and technological achievements in universities and its cracking mechanism," *Sci. Technol. Prog. Policy*, vol. 40, no. 4, pp. 106–113, Mar. 2023, doi: 10.6049/kjjbydc.2021120729.

[15] W. Yang and Q. Sun, "Coupling drive research on key core patents and technology lock-in," *J. Intell.*, vol. 42, no. 3, pp. 102–109, Mar. 2023, doi: 10.3969/j.issn.1002-1965.2023.03.014.

[16] R. B. Yesilay and U. Halac, "An assessment of innovation efficiency in EECA countries using the DEA method," *Contemp. Issues Bus. Econ. Finance*, vol. 104, pp. 203–215, Sep. 2020, doi: 10.1108/S1569-375920200000104014.

[17] X. Zhao, M. Chen, C. Long, J. Zhao, and S. Wang, "Research on evaluation of scientific and technological innovation efficiency in Chinese universities—An analysis based on sample data of 31 provinces," *IEEE Access*, vol. 11, pp. 16297–16309, 2023, doi: 10.1109/ACCESS.2023.3241343.

[18] Q. Bao, H. Chai, S. Zhao, and J. Wang, "Model of technology opportunity mining using machine learning algorithm and its application," *J. Shanghai Jiaotong Univ.*, vol. 54, no. 7, pp. 705–717, Sep. 2020, doi: 10.16183/j.cnki.jsjtu.2020.99.007.

[19] F. Han, S. Zhang, L. Feng, and J. Yuan, "Identifying breakthrough patent topics by measuring technological convergence—Case study of solar PV domain," *Data Anal. Knowl. Discov.*, vol. 5, no. 12, pp. 137–147, Feb. 2022, doi: 10.11925/infotech.2096-3467.2021.0240.

[20] K. H. Kim, Y. J. Han, S. Lee, S. W. Cho, and C. Lee, "Text mining for patent analysis to forecast emerging technologies in wireless power transfer," *Sustainability*, vol. 11, no. 22, p. 6240, Nov. 2019, doi: 10.3390/su11226240.

[21] Y. Lu, Q. Mei, and C. Zhai, "Investigating task performance of probabilistic topic models: An empirical study of PLSA and LDA," *Inf. Retr.*, vol. 14, no. 2, pp. 178–203, Apr. 2011, doi: 10.1007/s10791-010-9141-9.

[22] L. E. George and L. Birla, "A study of topic modelling methods," in *Proc. 2nd Int. Conf. Intell. Comput. Control Syst. (ICICCS)*, Madurai, India, Mar. 2019, pp. 109–113, doi: 10.1109/ICCONS.2018.8663152.

[23] L. Ma, J. Wang, and H. Chen, "Business data security of system-wide information management based on content mining," *J. Comput. Appl.*, vol. 39, no. 2, pp. 488–493, Mar. 2019, doi: 10.11772/j.issn.1001-9081.2018071449.

[24] T. Kim, M. Min, T. Yoon, and J. Lee, "Semantic analysis of Twitter contents using PLSA, and LDA," in *Proc. SCIS & ISIS*, Okayama, Japan, Mar. 2012, pp. 1189–1922, doi: 10.14864/softscis.2010.0.189.0.

[25] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003, doi: 10.5555/944919.944937.
[26] Z. Miao, J. Du, F. Dong, Y. Liu, and X. Wang, "Identifying technology evolution pathways using topic variation detection based on patent data: A case study of 3D printing," *Futures*, vol. 118, Apr. 2020, Art. no. 102530, doi: 10.1016/j.futures.2020.102530.
[27] B. Yu, Y. Chen, and Y. Yang, "Research on evaluation of technical innovation based on topic model and patent data," *Mod. Inf.*, vol. 39, no. 1, pp. 111–117, Jan. 2019, doi: 10.3969/j.issn.1008-0821.2019.01.014.
[28] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, San Francisco, CA, USA, Aug. 2016, pp. 785–794, doi: 10.1145/2939672.2939785.
[29] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.
[30] V. Cherkassky, "The nature of statistical learning theory~," *IEEE Trans. Neural Netw.*, vol. 8, no. 6, p. 1564, Nov. 1997, doi: 10.1109/TNN.1997.641482.
[31] A. Nigrin, "Neural networks for pattern recognition," *Adv. Comput.*, vol. 37, pp. 119–166, Dec. 1993, doi: 10.1016/S0065-2458(08)60404-0.
[32] Z. H. Ali and A. M. Burhan, "Hybrid machine learning approach for construction cost estimation: An evaluation of extreme gradient boosting model," *Asian J. Civil Eng.*, vol. 24, no. 7, pp. 2427–2442, Nov. 2023, doi: 10.1007/s42107-023-00651-z.
[33] S. Lundberg and S. I. Leem, "A unified approach to interpreting model predictions," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, San Francisco, CA, USA, May 2017, pp. 4768–4777, doi: 10.48550/arXiv.1705.07874.
[34] L. S. Penrose, "A value for N-person games," *Contrib. Theory Games*, vol. 2, no. 28, pp. 307–317, Mar. 1952.
[35] S. B. Jabeur, S. Mefteh-Wali, and J. L. Viviani, "Forecasting gold price with the XGBoost algorithm and SHAP interaction values," *Ann. Oper. Res.*, Jul. 2021, doi: 10.1007/s10479-021-04187-w.
[36] D. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," *J. Mach. Learn. Technol.*, vol. 2, no. 1, pp. 37–63, Oct. 2020, doi: 10.48550/arXiv.2010.16061.
[37] Y. Song, J. Chang, and L. Chen, "Research on the impact of patent quality on enterprise performance- the moderating effect of technological innovation type," *Stud. Sci. Sci.*, vol. 39, no. 8, pp. 1459–1466, Aug. 2021, doi: 10.16192/j.cnki.1003-2053.20200915.007.
[38] Y. Sun and Q. Luan, "'Three stages-two dimensions' model of patent quality measuring and its empirical study: A case study of C9 league," *Sci. Manag.*, vol. 37, no. 6, pp. 23–32, Aug. 2016.
[39] F. Gäßler, "The timing of patent transfers in Europe," in *Enforcing and Trading Patents: Evidence for Europe*. Wiesbaden: Springer Fachmedien Wiesbaden, 2016, pp. 105–145, doi: 10.1007/978-3-658-13375-7_4.
[40] W. Du, Y. Wang, W. Xu, and J. Ma, "A personalized recommendation system for high-quality patent trading by leveraging hybrid patent analysis," *Scientometrics*, vol. 126, no. 12, pp. 9369–9391, Nov. 2021, doi: 10.1007/s11192-021-04180-x.
[41] Y. Li, H. Huijuan, X. U. Lijing, X. U. Mingkai, S. Gehao, and J. Xiuchen, "Transformer fault diagnosis with unbalanced samples based on neighbourhood component analysis and k-nearest neighbours," *High Voltage Eng.*, vol. 47, no. 2, pp. 472–479, Apr. 2021, doi: 10.13336/j.1003-6520.hve.20200373.
[42] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.
[43] Y. Sun, C. Zhang, and R. A. W. Kok, "The role of research outcome quality in the relationship between university research collaboration and technology transfer: Empirical results from China," *Scientometrics*, vol. 122, no. 2, pp. 1003–1026, Feb. 2020, doi: 10.1007/s11192-019-03330-6.
[44] H. Zhang, "To identify Chinese university patents' transferability: Indicators from the perspective of bibliometrics," *Data Sci. Inf.*, vol. 2, no. 1, pp. 26–41, Feb. 2022.
[45] J. Choi, D. Jang, S. Jun, and S. Park, "A predictive model of technology transfer using patent analysis," *Sustainability*, vol. 7, no. 12, pp. 16175–16195, Dec. 2015, doi: 10.3390/su71215809.
[46] T. Manu, "Challenging the validity of patents: Stepping in line with EPO and U.S. jurisprudence," *IIC Int. Rev. Intellectual Property Competition Law*, vol. 48, no. 7, pp. 813–837, Nov. 2017, doi: 10.1007/s40319-017-0633-7.
[47] P. Drahos, *The Global Governance of Knowledge: Patent Offices and Their Clients*. Cambridge, U.K.: Cambridge Univ. Press, May 2010, doi: 10.1017/CBO9780511676581.
[48] R. Li and J. Fan, "To study the relationship between university patents transformation with their bibliometrics characteristics," *Progr. Policy*, vol. 38, no. 17, pp. 43–50, Sep. 2021, doi: 10.6049/kjjbydc.2020080061.
[49] Z. Bao and S. Li, "Research on effectiveness of patent quality indicators from the perspective of technology transfer," *Tech. Intell. Eng.*, vol. 5, no. 5, pp. 37–45, Mar. 2020, doi: 10.3772/j.issn.2095-915x.2019.05.004.
[50] L. Aristodemou and F. Tietze, "Citations as a measure of technological impact: A review of forward citation-based measures," *World Pat. Inf.*, vol. 53, pp. 39–44, Jun. 2018, doi: 10.1016/j.wpi.2018.05.001.
[51] H. Lin, "Patent quality valuation with deep learning models," in *Database Systems for Advanced Applications* (Lecture Notes in Computer Science), J. Pei, Y. Manolopoulos, S. Sadiq, and J. Li, Eds. May 2018, pp. 474–490, doi: 10.1007/978-3-319-91458-9_29.
[52] R. Li and H. Zhang, "Contrastive analysis on expired patents' national differences and bibliometric characteristics," *J. China Soc. Sci. Tech. Inf.*, vol. 40, no. 12, pp. 1301–1311, Jan. 2022, doi: 10.3772/j.issn.1000-0135.2021.12.008.
[53] F. J. L. Bertran, "Patents, citations, and the market value of innovations," Ph.d. dissertation. Univ. Rochester, Rochester, NY, USA, Dec. 2005.

DISHA DENG received the bachelor's degree in accounting from the Hubei University of Technology, China, in 2017, and the master's degree in finance from McMaster University, Canada, in 2019. She is currently pursuing the Ph.D. degree in management with the Wuhan University of Science and Technology. Her current research interests include technology transfer, data mining, natural language processing, and machine learning.

TAO CHEN received the Ph.D. degree in management from the Huazhong University of Science and Technology, in 2009. He is currently a Professor with the Marketing Department, Wuhan University of Science and Technology. His main research interests include, but not limited to, technology transfer, new product diffusion, and supply chain management. His research has supported by the Hubei Industrial Policy and Management Research Center and Hubei Provincial Department of Education.