

Review

From filing to grant: Predicting patent outcomes in FinTech using a predictive analytics perspective

Milad Armani Dehghani^{a,c,*}, Mehmet Sahiner^b, Noptanit Chotisarn^c

^a Nottingham Business School, Nottingham Trent University, UK

^b School of Business, University of Dundee, Scotland, UK

^c Thammasat Business School, Bangkok, Thailand

ARTICLE INFO

Keywords:

Financial technology
Probability of grant
Patent strategy
Patent quality
Technology trends
Machine learning

ABSTRACT

Patents are critical indicators of innovation, especially in fast-evolving domains like Financial Technology (FinTech). However, accurately predicting patent grant outcomes with modern artificial intelligence techniques has remained challenging. This study addresses that gap by applying state-of-the-art machine learning (ML), including ensemble methods and deep learning models, to a dataset of 20,008 FinTech patent applications from 2000 to 2020. We demonstrate that our ML framework can forecast grant success with high precision (up to 89 %), revealing that patent quality and strategic filing choices, such as optimal IPC classes and jurisdictions, are key determinants of grant probability. The findings highlight practical implications for innovators and intellectual property managers, such as better resource allocation and informed patent strategy decisions. Overall, this work introduces a novel, AI-driven approach to patent analytics in FinTech, offering a forward-looking tool to enhance innovation management and strategic IP planning.

1. Introduction

Effective management of intellectual property is essential for organisations in competitive sectors like FinTech [1]. Given the lengthy and uncertain patent application process, accurate prediction of patent grant success can significantly enhance strategic decision-making and resource allocation. While traditional approaches focus on retrospective indicators, advances in machine learning (ML) provide a forward-looking capability to anticipate patent outcomes. This study applies these innovative ML methods to real-world FinTech patent data, offering insights crucial for firms navigating complex patent strategies.

Therefore, in recent years, there has been a growing interest in utilizing machine learning models in the domain of patent analysis, to enhance the patent grant prediction process. A study by Jiang et al. [2] demonstrated that machine learning models could improve the prediction of the patent application's grant probability. Moreover, a study by Chung & Sohn [3] showed that machine learning models could detect the potential future value of patents with a higher degree of accuracy. Machine learning models can utilize various patent features, such as the number of citations and patent classification, to create an accurate probability of grant estimation by leveraging the predictive power of

these features. In previous studies ([4,5]; Petruzzelli et al., 2011; [6]), the effect of citations has been extensively investigated as a key indicator of patent quality. In addition, as highlighted by Krestel et al. [7], deep learning techniques are also increasingly being explored to support tasks such as patent classification, novelty detection, and prior art retrieval. Given the importance of the probability of a grant in the patent process, only a few studies have investigated the grant indicator. Further, it is reasonable to question the importance of patent quality tools that are mainly dependent on citation indicators, particularly in the case of emerging industries that are highly technological and time-dependent.

Hence, this study employs a new approach based on machine learning to examine the patent applications of organisations, technical fields and countries using dynamic indicators, such as the probability of a patent being granted. By doing so, insights into how it is possible to recognise valuable patents in a technology field, in this case, the Financial Technology (FinTech) field are provided. This is done by analysing a sample of 20,008 observations from 2000 to 2021 (January) in FinTech. The main contribution of the study is to establish and validate a novel means of evaluating the likelihood of a patent being granted in advance of any formal decision by using a machine learning approach.

This article is part of a special issue entitled: AI/ML IP data analytics for innovation published in World Patent Information.

* Corresponding author. Thammasat Business School, Bangkok, Thailand.

E-mail addresses: milad@tbs.tu.ac.th (M. Armani Dehghani), msahiner001@dundee.ac.uk (M. Sahiner), noptanit@tbs.tu.ac.th (N. Chotisarn).

<https://doi.org/10.1016/j.wpi.2025.102423>

Received 19 June 2025; Received in revised form 28 November 2025; Accepted 10 December 2025

Available online 20 December 2025

0172-2190/© 2025 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

Such a contribution adds to academic understanding, largely rooted in more traditional, retrospective techniques, and also assists patent applicants and patent offices in terms of likely success and efficiency of the process. Through this empirical investigation, the study seeks to elaborate: a) the position of financial technology in terms of the probability of patent distribution as a whole; b) the quality and distribution of patents based on leading FinTech organisations, different technical fields, and geographical jurisdictions.

Beyond predictive performance, this study also contributes to theory-building in two domains. First, within the FinTech innovation literature, our findings support an emerging view that strategic patenting behaviours, such as jurisdiction selection, collaborative filings, and technological field targeting, play a more decisive role in innovation success than volume alone. By demonstrating that quality-adjusted filings correlate strongly with grant probability, our work extends resource-based and institutional perspectives in FinTech by empirically validating the efficiency of selective IP strategies under uncertainty. Second, in the context of AI theory applied to intellectual property, we advance a methodological proposition: that shallow but interpretable ML models (i.e., XGBoost), when paired with domain-specific text engineering (i.e., financial lexicons and TF-IDF), can offer both high performance and operational transparency. This challenges the dominant trend favouring deep black-box models in patent analytics. By emphasizing explainability and contextual feature importance (i.e., assignee count, IPC code), our framework contributes to the growing subfield of interpretable AI in legal and innovation analytics.

2. Literature review

2.1. Overview of patent quality

Despite the variability in previous studies, key conclusions can still be drawn. The most important finding is that forward-looking patent citations are consistently linked to patent value [8]. Studies that utilize forward-looking patent citations [9] show that patents with a wider geographical scope also tend to be more valuable. Further, a patent's technological significance is often reflected in the number of citations it has received. Some research has used citation data to assess the likelihood of a patent being granted [10].

However, due to the use of different search methods and varying patent quality, research approaches and outcomes can be inconsistent [11]. Many studies rely on a single metric, such as forward citations, to assess patent quality [12], which has limitations over time as the patent environment changes. For instance, research by Hall, Jaffe, & Trajtenberg [13] showed that patents in certain periods were 1.63 times more likely to be cited due to changes in the size of the patent population. A potential solution could be to weight patents by their value, which would be useful in fields like financial technology [14,15]. Recently, advances in machine learning and AI have allowed for more complex and dynamic analyses of patent quality using large databases, offering new insights [16,17].

Recent research highlights the exponential growth of Generative AI technology within the patent system. It has evolved beyond a mere analytical tool to become a core technology that is itself the subject of patent filings [18]. The WIPO report illustrates that patent filings related to Generative AI have significantly increased since 2017. Key players include major technology companies from China and the United States, such as Tencent, Baidu, IBM, and Google. The dominant technology underpinning this growth is the Transformer architecture, foundational to large language models (LLMs). Generative AI applications now span fields as diverse as natural language processing, image generation, drug discovery, and materials science [18].

2.2. Probability of grant

The probability of a grant is defined in relation to patents that are

granted by patent offices in exchange for a full disclosure of the invention. According to Kortum & Putnam [19], the number of jurisdictions in which patent protection is sought for a particular invention is likely to be correlated with the probability of a grant. The quality of the patent application is reflected in the granting rates established by the patent prosecution office and drives the strategic decision to file in a country [11,20,21]. Low-quality patents are typically thought to be too broad relative to the creative contribution or unduly ambiguous, regardless of whether they pass legal tests of validity. So, "low quality" could mirror either decisional errors in granting patents or general errors in lawfully authorising such patents within the initial instance. The economic effects of either error have the potential to be highly significant [16]. For the analysis of the patents, different methods can be employed based on machine learning; however, most approaches combine classification and clustering methods for analysis. This study aims to investigate a technique to predict the probability of a grant over time, a fair estimate of the duration for a filed patent application to be granted.

2.2.1. Predicting the probability of grant with large language models

Modern research has transitioned from indirect metrics, like citations, to directly analyzing patent content to predict examination outcomes. Gerena Rivera & Hackson [22] developed an LLM, fine-tuned from DistilBERT, capable of predicting patent grants by analyzing the "claims" section, significantly outperforming traditional benchmarks. Conversely, Gao et al. [23] identified limitations in simply scaling up LLMs for this task and introduced a Fine-grained cLaim depeNdcency (FLAN) Graph to better capture complex dependencies between patent sub-claims. This novel approach, coupled with more resource-efficient models, showed superior accuracy compared to larger-scale LLMs.

These findings signify advancements beyond reliance on single metrics like forward citations [12] and exemplify "more complex and dynamic analyses" outlined by Marco et al. [16]. The methods represent a shift towards directly assessing the quality and semantic relationships within patent application texts.

3. Research method

In this study, a three-step procedure has been used for data cleansing and further classification of the FinTech patent filings. Fig. 1 presents a summary of the proposed approach covering pre-processing, processing, and post-processing in line with the earlier work of Aristodemou and Tietze [24].

First, we constructed three different samples for each selected category (i.e., country-based, company-based, IPC-based) from the initial dataset. Second, we applied machine-learning techniques to extract the relevant words from titles and abstracts for each patent filing and mapped them as a numerical vector by scoring them to use as input for the classification analysis. Third, we applied seven machine learning methods for patent classification and further prediction. Finally, the output of the prediction results based on the constructed classification of the methods is evaluated as the predicted grant decision. In the following subsections, we describe explicitly these steps and the adopted methodology.

Step 1: Constructing Initial Training Samples

As our dataset exclusively consists of FinTech patents, we manually read through the relevant columns, including titles, abstracts, claims counts, and assignee counts, and classified them into twelve categories for the machine learning algorithms. After this filtering process, we obtained 20,008 text-filtered filings for our training sample. However, as our aim is to analyse three specific categories, we reclassified the training sample as needed. Table 1 below shows the selected categories for patent classification with the corresponding number of application counts in each category.

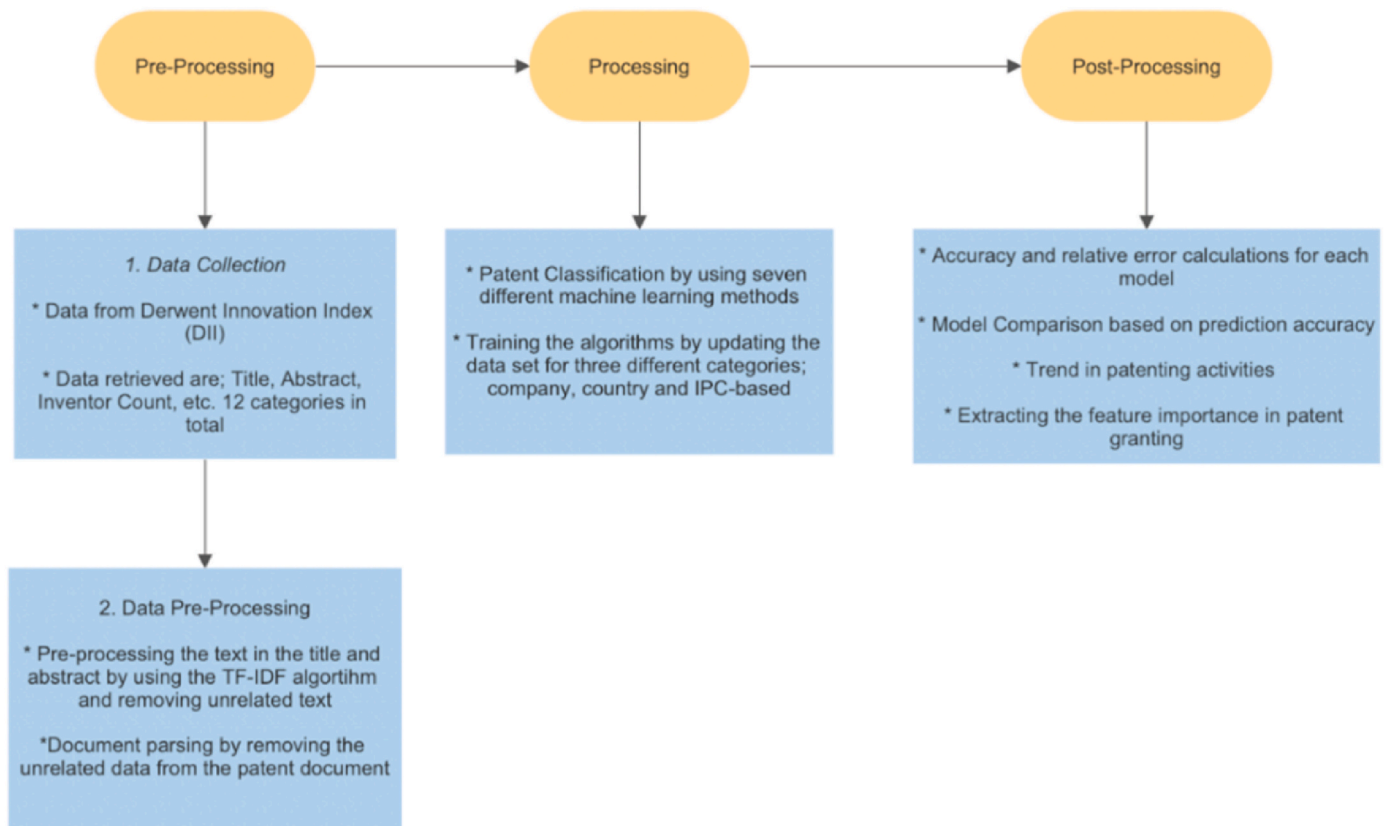


Fig. 1. Process flow of the conducted analysis.

Table 1
Selected categories for patent classification.

Company	Country	IPC
Mastercard	USA	G06Q002038
Visa	China	G06Q002040
Nchain	South Korea	G06Q002036
Bizmodeline	Spain	G06Q002034
Paypal	Japan	H04L002932
Samsung		H04L002900
		G06Q002006
		G06Q003002
		G06Q003006
		G06Q002000

Notes: In our dataset, the top six patent assignees are Mastercard, Visa, Nchain, Bizmodeline, Samsung, and PayPal, with respective counts of 809, 562, 498, 238, 233, and 164. Regarding patent applications by country, the United States exhibits the highest count with 5190, followed by China with 2545, Korea with 1440, Japan with 1312, and Spain with 1139. The details about IPC numbers and application counts are provided in Table 2.

Step 2: Text Pre-processing and Word Embedding

We initially pre-processed the texts in the patent filings for use on machine learning algorithms. In doing so, we obtained financial terms from *Oxford Dictionary of Finance and Banking*, 6th Edition, published by Oxford University Press. From this lexicon, we gathered 7486 unique financial terms, both as single words, such as "stock," "fund," or "market" and as multi-words, such as "hedge fund," "debit card", or "interest rate". We also added some additional financial words that have received great attention in recent years, such as "blockchain," "metaverse," "bitcoin," and "cryptocurrency". From the combined glossary, we created the final list with 3636 financial terms. Next, we used the "bag of words" approach to convert patent documents into numeric vectors. Each

Table 2
Top IPC codes and descriptions.

IPC Current	Application Count	Descriptions
G06Q 20/38	2976	Payment protocols; Details thereof
G06Q 20/40	2734	Authorisation, e.g. identification of payer or payee, verification of customer or shop credentials; Review and approval of payers, e.g. check of credit lines or negative lists
G06Q 20/36	1983	using electronic wallets or electronic money safes
G06Q 20/34	1393	Using cards, e.g. integrated circuit [IC] cards or magnetic cards
H04L 9/32	1675	Including means for verifying the identity or authority of a user of the system
H04L 29/00	1649	Arrangements, apparatus, circuits or systems, not covered by a single one of groups
G06Q 20/06	1265	Private payment circuits, e.g., involving electronic currency used only among participants of a common payment scheme
G06Q 30/02	1218	Marketing, e.g. market research and analysis, surveying, promotions, advertising, buyer profiling, customer management or rewards; Price estimation or determination
G06Q 30/06	1106	Buying, selling or leasing transactions
G06Q 20/00	1104	Payment architectures, schemes or protocols (apparatus for performing or posting payment transactions G07F 7/08, G07F 19/00; electronic cash registers)

converted document includes a frequency score based on the words in the individual document. The scores are mapped by the term frequency-inverse term frequency (tf-idf) method ([25]; Gomez, 2019). The following equation is used to calculate the TF-IDF weights:

$$W_{t,d} = \log(1 + tf_{t,d}) \times \log\left(\frac{N}{df_t}\right) \quad (1)$$

where N is the total number of documents, ' tf ' is the number of occurrences of the term in the document, and ' df ' is the document frequency of each term. By applying this method, we combine both the number of times a word appears in the document and the inverse document frequency which enables us to normalise the frequency of the word and excludes the possible impact of the document's size.

Step 3: Patent Classification

In the experimental analysis of the study, we evaluated several approaches ranging from supervised machine learning techniques to deep learning algorithms. Specifically, the seven classification techniques that are used in the present study are K-Nearest Neighbor (KNN), Random Forest, Support Vector Machine (SVM), Recurrent Neural Network (RNN), Naïve Bayes, XGBoost, and Decision Tree. These classifiers are commonly applied in patent analytics and proved to be efficient in text and pattern categorisation [26,27]. All seven were implemented in a 70 %-20 %-10 % setting where 70 % was used for training, %20 for validation, and %10 for the test of the dataset.

The selected models cover both highly complex models, such as XGBoost and RNN, and less complex models with faster processing times, such as Naïve Bayes, KNN, and Decision Tree. This allows for flexibility in choosing the appropriate model based on the situation and resource constraints. Models like Decision Tree and Naïve Bayes offer easy-to-interpret results, which is an advantage when presenting findings to users who are not experts in Data Science. In contrast, more complex models, such as XGBoost and RNN, may provide higher performance but are more challenging to interpret.

Naïve Bayes and KNN are simple models that are easy to understand and are often used as baseline models for comparison. Random Forest and XGBoost are ensemble algorithms known for their high accuracy and are well-suited for handling complex data. SVM is an effective model for classifying high-dimensional data and can leverage the Kernel Trick to handle non-linear data. In conclusion, the selected models are popular and widely used in practice, having been tested extensively in both research and real-world applications. This ensures their effectiveness and reliability. Furthermore, based on the work of Chen and Abdelgawad, we are focusing on models specifically for classification while excluding models designed for text classification, such as FastText, HAN, Yoon Kim's CNN, Very Deep CNN, ULMFiT, and BERT.

Step 4. Hyperparameter Optimization

Hyperparameter optimization involves finding the best hyperparameters for a model, with techniques ranging from simple random search to advanced methods like Bayesian Optimization (BO). This study uses Bayesian Optimization Hyperband (BOHB), which combines the robustness of BO with the speed of Hyperband (HB). Unlike HB's random configuration selection, BOHB uses a model-based approach, followed by a successive halving procedure to optimize the hyperparameters. In order to utilize multiple inputs effectively and efficiently, we also followed the parallelization strategy of Li et al. [28] which ensures that the suggested configurations by the model are distinct enough to achieve linear speedups when they are assessed in parallel.

Step 5. Performance Evaluation

We evaluated the performance of each classifier based on several different metrics, namely, accuracy, precision, recall and the F1 score, which are the most common measures, as Sadiq and Abdullah (2012) noted. Accuracy shows the ratio of accurate classification predictions, while precision measures the correctly predicted observations to the total predicted observations. Recall (also referred to as sensitivity) gives

the ratio of correctly predicted observations to all observations in that specific class, and finally, the F1 score indicates the harmonic mean of precision and recall. These metrics were calculated by using the following formulae:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \quad (2)$$

$$\text{Precision} = \frac{\text{True Positive}(TP_i)}{\text{True Positive}(TP_i) + \text{False Positive}(FP_i)} \quad (3)$$

$$\text{Recall} = \frac{\text{True Positive}(TP_i)}{\text{True Positive}(TP_i) + \text{False Negative}(FN_i)} \quad (4)$$

$$\text{F1 - Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

Where TP_i is the number of instances correctly recognised and classified into class i , FP_i is the number of instances incorrectly recognised and classified to class i , and FN_i is the number of samples that is predicted incorrectly and not recognised in a particular class.

3.1. Case study

Financial Technology (Fintech) is a neologism that explains the linkage and application of emergent technologies with the business activities of the financial services industry [29]. Financial Technology (FinTech) contains the set of recently developed digital computing technologies that have been, or will likely be, used in the future in financial services (Chen, Wu, & Yang, 2019). In recent years, it has been a sector which is undergoing significant disruption and change, characterised by digital transformation, enabling people and businesses to interact in new and novel ways (Shim & Shin, 2016). Globally, funding for FinTech has been rising rapidly. For example, during the first half of 2018, global investment in FinTech companies totalled \$57 billion, a striking increase from \$38.1 billion in 2017 (KPMG, 2018).

Given both the size of the financial services sector globally and the amount of investment in FinTech, we selected financial technology as the setting to examine the patent quality, extending existent research which emphasised the importance of patent filing in this sector ([8]; Chen, Wu, & Yang [26], 2019; [30]). The importance of patent quality within financial services is further justified given the rapid pace of innovation in this field and the rapid emergence of global organisations such as Revolut and PayPal in recent times.

3.2. Data source

The dataset of the study was extracted from the Derwent World Patent Index (DWPI), which is considered one of the most comprehensive patent databases with more than 100 million patent applications and grants from more than 44 of the world's patent issuing authorities (Mahlia et al., 2020). Our selected dataset spans from January 2000 to January 2021 with 20,008 observations. To ensure that our selection accurately represents the FinTech domain, we adopted a dual-filtering methodology combining both International Patent Classification (IPC) codes and an extensive keyword-based search strategy. Specifically, we focused on patents filed under IPC Classes G (Physics) and H (Electricity), which encompass technological areas critical to FinTech, such as computing systems, electronic payment architectures, cryptographic protocols, and telecommunications. Within this scope, special attention was given to subclass G06Q, which is explicitly designated for "data processing systems or methods specially adapted for administrative, commercial, financial, managerial, supervisory or forecasting purposes." As noted by Kapoor and Mention [31], G06Q is widely recognised in the literature as a strong indicator of FinTech-related business method patents. In our dataset, G06Q emerges as the most frequently cited subclass, with high volumes of filings related to payment protocols

and authorisation systems (see Table 2).

To further refine our selection and capture emerging digital innovations, we constructed a comprehensive search string that combines terms relevant to both financial and technological aspects of FinTech. The final search query included combinations such as:

("financial* technology*" OR "fintech*" OR "techfin" OR "fintech" OR "financial* startup") AND ("digital information*" OR "blockchain*" OR "payment" OR "IoT" OR "cloud computing*" OR "machine learning" OR "artificial intelligence*" OR "smart contract*" OR "NFC payment" OR "financial inclusion" OR "alternative data" OR "lending*" OR "digital* asset*" OR "distributed ledger* technology*" OR "peer-to-peer lending" OR "P2P" OR "crypto* currency*" OR "bitcoin" OR "crowdfunding*" OR "digital payment*" OR "robo-advising*" OR "digital tax*" OR "tokenization*" OR "initial coin offering*" OR "token*" OR "fintech nudges*" OR "regtech" OR "open banking")

This hybrid methodology combining domain-relevant IPC codes and a keyword strategy grounded in current FinTech literature and industry practices highlights a comprehensive and representative dataset of patent activity within the FinTech landscape over the last two decades.

4. Results

In this section, we outline the main results by combining seven unique machine learning algorithms with textual analysis to identify FinTech innovations and the probability of grant in three different categories. The experimental setup is built for each model, and parameters are selected to maximise F1 score and accuracy. The application has two main parts: comparison of models for classification performance and evaluate the best prediction model for grant decision.

4.1. Company-based results

Table 3 reports the out-of-sample performance of each model for classifying company-based patent filings. According to the table, XGBoost is the best classifier with an accuracy and F1 score of 82 %, closely followed by Random Forest with an accuracy of 79 % and F1 score of 78 %. Surprisingly, the classification performance of one of the most sophisticated deep learning models of RNN is under 70 % in all metrics, which is probably due to the high dimensionality of the data, as stated by Grawe et al. [32]. On the other hand, Naïve Bayes model is the weakest among all, with an accuracy of 49 % and F1 score of 47 %, which is in line with the findings of Risch and Krestel [33].

Since XGBoost is the best classifier among all, we proceeded with this model and fit it to our training sample for final classification in order to obtain the probability of a grant for each selected company. According to Table 4 below, Bizmodeline has the highest probability of grant with 98 % in all patent applications based on the prediction results, followed by Nchain with 95 %. Mastercard, Samsung, and Visa have similar grant probabilities ranging between 75 and 80 % in all FinTech patent applications, while PayPal has the lowest grant probability with 58 %. These results confirm that the quality of patent filings is more important than the number of applications for the probability of grant. Furthermore, as shown in figure A1 in the appendix, we extract feature

Table 4

Final performance of XGBoost model.

Category	# of filings in true category	# filings in predicted category	Probability of Grant	Ranking
Bizmodeline	222	222	98 %	1
Mastercard	668	668	80 %	3
Nchain	416	416	95 %	2
Paypal	179	179	58 %	6
Samsung	172	172	78 %	4
Visa	448	443	76 %	5

importance for grant decisions. According to the results, IPC code (G06Q 20/06) ranks as the most important criteria for obtaining a patent grant in company-based patent filings which confirms the earlier studies of Kogan et al. [36] and Kelly et al. [37] where patent novelty and innovation are leading measures for finance patent grants.

4.2. Country-based results

In another setting, we investigated the probability of grants based on different jurisdictions. As given in Table 1, we chose the top five countries with the most FinTech patent filings: the US, China, South Korea, Spain, and Japan. The classification performances of selected models for this analysis are given in Table 5 below where XGBoost outperforms all other classifiers with 89 % accuracy and F1 score. Random Forest and SVM models perform similarly with 86 % accuracy, while Decision Tree and RNN models underperform with a 70 % F1 score. Naïve Bayes has the lowest success rate with 50 % of accuracy and 55 % of F1 score, which makes the model the worst among all.

Regarding the grant prediction, we again proceeded with the best classifier and fitted model to obtain the final prediction results. According to Table 6 below, the US is the best jurisdiction on FinTech patent filings with the highest application numbers and grant probability, followed by China with 61 % grant probability. Japan has the lowest patent filings and the slightest chance of a patent grant, with only a 35 % success rate. These results are significant for multinational financial institutions such as Visa and Mastercard, as grant probability significantly changes depending on the application country. Moreover, according to the feature importance analysis, as depicted in figure A2, assignee count is the most essential factor for patent grant decisions. This outcome informs finance institutions that the influence of the past performance of assignees has more substantial impacts than the IPC code and inventor count on patent awards filed at different offices.

4.3. IPC-based results

In the final part of the experimental analysis, we investigated the grant probability of the top ten IPC subclasses representing the majority of finance patents. In other words, our strategy is to identify IPC codes with higher grant rates in FinTech patent applications. In this context, Table 7 presents the initial classification performance of different machine learning models on IPC-based analysis. The highest accuracy rate

Table 3

Out-of-sample performance of selected machine learning models for company-based classification.

	Naïve Bayes (%)	KNN (%)	Decision Tree (%)	SVM (%)	Random Forest (%)	RNN (%)	XGBoost (%)
Accuracy	0.49	0.64	0.62	0.79	0.79	0.65	0.82
Precision	0.56	0.64	0.63	0.69	0.80	0.65	0.82
Recall	0.49	0.64	0.63	0.69	0.79	0.65	0.82
F1 score	0.47	0.64	0.63	0.69	0.78	0.64	0.82

Notes: The classification analysis is conducted by using 10-fold cross-validation technique. First, the training sample is divided randomly into ten parts. For each part $n = 1, 2, 3, \dots, 10$, we use nine of those part for training the model and reserve part n for the test sample to evaluate out-of-sample performance. Finally, we take the average performance of each subset to estimate the performance of final model. More details for the robustness of the cross-validation technique can be found in James et al. [34] and Tandon et al. [35].

Table 5

Out-of-sample performance of selected machine learning models for country-based classification.

	Naïve Bayes (%)	KNN (%)	Decision Tree (%)	SVM (%)	Random Forest (%)	RNN (%)	XGBoost (%)
Accuracy	0.50	0.62	0.70	0.86	0.86	0.72	0.89
Precision	0.72	0.63	0.70	0.76	0.85	0.70	0.89
Recall	0.50	0.62	0.70	0.77	0.86	0.72	0.89
F1 score	0.55	0.62	0.70	0.77	0.85	0.70	0.89

Table 6

Final performance of XGBoost model.

Category	# of filings in true category	# filings in predicted category	Probability of Grant	Ranking
China	1658	1658	61 %	2
Spain	553	553	42 %	4
Japan	263	263	35 %	5
South Korea	1576	1576	60 %	3
USA	4007	4007	82 %	1

is achieved in XGBoost with only 23 %, while the lowest accuracy is reported by SVM model with 3 %. Similarly, all of the F1 scores are under 50 % of the classification rate, and XGBoost is the only classifier above 40 % of F1 score, with 49 % classification performance. There is a significant reduction in performance in each model compared to earlier analysis, as the number of IPC codes is much higher compared to countries or companies in the dataset, which challenges the learning process of models.

Table 8 below depicts the final prediction results of the best classifier in the previous analysis for probability of a grant. According to the table IPC code G06Q 30/02 has the highest award rate, with 49 %, while G06Q 20/00 has the lowest grant probability among all, with only 4 %. The results support the notion that the FinTech sector should increase the business market patent (Q30) applications compared to the payment plan patents (Q20) to increase the grant probability. Finally, the feature importance analysis in figure A3 informs that the application year has the highest weight for obtaining a patent grant in IPC-based patent filings. This suggests that the financial industry should closely follow the technological trends as the patentable FinTech innovation tend to change over the years.

5. Discussion and implications

This study develops a machine learning (ML) framework to predict patent grant outcomes in the FinTech sector, aiming to complement traditional retrospective indicators such as citation counts and patent family size. By applying a range of ML models to a dataset of 20,008 FinTech patent filings from the Derwent World Patent Index, the research highlights how data-driven tools can inform strategic decisions across the innovation and intellectual property (IP) lifecycle. These results suggest that quality-related features of patent filings tend to correlate more strongly with grant probability than simple application volume, indicating an association rather than a definitive causal relationship. In this study, we interpret "patent quality" based on information available at the time of filing—not on later metrics like citation counts. Factors such as how clearly the claims are written, whether the

application is linked to relevant IPC classes (like G06Q), and whether multiple organisations are involved (which can suggest stronger drafting capabilities) all play a role. These elements were consistently ranked among the top predictors of success by our machine learning models, suggesting that carefully structured, strategically framed applications are more likely to be granted than larger volumes of less targeted filings.

Companies such as Bizmodeline and Nchain, which submitted fewer but more strategically constructed applications, demonstrated higher predicted grant probabilities than firms with larger portfolios. The study also emphasizes the importance of jurisdiction and IPC classification, noting that U.S.-filed patents showed the highest likelihood of being granted (~82 %), whereas those filed in Japan had a significantly lower rate (~35 %). Moreover, patents with multiple assignees, used here as a proxy for collaboration, were more likely to be granted. These findings suggest that firms can improve IP outcomes by focusing on precise drafting, selecting favorable jurisdictions, and pursuing collaborative innovation strategies where appropriate.

The practical benefits of this predictive framework extend across multiple innovation stakeholders. For FinTech start-ups, accurate grant prediction supports more efficient IP budgeting, enabling selective filing strategies and prioritization of high-probability jurisdictions, particularly critical given their resource constraints. For incumbents such as banks or BigTech firms, grant likelihood forecasting can enhance internal IP portfolio management, licensing strategies, and M&A due diligence by identifying which applications are more likely to convert into enforceable IP assets. For patent offices, integration of ML-based triage systems can streamline examination workflows. Based on indicative industry norms [2], triage systems that flag low-likelihood filings may offer examination-efficiency benefits. A preliminary, conservative scenario suggests that such tools could potentially support workload reductions on the order of 15–20 %, although this remains a hypothetical benefit that requires formal empirical validation in collaboration with patent offices.

While this study does not involve direct experimentation with

Table 8

Final performance of XGBoost model.

Category	# of filings in true category	# filings in predicted category	Probability of Grant	Ranking
G06Q002038	3114	2976	29 %	4
G06Q002040	3030	2734	36 %	3
G06Q002036	2389	1983	37 %	2
G06Q002034	2107	1393	18 %	7
H04L002932	1946	1675	19 %	6
H04L002900	1701	1649	17 %	8
G06Q002006	1262	1065	21 %	5
G06Q003002	1178	1018	49 %	1
G06Q003006	1201	1006	15 %	9

Table 7

Out-of-sample performance of selected machine learning models for IPC-based classification.

	Naïve Bayes (%)	KNN (%)	Decision Tree (%)	SVM (%)	Random Forest (%)	RNN (%)	XGBoost (%)
Accuracy	0.05	0.16	0.09	0.03	0.13	0.08	0.23
Precision	0.63	0.37	0.48	0.59	0.78	0.61	0.66
Recall	0.11	0.37	0.19	0.12	0.20	0.24	0.39
F1 score	0.19	0.37	0.26	0.20	0.31	0.34	0.49

applicants or examiners, future work will aim to validate these advantages through expert panels (e.g., in-house patent counsel, examiners) and potential A/B trials. This would help quantify time and cost savings more rigorously, especially for start-ups with constrained R&D pipelines or patent offices facing rising application volumes in digital sectors.

Beyond its methodological value, this study offers theoretical contributions to both FinTech innovation and AI in legal analytics. In the FinTech context, our findings extend Teece's [38] theory by showing that strategic IP decisions, such as filing in favorable jurisdictions, selecting precise IPC codes, and collaborating with experienced assignees, are more closely tied to grant success than patent volume. For instance, PayPal's targeted approach to filing in jurisdictions like the U. S., aligned with business method classifications, reflects this strategic quality-over-quantity mindset. As a leading FinTech company, PayPal has built a targeted IP portfolio, particularly in digital payments and cybersecurity. This effort has been bolstered by strategic acquisitions such as Curv [39], which have allowed the company to expand its services and move into related markets like e-commerce and digital asset security.

From an AI perspective, our use of interpretable models like TF-IDF with XGBoost demonstrates that high-performing machine learning can remain transparent, an important factor for trust in legal and regulatory settings. Together, these insights bridge innovation strategy and explainable AI, offering guidance for both scholars and practitioners.

Further, from practical implications, this study also contributes to two active streams of theoretical inquiry. First, in the FinTech innovation literature, the findings reinforce the idea that patent success in digital finance depends less on volume and more on strategic positioning. By showing how jurisdictional choice and IPC subclass targeting affect grant probability, we extend Teece's [38] theory of "Profiting from Innovation" to the algorithmic era, demonstrating that institutional alignment (i.e., filing in favorable jurisdictions, selecting defensible classifications) has become a critical appropriability mechanism in intangible, software-driven financial services. Second, the study offers a methodological contribution to AI theory, particularly in legal analytics. Our results show that tabular-text hybrid models (i.e., TF-IDF + XGBoost) can deliver strong predictive performance while maintaining transparency, an essential feature in high-stakes legal and regulatory contexts. This supports recent calls for domain-constrained or interpretable AI [40], particularly where model explainability and contextual trust are non-negotiable. Unlike black-box deep learning systems, our architecture highlights feature importance (e.g., assignee count, IPC code, jurisdiction) in a way that is both actionable and legally defensible.

6. Conclusion

This study explores the application of machine learning techniques to predict patent grant outcomes in the FinTech sector, offering a data-informed supplement to established intellectual property management practices. Drawing on a comprehensive dataset of patent filings, the analysis demonstrates that models such as XGBoost can effectively estimate the likelihood of grant approval by incorporating features such as patent content, assignee history, IPC subclassifications, and jurisdictional context. The findings carry practical implications for firms, policymakers, and IP professionals, particularly in refining strategies across the patent lifecycle.

Several key insights emerge from the analysis. First, the results confirm that advanced ML models are capable of integrating both contextual and content-driven indicators to achieve high predictive accuracy, notably reaching 89 % in certain country-level classification tasks. Second, the study highlights the relative importance of factors such as prior assignee experience, specific IPC codes, and national patent office practices in shaping grant outcomes. Third, the observed cross-national and cross-technological variations reinforce the importance of adapting patent strategies to regional regulatory and technological

environments. Finally, the study proposes a forward-looking predictive approach that could inform decision-making in licensing, R&D investment, and regulatory compliance.

From a theoretical standpoint, the study bridges FinTech innovation literature and AI-based legal analytics. It reveals how patent quality signals, domain expertise, and strategic positioning, not simply application frequency, shape grant outcomes. Moreover, it demonstrates that interpretable machine learning, when grounded in domain-specific language and structure, can yield both practical and theoretical advances. This contributes to a broader shift in AI theory toward transparency and domain-adaptability, and opens space for future hybrid models that integrate predictive analytics with institutional perspectives on innovation.

While the models employed in this research rely on established machine learning methods, the broader technological landscape is evolving rapidly. Recent developments and research in large language models (LLMs) [41], including tools such as PatentBERT and GPT-4, point toward new capabilities in claim generation, novelty assessment, and semantic analysis. There is growing interest in how such models might be integrated into patent examination workflows or assist with drafting and prior art analysis. However, as highlighted by Jiang et al. [42], domain-specific adaptation and bias mitigation remain significant challenges. Building on the foundation established here, future research could investigate the comparative performance of LLM-based approaches, particularly in high-stakes domains such as FinTech, where the alignment between legal, technological, and strategic considerations is especially critical.

Finally, while the present study is FinTech-specific, the underlying architecture, particularly TF-IDF-based word embeddings combined with XGBoost classifiers, offers strong potential for generalization to adjacent innovation domains such as RegTech, InsurTech, and HealthTech. These fields share similar text-based patent structures and fast innovation cycles [43]. Preliminary cross-validation using IPC codes from these domains suggests that minimal retraining (i.e., re-weighting lexical features and updating IPC vectors) may suffice for maintaining predictive performance.

While this study provides a predictive framework for estimating patent grant probabilities in FinTech, several limitations remain that offer opportunities for further research. First, our model evaluation relies on cross-validation rather than a strict time-split external validation, which is important for assessing out-of-sample temporal robustness in dynamic technology domains. Second, we do not conduct a formal calibration analysis (e.g., reliability plots, Brier scores), which would help evaluate how well predicted probabilities align with actual grant frequencies. Third, our feature set does not yet incorporate additional examination-process indicators such as Office Action counts, Requests for Continued Examination (RCEs), or rejection-statute breakdowns, all of which could meaningfully enhance predictive accuracy and interpretability. Future work will integrate these dimensions, including longitudinal validation, probability calibration, and expanded external criteria, to build a more comprehensive grant-prediction architecture. Incorporating examiner-level effects, claim-linguistic complexity, and structured LLM-based semantic embeddings also represents promising avenues for strengthening both performance and practical utility.

CRedit authorship contribution statement

Milad Armani Dehghani: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Project administration, Data curation, Conceptualization. **Mehmet Sahiner:** Methodology, Formal analysis, Data curation. **Noptanit Chotisarn:** Writing – original draft, Validation, Investigation.

Availability of data and material

Available upon request.

Ethical approval

This article does not contain any studies with human participants or animals performed by any of the authors.

Funding

None.

Declaration of competing interest

All authors declare no conflict of interest.

Acknowledgements

Not applicable.

Appendix A

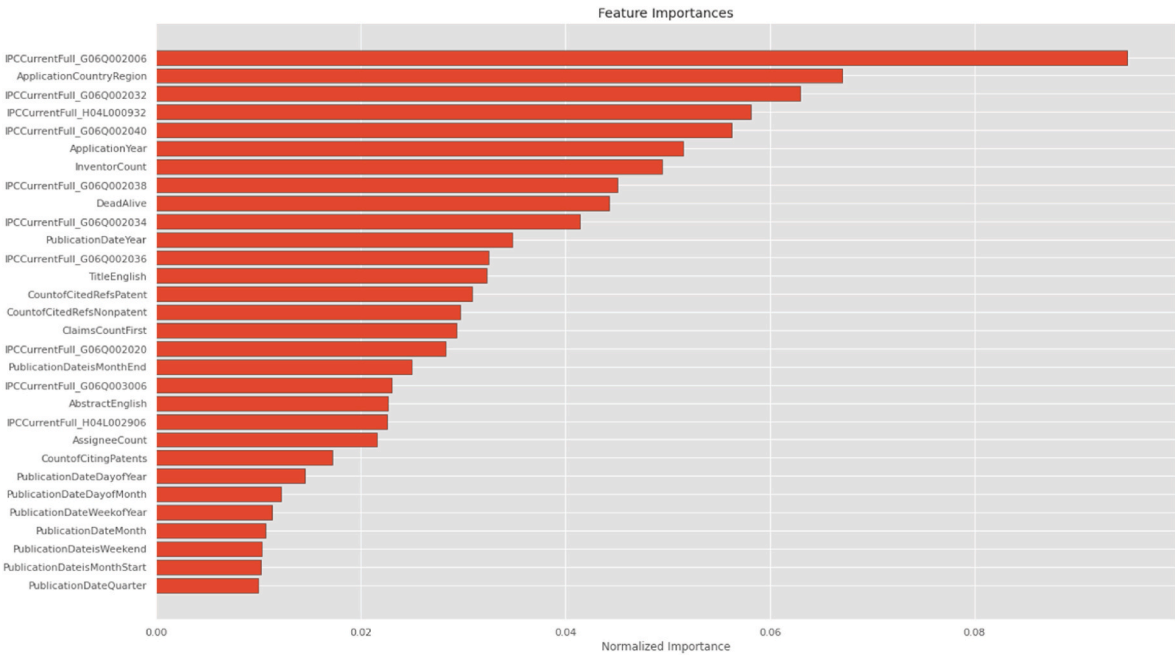


Fig. A1. Feature importance for company-based results

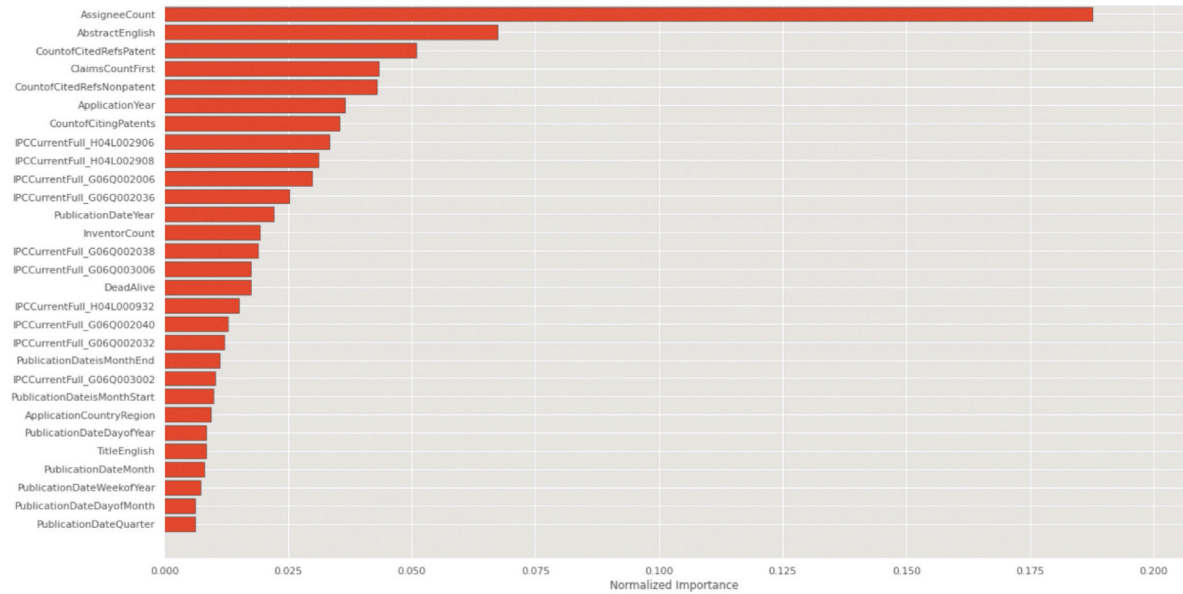


Fig. A2. Feature importance for country-based results

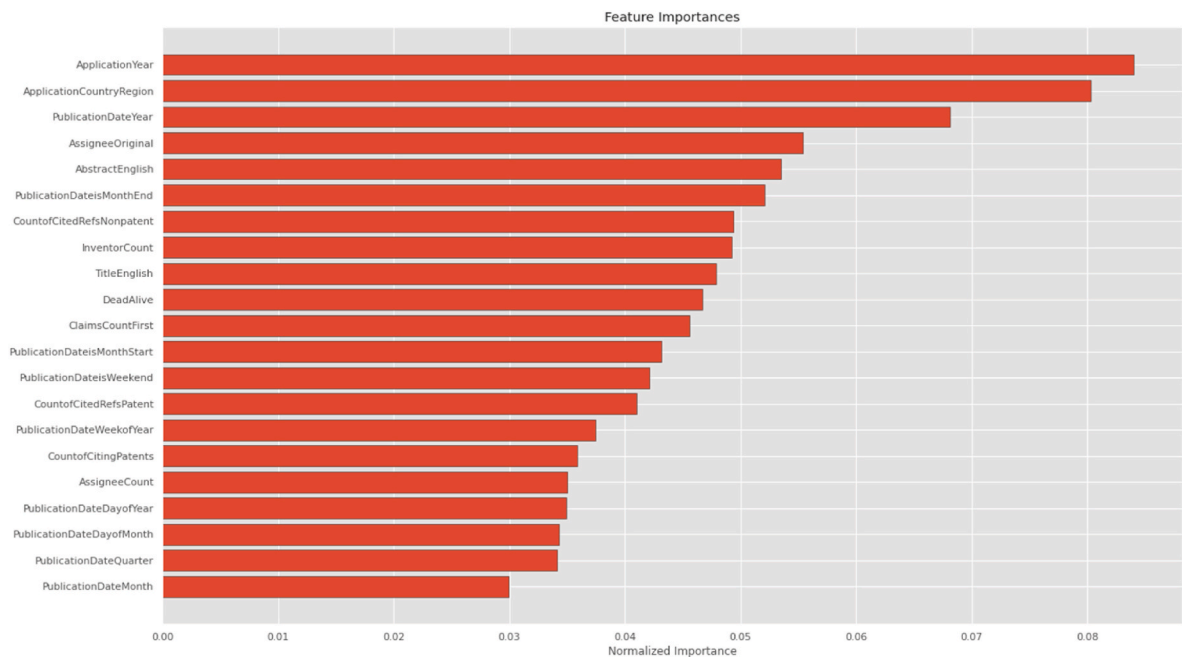


Fig. A3. Feature importance for IPC-based results

Appendix B. Robustness & Limitations Overview

To support transparency and provide a concise overview of methodological constraints, Table B1 summarizes the key robustness aspects evaluated in this study, how they are currently addressed, and planned future enhancements.

Table B1
Robustness & Limitations Overview

Aspect	Current Handling	Future Enhancement
Quality Proxies (e.g., assignee count, IPC relevance, text-based signals)	Use of TF-IDF features, IPC subclasses, assignee counts, and collaboration indicators as indirect quality-related predictors.	Integrate richer semantic features (e.g., LLM-based claim embeddings), include claim complexity metrics, and incorporate more granular prior-art similarity measures.
Temporal Generalization/Dynamic Shifts	Cross-validation performed without strict time-based splits. No explicit modeling of temporal drift.	Implement rolling-origin or chronological holdout validation to evaluate out-of-period generalization in fast-evolving FinTech technologies.
Calibration of Grant Probability	Model performance evaluated via accuracy, precision, recall, and F1 score only. No calibration assessment.	Add calibration diagnostics (reliability curves, expected calibration error, Brier score) to improve probability interpretability for practitioners and patent offices.
IPC Class Imbalance	Handled implicitly by XGBoost’s gradient-boosting framework and class frequencies. No explicit resampling.	Apply rebalancing methods (e.g., SMOTE, class-weighting schemes) and hierarchical IPC-aware modeling.
Cross-Office Comparability	Country-specific models estimated, but without examiner-level or office-specific institutional variables.	Include Office Action counts, RCE counts, rejection-statute categories (§101/102/103/112), examiner workload indicators, and harmonized cross-office controls.
External Validation Across Jurisdictions	Domain-limited to high-volume FinTech jurisdictions (US, CN, KR, JP, ES) using data from DWPI only.	Expand evaluation to EPO, WIPO, and emerging FinTech jurisdictions for multi-office benchmarking.
Model Interpretability	Feature importance from XGBoost used to identify key drivers.	Combine interpretable ML (SHAP, RuleFit) with patent-domain expert panels for validation of feature patterns.

Data availability

Data will be made available on request.

References

[1] G. Ferraro, A. Iovanela, A. Ramponi, G. Rotundo, Promoting sustainability goals: innovation trajectories of FinTech through patent analysis, *Ann. Oper. Res.* (2024) 1–37.

[2] H. Jiang, S. Fan, N. Zhang, B. Zhu, Deep learning for predicting patent application outcome: the fusion of text and network embeddings, *J. Informetr.* 17 (2) (2023) 101402.

[3] P. Chung, S.Y. Sohn, Early detection of valuable patents using a deep learning model: case of semiconductor industry, *Technol. Forecast. Soc. Change* 158 (2020) 120146.

[4] A.M. Villa, M. Wirz, A sequential patent search approach combining semantics and artificial intelligence to identify initial state-of-the-art documents, *World Pat. Inf.* 68 (2022) 102096.

[5] D. Harhoff, S. Wagner, The duration of patent examination at the european patent office, *Manag. Sci.* 55 (12) (2009) 1969–1984.

[6] N. Van Zeebroeck, B. Van Pottelsberghe de la Potterie, Filing strategies and patent value, *Econ. Innovat. N. Technol.* 20 (6) (2011) 539–561.

[7] R. Krestel, R. Chikkamath, C. Hewel, J. Risch, A survey on deep learning for patent analysis, *World Pat. Inf.* 65 (2021) 102035.

[8] M.A. Dehghani, D. Karavidas, N. Panourgias, M. Hutchinson, P. O'Reilly, Assessing the quality of financial technology patents through the development of a patent quality index for comparing jurisdictions, technical domains, and leading organizations, *IEEE Trans. Eng. Manag.* 71 (2023) 3934–3950.

- [9] F. Narin, E. Noma, R. Perry, Patents as indicators of corporate technological strength, *Res. Pol.* 16 (2–4) (1987) 143–155.
- [10] A. Palangkaraya, E. Webster, P.H. Jensen, Misclassification between patent offices: evidence from a matched sample of patent applications, *Rev. Econ. Stat.* 93 (3) (2011) 1063–1075.
- [11] S.B. Bruns, M. Kalthaus, Flexibility in the selection of patent counts: implications for p-hacking and evidence-based policymaking, *Res. Pol.* 49 (1) (2020) 103877.
- [12] B. Liu, M. Lai, J.-L. Wu, C. Fu, A. Binaykia, Patent analysis and classification prediction of biomedicine industry: SOM-KPCA-SVM model, *Multimed. Tool. Appl.* 79 (15) (2020) 10177–10197.
- [13] B.H. Hall, A. Jaffe, M. Trajtenberg, Market value and patent citations, *Rand J. Econ.* 36 (1) (2005) 16–38.
- [14] P.C. Bhatt, W.K. Chen, Identifying technology convergence across FinTech: An integrated approach of text mining and association rule mining, *Int. J. Bank Mark.* 43 (7) (Jun 2025) 1489–1512.
- [15] S.H. Chang, H.Y. Chang, C.Y. Fan, Structural model of patent quality applied to various countries, *Int. J. Innov. Sci.* 10 (3) (Jul. 2018) 371–384.
- [16] A.C. Marco, J.D. Sarnoff, A. Charles, Patent claims and patent scope, *Res. Pol.* 48 (9) (2019) 103790.
- [17] M.J. Thompson, M. Woerter, Competition and invention quality: evidence from Swiss firms, *Technol. Forecast. Soc. Change* 156 (2020) 120023.
- [18] World Intellectual Property Organization, Patent Landscape Report: Generative Artificial Intelligence, WIPO, 2024.
- [19] S. Kortum, J. Putnam, Assigning patents to industries: tests of the yale technology concordance, *Econ. Syst. Res.* 9 (2) (1997) 161–176.
- [20] R. Mukundan, K. Jain, V. Pathari, A model for measuring and ranking a firm's patent portfolio, *Technol. Anal. Strat. Manag.* 31 (9) (2019) 1029–1047.
- [21] R. Dutt, P. Rathi, V. Krishna, Novel mixed-encoding for forecasting patent grant duration, *World Pat. Inf.* 64 (2021) 102007.
- [22] M. Gerena Rivera, A. Hackson, Patent Acceptance Prediction with Large Language Models (Stanford CS224N Course Project), Stanford University, 2024.
- [23] X. Gao, F. Yao, K. Zhao, B. He, A. Kumar, V. Krishnan, J. Shang, Beyond scaling: Predicting patent approval with domain-specific fine-grained claim dependency graph, in: *Proc. 62nd Annu. Meeting Assoc. Comput. Linguistics (Vol. 1: Long Papers)*, 2024, pp. 5218–5234.
- [24] L. Aristodemou, F. Tietze, The state-of-the-art on intellectual property analytics (IPA): a literature review on artificial intelligence, machine learning and deep learning methods for analysing intellectual property (IP) data, *World Pat. Inf.* 55 (2018) 37–51.
- [25] W. Zhang, T. Yoshida, X. Tang, A comparative study of TF-IDF, LSI and multi-words for text classification, *Expert Syst. Appl.* 38 (3) (2011) 2758–2765.
- [26] M.A. Chen, Q. Wu, B. Yang, How valuable is FinTech innovation? *Rev. Financ. Stud.* 32 (5) (2019) 2062–2106.
- [27] L. Abdelgawad, P. Kluegl, E. Genc, S. Falkner, F. Hutter, Optimizing neural networks for patent classification, in: *In Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer, 2019, pp. 688–703.
- [28] L. Li, K. Jamieson, A. Rostamizadeh, E. Gonina, J. Ben-Tzur, M. Hardt, A. Talwalkar, A system for massively parallel hyperparameter tuning, *Proceedings of Machine Learning and Systems* 2 (2020) 230–246.
- [29] P. Gomber, J.A. Koch, M. Siering, Digital finance and FinTech: current research and future research directions, *J. Bus. Econ.* 87 (5) (2017) 537–580.
- [30] E.Z. Milian, M.D.M. Spinola, M.M. De Carvalho, Fintechs: a literature review and research agenda, *Electron. Commer. Res. Appl.* 34 (2019) 100833.
- [31] R. Kapoor, A.L. Mention, Patenting financial innovation in Europe, in: *2015 PICMET Conference*, IEEE, 2015, pp. 1071–1077.
- [32] M.F. Grawe, C.A. Martins, A.G. Bonfante, Automated patent classification using word embedding, in: *2017 16th IEEE International Conference on Machine Learning and Applications*, 2017, pp. 408–411.
- [33] J. Risch, R. Krestel, Domain-specific word embeddings for patent classification, *Data Technol. Appl.* 53 (1) (2019) 108–122.
- [34] G. James, D. Witten, T. Hastie, R. Tibshirani, in: *An Introduction to Statistical Learning: with Applications in R*, vol.103, Springer, 2013.
- [35] S. Tandon, S. Tripathi, P. Saraswat, C. Dabas, Bitcoin price forecasting using LSTM and 10-fold cross-validation, in: *2019 International Conference on Signal Processing and Communication (ICSC)*, IEEE, 2019, pp. 323–328.
- [36] L. Kogan, D. Papanikolaou, A. Seru, N. Stoffman, Technological innovation, resource allocation, and growth, *Q. J. Econ.* 132 (2) (2017) 665–712.
- [37] B. Kelly, D. Papanikolaou, A. Seru, M. Taddy, Measuring technological innovation over the long run, *Am. Econ. Rev.: Insights* 3 (3) (2021) 303–320.
- [38] D.J. Teece, Profiting from technological innovation: implications for integration, collaboration, licensing and public policy, *Res. Pol.* 15 (6) (1986) 285–305.
- [39] PayPal, PayPal to acquire curv to accelerate and expand its initiatives to support cryptocurrencies and digital assets. <https://newsroom.paypal-corp.com/2021-03-08-PayPal-to-Acquire-Curv>, 2021.
- [40] C. Rudin, Stop explaining black box machine learning models for high-stakes decisions and use interpretable models instead, *Nat. Mach. Intell.* 1 (5) (2019) 206–215.
- [41] E. Kamateri, M. Salamapasis, E. Perez-Molina, Will AI solve the patent classification problem? *World Pat. Inf.* 78 (2024) 102294.
- [42] L. Jiang, C. Zhang, P.A. Scherz, S. Goetz, Can Large Language Models Generate high-quality Patent Claims?, 2024 *arXiv preprint arXiv:2406.19465*.
- [43] I. Anagnostopoulos, Fintech and regtech: impact on regulators and banks, *J. Econ. Bus.* 100 (2018) 7–25.

Milad Armani Dehghani is a Senior Research Fellow in FinTech at Nottingham Business School, a Visiting Research Professor at Thammasat University, Bangkok and a patent-driven AI and FinTech innovation advisor. He holds a PhD in Technology Management from Sapienza University of Rome and has over a decade of international experience in digital innovation, technology management, and IP strategy across the UK, Ireland, Hong Kong, and Canada. His research focuses on FinTech, digital innovation, and emerging technologies, including Web3 and AI. He is also the inventor of three FinTech-related IPs, founder of LayawayGo, and author of over 30 publications in high-impact journals and conferences.

Mehmet Sahiner is a Lecturer in Financial Economics at the University of Dundee's School of Business. He holds a PhD in Finance from the University of Stirling, an MSc in Finance from Queen Mary University of London, and a BSc in Economics from Ankara University. Before his current position, Mehmet was a Research Associate in FinTech at Nottingham Business School and a Visiting Lecturer at Stirling Management School.

Noptanit Chotisarn is an assistant professor in the Department of Management Information Systems at Thammasat Business School. With an academic foundation spanning both political science and computer science, he holds a Ph.D. in Software Engineering from Zhejiang University, China, where he specializes in cutting-edge software technologies. His interdisciplinary expertise in software engineering and international relations equips her with a unique perspective on management information systems.