BETA
Bureau
d'économie
théorique
et appliquée

Documents de travail

« Exploring Academic Patent–Paper Pairs in Japan: Benchmarking Existing Detection Models »

Auteurs

Van-Thien Nguyen, Rene Carraz

Document de Travail n° 2025 – 27

Juillet 2025

Bureau d'Économie
Théorique et Appliquée
BETA

https://www.beta-economics.fr/

Contact :
jaoulgrammare@beta-cnrs.unistra.fr

Université de Strasbourg · cnrs · UNIVERSITÉ DE LORRAINE · INRAE · AgroParisTech · UNIVERSITÉ HAUTE-ALSACE

Exploring Academic Patent–Paper Pairs in Japan: Benchmarking Existing Detection Models.

Van-Thien Nguyen[1] and Rene Carraz[2,3*]

[1]Graduate School of Frontier Sciences, The University of Tokyo, Japan.
[2]Department Global Innovation Studies, Toyo University, Japan.
[3]BETA, UMR 7522 CNRS-Unistra, France.

*Corresponding author(s). E-mail(s): carraz@toyo.jp;

Abstract

This study expands on the patent-paper pair (PPP) detection model developed by Nguyen and Carraz (2025, Scientometrics) by systematically comparing it with two prominent large-scale approaches: Marx and Scharfmann (2024) and Wang et al. (2025). Although these models all aim to identify instances where the same research result is disclosed through both a patent and a scientific paper, they differ substantially in scope, design, and methodological assumptions. The Nguyen and Carraz model is designed for the Japanese academic context and integrates inventor–author matching, citation overlap, and semantic and lexical similarity within a supervised learning framework. In contrast, Marx and Scharfmann rely on detecting long identical word sequences ("self-plagiarism") via a random forest classifier, and Wang et al. implement an inventor-centric clustering method with logistic regression applied to title and abstract similarity. We directly compare the Nguyen and Carraz dataset with those of Marx and Scharfmann and Wang et al., focusing on PPPs involving Japanese academic assignees. Despite the shared national context, there is minimal overlap: only 168 PPPs overlap with the Marx and Scharfmann model and 425 overlap with the Wang et al. model. When evaluated on a shared validation set, the Nguyen and Carraz model outperforms both alternatives in the Japanese academic context, especially with logistic regression features. Feature extensions such as self-plagiarism and geographic distance offer only modest improvements under non-linear models. These findings highlight the importance of designing context-specific models and exercising caution when applying global PPP datasets to localized settings.

Keywords: Patent Paper Pair; Methodology; Matching algorithm; Academic patent; Japan

JEL Classification: 031 , 034 , 05

1 Introduction

Patent-paper pairs (PPPs), in which the same research result is protected by a patent and published as a scientific paper, have emerged as a valuable analytical tool for studying the interface between scientific research and technological innovation. By capturing instances of dual disclosure, PPPs provide concrete evidence of cases where scientific discovery and inventive application arise from a common research project. The potential of PPPs was first demonstrated in the work of Murray and Stern (2007), and was subsequently expanded by Lissoni et al. (2013) and Magerman et al. (2015).

More recently, the field has seen significant methodological advances, driven by the availability of structured data on non-patent literature (Marx and Fuegi 2020, 2022) and the increasing sophistication of text mining and machine learning techniques. For example, Marx and Scharfmann (2024) introduced a large-scale PPP detection model that exploits the identification of long identical word sequences, a way to mimic "self-plagiarism" between patent and publication abstracts, coupled with a random forest classifier trained on a dataset of over 800 manually reviewed PPPs. In parallel, Wang et al. (2025) developed a two-step approach that first constructs an academic inventor database through name disambiguation and affiliation matching, and then applies domain-specific clustering along with logistic regression to assess document similarity. These improvements significantly increase the scope of PPP matching beyond previous rule-based or manually curated approaches and have enabled the emergence of large open datasets.

As a result, a new wave of large-scale PPP studies has emerged over the past two years, including contributions by Kwon (2024), Lippert and Förstner (2024), Marx and Scharfmann (2024), Raiteri and Buettner (2024), Nguyen and Carraz (2025) and Wang et al. (2025). Collectively, these studies underscore the growing interest in PPPs as an analytical construct for understanding the co-evolution of science and technology across institutional, disciplinary, and national contexts.

The present study contributes to this growing body of work by extending previous research specifically tailored to the academic context of Japan, as developed by Nguyen and Carraz (2025). Although narrower in geographic scope than the models proposed by Marx and Scharfmann (2024) or Wang et al. (2025), the Van Thien and Carraz approach prioritizes a cohesive institutional framework and incorporates extensive validation procedures. emphasizes a cohesive institutional context and extended validation exercises. Using a dataset of 115 universities and 22 public research institutes in Japan, the model employed a scoring system that integrates inventor-author matching, semantic and lexical similarity, and citation overlap. The model was trained on a manually validated dataset supported by over 700 PPPs and achieves high predictive performance. The resulting open dataset consists of 16,899 high-confidence PPPs identified between 2004 and 2018.

Considering the innovative nature and extensive scale of Marx and Scharfmann (2024)'s model, this study conducts a comprehensive comparative analysis to evaluate the validity, specificity, and robustness of the model developed by Nguyen and Carraz (2025). Additionally, the inventor-focused methodology by Wang et al. (2025) provides another critical benchmark, allowing us to clearly articulate the methodological divergences and their implications. The results of this comparison highlight how

distinct design choices, ranging from feature selection to model scope, substantially shape PPP detection outcomes, reinforcing the need for methodological frameworks that are carefully adapted to specific research contexts.

The remainder of the paper is organized as follows. Section 2 outlines the methodological framework developed by Nguyen and Carraz (2025), including dataset construction, the design of the matching algorithm, and the supervised learning model used to detect academic PPPs in the Japanese context. Section 3 offers a comparative evaluation of this approach alongside the models proposed by Marx and Scharfmann (2024) and Wang et al. (2025), with particular attention to methodological design, predictive performance, and contextual applicability. Section 4 concludes by examining the findings' implications and emphasizing the trade-offs and complementarities among PPP detection strategies.

2 Methodology Overview

This study develops and evaluates the Nguyen and Carraz (2025) model for detecting academic patent–paper pairs (PPPs). This section provides a summary of the methodology developed by Van Thien and Carraz; a full description is available in the original study. The corresponding code and dataset are openly accessible at: https://github.com/ReneCarraz/Patent-Paper-Pair.

2.1 Institutional Scope and Dataset Construction

The sample was constructed from 115 universities and 22 National Public Research Institutes (PRIs) in Japan, each of which had at least one patent granted by the United States Patent and Trademark Office (USPTO) between 2004 and 2018. The choice to rely on USPTO data was motivated by the richer availability of non-patent literature (NPL) citations compared to Japanese patent filings. Patent data were obtained from the *PatentsView* platform, while bibliometric records were obtained from *OpenAlex*. After applying filters based on institutional affiliation and authorship, the final dataset included 10,896 patents and 652,610 publications.

2.2 Identification of Candidate Matches

Potential PPPs were initially identified by matching inventors listed in patent applications to the first or last authors of publications affiliated with the same institution. A publication was considered a candidate if its date fell within a three-year window from the priority date of the patent (-1 to +2 years). To refine the dataset, we excluded review articles and applied institutional filters to ensure consistency. This process resulted in 467,669 potential matches.

2.3 Scoring System

To assess the similarity between candidate PPPs, Nguyen and Carraz (2025) develop a four-dimensional scoring system that integrates both lexical and contextual indicators:

- ***Inventor Score:*** Measures the proportion of overlapping inventors and authors, requiring that either the first or last author be an inventor.

- ***Semantic Similarity Score:*** Computed via S-BERT embeddings using cosine similarity between the title and abstract of patents and papers, capturing context-aware semantic proximity.
- ***Word Overlap Score:*** Calculates the lemmatized word overlap between patent and paper abstracts and titles, normalized by the total word count.
- ***Citation Overlap Score:*** Measures bibliographic citation overlap using NPL citations from the Marx and Fuegi datasets (Marx and Fuegi 2020, 2022).

Each indicator was normalized to a range of 0 to 1, enabling a multi-dimensional assessment of relatedness between documents.

2.4 Labeled Dataset and Model Development

Nguyen and Carraz (2025) develop a supervised machine learning model to classify true PPPs using the four similarity scores described above. To generate training data, they validate 722 pairs using two methods: (1) direct author feedback from a sample of 90 researchers (143 confirmed matches, 247 false matches), and (2) manual visual validation of 600 randomly selected pairs based on shared thematic content, author affiliations, and graphical elements. A balanced training dataset of 361 positive and 361 negative matches is constructed, the latter including a control group of semantically similar but unrelated works identified via OpenAlex's `related_works` function.

Then a logistic regression model is trained using the four scores. The model demonstrates strong performance under stratified fivefold cross-validation, achieving an F1 score of 0.80. Comparative benchmarking against prior models (e.g., Lissoni et al. 2013; Magerman et al. 2015) confirms the enhanced precision and recall of this approach.

2.5 Results of Matching Process

Applying the trained model to the full dataset of 467,669 potential matches results in the identification of 16,899 high-confidence academic PPPs. Of these, 1,280 are one-to-one matches, while 2,726 patents are associated with multiple publications. Descriptive statistics show that most PPPs are concentrated in materials science, chemistry, and physics. The most frequently represented institutions include the University of Tokyo, Kyoto University, Tohoku University, and the National Institute of Advanced Industrial Science and Technology (AIST).

An analysis of feature importance indicated that citation overlap score and word overlap score were the most predictive indicators in the model. Specifically, the logistic regression coefficients showed strong statistical significance for citation overlap ($\beta = 1.24$, $p < 0.001$) and word overlap ($\beta = 0.85$, $p < 0.001$), while the inventor score also played a significant but comparatively smaller role ($\beta = 0.45$, $p < 0.001$). The semantic similarity score, although not statistically significant on its own, contributes to the overall robustness and interpretability of the model.

3 Comparison

To further assess the validity and distinctiveness of Nguyen and Carraz (2025) model, we conducted a structured comparison with two prominent large-scale PPP detection approaches: Marx and Scharfmann (2024) and Wang et al. (2025). This section is divided into four parts. Section 3.1 outlines the main methodological differences between the three models, focusing on dataset coverage, feature design, and model architecture. Section 3.2 compares overlapping PPPs identified in the Japanese context, using four core metrics: Inventor Score, Citation Overlap, Word Overlap, and Semantic Similarity. Robust statistical tests are applied to highlight where the models diverge in what they detect. Section 3.3 uses a validated dataset of 722 PPPs to benchmark model performance. Each model's feature set is applied under identical conditions, and the models are evaluated using both logistic regression and random forest classifiers. This enables a direct performance comparison applied to a shared evaluation framework. Section 3.4 expands on this by examining whether integrating additional features inspired by Marx and Scharfmann, such as self-plagiarism indicators and inventor–assignee distance, into the Van Thien and Carraz model adds predictive value.

3.1 Descriptive Statistics

Table 1 summarizes the key methodological differences among three PPP detection models: the model of Nguyen and Carraz (2025), Marx and Scharfmann (2024) and Wang et al. (2025). While all models rely on open-access patent and publication metadata, they differ substantially in scope, feature design, and methodological approach.

The model proposed by Nguyen and Carraz (2025), evaluated in the present study, targets Japan's academic sector and combines inventor–author matching, semantic and lexical text similarity, and citation overlap within a supervised learning framework. By contrast, Marx and Scharfmann (2024) adopt a global approach centered on detecting high-confidence "self-plagiarism" via long identical sequences, author name overlap, and citation-based features, using a random forest classifier. Wang et al. (2025) implement an inventor-centric model, clustering disambiguated inventor profiles and applying logistic regression on title and abstract similarity.

Each approach emphasizes distinct dimensions of PPP detection. Nguyen and Carraz (2025) model focuses on identifying academic PPPs within a national research context. Marx and Scharfmann (2024) prioritize textual reuse, emphasizing direct phrase overlap between documents in a global setting. Wang et al. (2025) , by contrast, capture broader inventor-centric connections through clustering techniques and abstract similarity evaluation. These differences influence both the composition of detected PPPs and the types of analytical conclusions each model supports.

Table 1: Comparison of the Models Methodology

Criteria	**Nguyen & Carraz (2025)**	**Marx & Scharfmann (2024)**	**Wang et al. (2024)**
Dataset Source	Open-access: Yes; 16,899 PPPs; Patents USPTO + OpenAlex. Japan-focused, academic inventors	Open-access: Yes; ˜500,000 PPPs; USPTO + OpenAlex, self-plagiarism detection, author-inventor matching	Open-access: Yes; 14,137,072 PPPs; PATSTAT + Scopus; inventor-based clustering
Sample Selection Criteria	Manual validation + supervised model using inventor-author match, semantic and lexical similarity, citation overlap	Self-plagiarism detection: long exact sequences (> 10 words) between patent and paper abstracts; author-inventor name overlap; citation; combined in a random forest classifier	Candidates generated via author-inventor disambiguation and domain clustering; logistic regression on abstract/title similarity applied in classification stage
Benchmark Pairing (Non-PPPs)	Semantically similar but manually verified false pairs; unrelated but similar documents using OpenAlex related_works	Pairs published > 5 years apart; random combinations of publications with overlapping names	Not explicitly detailed; inferred via clustering and logistic modeling
"Golden Goose" - Training Set	Validated 722 pairs (researcher survey + manual inspection)	800+ hand-coded PPPs for validation	Used disambiguated author-inventor data as training proxy
Geographical Restriction	Academic Japan-based institutions	Global scope, no specific geographical restriction	Global scope with focus on inventor-level profiles
Feature Types Used	Inventor score, S-BERT-based semantic similarity, word overlap, citation overlap, publication time gap (3-year window, -1 to +2)	Textual overlap (self-plagiarism), author name match, citation presence (patent cites the paper), institution–assignee distance, publication time gap (4-year window, -2 to +2)	Title/abstract similarity, author-inventor affiliation, cluster proximity, publication time gap (5-year window, -3 to +2)
Feature Importance	Most important: citation overlap and word overlap	Most important: Textual overlap (self-plagiarism) and author name match	Most important: author-inventor link and abstract similarity

Criteria	Nguyen & Carraz (2025)	Marx & Scharfmann (2024)	Wang et al. (2024)
Other Notable Features	Focus on institutional-level validation; contextual model tuned to Japanese academic system	Emphasis on high-confidence self-citations and large-scale global matching	Inventor-centric view; large-scale global matching

3.2 Feature-Based Comparison of Overlapping PPPs

To enable a direct comparison with the model developed by Nguyen and Carraz (2025), the analysis focused on the subsets of PPPs from Marx and Scharfmann (2024) and Wang et al. (2025) that intersect with Japanese assignees, given the national specificity of the Van Thien and Carraz dataset. We obtained the complete lists of PPPs from Marx and Scharfmann and Wang et al., filtering each to include only cases where the patent's assignee is in Japan. Specifically, first, we obtained the PPP data published by Marx and Scharfmann from the Reliance on Science platform (https://relianceonscience.org/patent-paper-pairs). This dataset provides patent IDs, and we used USPTO data to extract the country of each assignee. We then filtered out pairs where the assignee country did not include Japan. Similarly, we downloaded the PPP data published by Wang et al. from Zenodo (https://zenodo.org/record/15478277). This dataset contains application IDs, and we used USPTO data to map application ID to patent ID. We then obtained the assignee country for each patent. Finally, we filtered out any pairs with an `assignee_country` other than Japan.

This yielded 4,387 PPPs from the Marx and Scharfmann (2024) dataset (out of 107,820, or 4.1%) and 99,675 PPPs from the Wang et al. (2025) dataset (out of over 14 million, or 0.7%). Among these, only a small fraction overlapped with our 16,899 PPPs. Specifically, 168 pairs were common between Nguyen and Carraz (2025) and Marx and Scharfmann's dataset, and 425 pairs were common with Wang et al. dataset (see Table 2). The remaining Japanese-context pairs – 4,219 were unique to Marx and Scharfmann, and 99,250 were unique to Wang et al. – did not appear in Nguyen and Carraz. This limited overlap highlights that each model, with its unique criteria, identifies a distinct set of PPP connections. Even within the same country and time span, many pairs identified by the global models were absent from our list, and vice versa.

Table 2: Overlap of PPP datasets for Japan (2004–2018)

Dataset	Total Entries	Entries in Common with Nguyen & Carraz (2025)	Unique Entries
Nguyen & Carraz (2025)	16 899	—	16 899
Marx & Scharfmann (2024)	4 387	168	4 219
Wang et al. (2024)	99 675	425	99 250

The dataset constructed by Nguyen and Carraz (2025) is designed to include only patents and publications affiliated with Japanese academic institutions. After rigorous filtering based on institutional affiliation and authorship, the dataset contained 10,896 patents and 652,610 publications. These were used to generate a pool of candidate matches. After applying a

supervised learning model based on four key similarity indicators, 16,899 high-confidence PPPs were identified. Consequently, each PPP in this dataset is explicitly linked to a Japanese academic institution. In contrast, the global models developed by Marx and Scharfmann (2024) and Wang et al. (2025) cover a much broader population of PPPs, only a small fraction of which involve Japanese academic assignees. Consequently, their capacity to detect Japan-specific academic PPPs seems to be limited in scale and coverage.

To evaluate the differences between the PPPs identified by Marx and Scharfmann (2024), Wang et al. (2025) , and those in the Nguyen and Carraz (2025) dataset, we examined the distributions of four key matching features used in the Nguyen and Carraz model — *inventor score, citation overlap, word overlap*, and *semantic similarity* —across three groups: (1) the complete Nguyen and Carraz dataset, (2) the subset overlapping with Marx and Sharfmann's Japan-related PPPs, and (3) the subset overlapping with Wang et al.'s (See Table 2). Statistical tests revealed that the distributions of these features differ significantly between the groups.

To assess distributional assumptions, we applied the Shapiro–Wilk test for normality and Levene's test for homogeneity of variance. Both tests rejected parametric assumptions. As a result, we used the Mann–Whitney U test (Nachar 2008) to compare feature distributions between the Nguyen and Carraz (2025) dataset (N = 16,899) and the subsets from Marx and Scharfmann (2024) (N = 168) and Wang et al. (2025) (N = 425). The results, summarized in Table 3, show statistically significant differences in several features.

Table 3: Mann–Whitney U Test Results

Feature	**Comparison**	**U statistic**	**p-value**	**Significant** ($\alpha = 0.05$)
Inventor Score	NC vs. Marx	1,023,852	$< .001$	Yes
	NC vs. Wang	3,616,683	0.738	No
Citation Overlap Score	NC vs. Marx	1,463,022	0.984	No
	NC vs. Wang	4,585,140	$< .001$	Yes
Word Overlap Score	NC vs. Marx	674,783	$< .001$	Yes
	NC vs. Wang	2,696,451	$< .001$	Yes
Semantic Similarity Score	NC vs. Marx	1,092,886	$< .001$	Yes
	NC vs. Wang	2,914,453	$< .001$	Yes

Note: For brevity, "Marx" refers to the intersection with Marx & Scharfmann (2024), "Wang" refers to the intersection with Wang et al. (2024), and "NC" to the full dataset of Japanese academic PPPs identified by Nguyen & Carraz (2025).

Feature-level comparisons confirm that design choices significantly influence how PPPs are identified. For example, Marx and Scharfmann (2024)'s reliance on strict text reuse signals may have skewed their dataset toward high-confidence inventor-linked PPPs, while Wang et al. (2025)'s broader clustering approach produces overlaps in semantic and lexical similarity but diverges in citation linkage with Nguyen and Carraz (2025) 's model.

These contrasting results underscore that PPP datasets are shaped by the priorities embedded in each model's design. The low overlap and distinct feature profiles suggest that no single methodology offers a complete and unified picture. This has important implications for studies relying on PPPs to trace knowledge transfer, commercialization, or dual disclosure. In the next section, we compare the predictive performance of the three models within a shared validation framework.

3.3 Comparison of Model Performances

To benchmark the Nguyen and Carraz (2025) model against those developed by Marx and Scharfmann (2024) and Wang et al. (2025), we replicated each feature set based on their published methodologies and publicly available code. Rather than comparing models on their original datasets, we applied each feature set to Nguyen and Carraz's labeled dataset of 722 PPPs, evenly balanced between 361 positive and 361 negative examples while covering 444 unique patents and 575 unique papers. This approach ensured that all models were evaluated under the same data conditions, allowing us to directly assess the contribution of feature design to classification performance. Nguyen and Carraz feature set includes four components: inventor score, semantic similarity score, word overlap score, and citation overlap score, as described in Section 2.3. Marx and Scharfmann's feature set consists of title similarity, abstract similarity, the number of shared author names between paper and patent, the proportion of inventors who are also authors, and the geographical distance between the patent assignee and the author's affiliated institution. Meanwhile, the feature set used by Wang et al. includes word overlap and semantic similarity measures for both titles and abstracts (see Table 1 for a concise description of all models).

For each feature set, we trained both a logistic regression model and a random forest classifier to evaluate predictive performance. While Nguyen and Carraz (2025) model employed logistic regression, we adopted Marx and Scharfmann (2024)'s use of random forests to examine whether non-linear classifiers offer improved accuracy. The models were trained using stratified fivefold cross-validation, and the results are summarized in Table 4.

Table 4: Model Performance across Different Feature Sets (5-fold cross-validation)

Type	Feature Set	Accuracy	Precision	Recall	F1
Logistic Regression	NC	0.8144	0.8259	0.7915	0.8080
	Marx	0.7659	0.7551	0.7813	0.7675
	Wang	0.7714	0.7875	0.7393	0.7594
Random Forest	NC	0.8352	0.8245	0.8492	0.8365
	Marx	0.7908	0.7803	0.8044	0.7919
	Wang	0.7424	0.7355	0.7594	0.7434

Note: For brevity, "Marx" refers to Marx & Scharfmann (2024), "Wang" refers to Wang et al. (2024), and "NC" refers to Nguyen & Carraz (2025).

Random forest classifiers show a slight performance advantage across all feature sets. When using Nguyen and Carraz (2025) features, the random forest achieved an F1 score of 0.8365 and demonstrated consistently high accuracy, precision, and recall. This suggests that non-linear models can more effectively capture complex feature interactions, especially when combining heterogeneous inputs such as semantic similarity, co-authorship, and bibliographic overlap.

Nevertheless, the feature set developed by Nguyen and Carraz (2025) outperformed those of Marx and Scharfmann (2024) and Wang et al. (2025) across all classifier types. This highlights the importance of incorporating various dimensions of similarity, such as lexical, semantic, citation-based, and institutional, into a single framework. Furthermore, the stable results across linear and nonlinear classifiers support the generalizability of their approach.

These findings highlight the methodological soundness and adaptability of the proposed scoring system for broader PPP research.

3.4 Improvements

Building on the comparative evaluation in the previous section, we aimed to extend the work of Nguyen and Carraz (2025) by exploring whether the integration of additional predictive features could further enhance model performance. In particular, we drew on elements introduced by Marx and Scharfmann (2024), notably their use of "self-plagiarism" detection and the inclusion of geographic distance between inventors and assignees—features that potentially capture latent dimensions of authorial overlap and institutional proximity. While these variables were not part of the original Nguyen and Carraz model, we hypothesized that, when combined with the existing indicators, they could provide complementary predictive value in identifying high-confidence PPPs.

To extend the original feature set developed by Nguyen and Carraz (2025), we constructed an augmented version that incorporates several additional variables. These include semantic similarity between titles and abstracts (using S-BERT embeddings), lexical word overlap for both titles and abstracts, inventor–author overlap, and citation overlap based on shared DOI references—features already present in the baseline model. To this, we added two elements inspired by Marx and Scharfmann (2024): geographical proximity (measured via Haversine distance between author and assignee affiliations) and the total length of the three longest shared textual sequences between the patent and publication abstracts, serving as a proxy for self-plagiarism.

When trained with logistic regression, Table 5 shows that the extended model achieved an F1 score of $0.8131(\pm 0.0168)$, with precision at 0.8270 and recall at 0.7918. These results are comparable to those of the original Nguyen and Carraz (2025) specification, suggesting that the inclusion of additional features does not substantially improve performance under linear assumptions. By contrast, the random forest model performed slightly better, reaching an F1 score of 0.8597, indicating that non-linear classifiers are more effective in capturing complex interactions among the expanded set of features.

Table 5: Model Performance with Extended Feature Set

Classifier	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Logistic Regression	0.8131	0.8270	0.7918	0.8087
Random Forest	0.8587	0.8545	0.8653	0.8597

Despite the comparable performance, feature importance varied notably between the two classifiers. In the logistic regression model, the most influential feature was citation overlap ($\beta = 1.182$, $p < 0.001$), followed by word overlap ($\beta = 0.734$, $p < 0.001$) and semantic similarity ($\beta = 0.466$, $p = 0.016$). Inventor–author overlap also showed a significant contribution ($\beta = 0.385$, $p < 0.005$), while the additional features, self-plagiarism ($\beta = 0.10$, $p = 0.441$) and geographic distance ($\beta = -0.08$, $p = 0.532$), had coefficients close to zero and were not statistically significant. These results suggest that the extensions add limited value in a linear setting. Table 6 reports the full set of coefficient estimates and significance levels.

In the random forest model, we assess predictor relevance with permutation feature importance, which quantifies the drop in predictive accuracy when a variable's values are randomly

Table 6: Feature Weights (Logistic Regression, with Extended Feature Set)

Feature	**Coefficient**
Citation Overlap	1.182***
Word Overlap	0.734***
Semantic Similarity	0.466
Inventor–Author Overlap	0.385**
Geographic Distance	−0.079
Self-Plagiarism	0.103

Significance levels: $^{*}p < 0.01$, $^{**}p < 0.005$, $^{***}p < 0.001$.

permuted (Breiman 2001). This metric substitutes for p-values, which are unavailable because random forests lack parametric coefficients with tractable sampling distributions, making classical hypothesis tests inapplicable (Altmann et al. 2010). In the extended version of the Nguyen and Carraz (2025) model, citation overlap remained the most influential variable (importance = 0.325), followed by semantic similarity (importance = 0.179) and word overlap (importance = 0.210). In contrast, the additional features, self-plagiarism (importance = 0.036) and geographic distance (importance = 0.101), played a comparatively minor role, indicating limited added value in the non-linear setting.

Table 7: Feature Importances (Random Forest, with Extended Feature Set)

Feature	**Importance**
Citation Overlap	**0.325**
Semantic Similarity	0.179
Word Overlap	0.210
Inventor–Author Overlap	0.150
Geographic Distance	0.101
Self-Plagiarism	0.036

Because self-plagiarism and geographic distance were relatively unimportant in both models, we conducted a feature ablation analysis to see if removing these variables would improve or at least preserve overall performance. In the logistic regression model, excluding geographic distance slightly increased the F1 score to 0.8122, while removing self-plagiarism marginally decreased it to 0.8055. Both outcomes remained close to the full model's performance. In the random forest classifier, removing self-plagiarism and geographic distance produced modest reductions in the F1 score, to 0.8504 and 0.8301, respectively. This indicates that, while these features contribute some value in nonlinear settings, their overall impact is limited.

Overall, these results demonstrate that the original feature set developed by Nguyen and Carraz (2025) —which includes semantic similarity, word overlap, inventor–author linkage, and citation overlap—is highly effective and robust across linear and nonlinear classification models. However, the addition of self-plagiarism and geographic proximity did not yield notable improvements in the logistic regression setting and may introduce marginal noise rather than signal. In the random forest model, however, these features contributed more meaningfully, though modestly, suggesting that they may be useful when optimizing for maximum predictive accuracy. These findings support a pragmatic hybrid strategy of employing

Table 8: F1 Scores After Dropping Features

Classifier	Dropped Feature	F1 Score
Logistic Regression	None (Full Model)	0.8087
	Self-Plagiarism	0.8055
	Geographic Distance	0.8122
Random Forest	None (Full Model)	0.8597
	Self-Plagiarism	0.8504
	Geographic Distance	0.8301

logistic regression with the core feature set for greater interpretability and leveraging random forest models with selective feature expansion when enhanced predictive performance is a priority.

4 Conclusion

This comparative analysis underscores that the three PPP detection models examined — Nguyen and Carraz (2025), Marx and Scharfmann (2024), and Wang et al. (2025) — yield markedly divergent results, as clearly demonstrated in Table 2, which highlights substantial differences in both the quantity and the specific PPPs identified by each model, emphasizing the distinct selection mechanisms and criteria employed. Notably, the Nguyen and Carraz approach identified substantially more PPPs explicitly linked to Japanese academic patents, likely due to its stringent selection criteria, which include exclusively academic institutions and their associated academic publications, significantly limiting noise and enhancing specificity. These findings underscore the importance of accounting for local context, including the relative homogeneity of academic scientists' behavior in the Nguyen and Carraz study, as well as the institutional norms and national research practices that influence patenting and publishing processes.

However, the degree of divergence between the Nguyen and Carraz (2025) dataset and those of Marx and Scharfmann (2024), and Wang et al. (2025) was surprising, highlighting a critical caution for researchers: global large-scale models can significantly underrepresent PPP phenomena when applied to localized or institutional contexts such as the academic landscape of Japan. This divergence prompted further methodological testing within this study, in which we extended the original Nguyen and Carraz model by incorporating additional predictive features inspired by Marx and Scharfmann, specifically, the random forest classifier method, self-plagiarism indicators, and geographic distance measures.

The results of this comparative analysis confirm the robustness and adaptability of the original Nguyen and Carraz (2025) feature set, demonstrating consistent and strong performance under both logistic regression and random forest classifiers. The stability of these results across different modeling approaches indicates that the core feature set effectively captures the key dimensions necessary for reliable PPP detection, confirming its suitability and generalizability across diverse analytical contexts. While the extended features offered modest incremental predictive power within random forest models, the original core features (semantic similarity, lexical overlap, citation overlap, and inventor-author linkage) remained central and effective. These findings reinforce the practical value of employing the Nguyen and Carraz framework as a foundational approach, adaptable to additional features when necessary. The results also suggest that large-scale models may produce a large number of

false negatives. This warrants caution when interpreting their outputs in ongoing analyses, particularly when the breadth of the approach comes at the expense of local specificity.

All resulting datasets —including the full set of Japanese academic PPPs identified through the extended Nguyen and Carraz (2025) approach— are publicly available on *GitHub*, supporting further validation and reuse by the broader research community. Future methodological work could explore the application of this approach to other national contexts, as well as the distinct characteristics of corporate versus academic PPPs. Additionally, the current training dataset used by Nguyen and Carraz is limited to 721 validated pairs. Further expansion of this dataset could strengthen the evaluation of model robustness and enhance its generalizability across settings.

Declarations

- Funding: This research was supported by the JSPS grant 24K05092.
- Conflict of interest: The authors have no relevant financial or non-financial interests to disclose.
- Data and code availability : The datasets and code used and/or analyzed during the current study are available here: https://github.com/ReneCarraz/Patent-Paper-Pair

References

Altmann, A., L. Toloşi, O. Sander, and T. Lengauer. 2010. Permutation importance: a corrected feature importance measure. *Bioinformatics 26*(10): 1340–1347. https://doi.org/10.1093/bioinformatics/btq134 .

Breiman, L. 2001. Random Forests. *Machine Learning 45*(1): 5–32. https://doi.org/10.1023/A:1010933404324 .

Kwon, S. 2024. Underappreciated government research support in patents. *Science 385*(6712): 936–938. https://doi.org/science.org/doi/10.1126/science.ado1078 .

Lippert, K. and K.U. Förstner. 2024. Patent-publication pairs for the detection of knowledge transfer from research to industry: reducing ambiguities with word embeddings and references. *arXiv preprint arXiv:2412.00978* .

Lissoni, F., F. Montobbio, and L. Zirulia. 2013. Inventorship and authorship as attribution rights: An enquiry into the economics of scientific credit. *Journal of Economic Behavior & Organization* 95: 49–69. https://doi.org/10.1016/j.jebo.2013.08.016 .

Magerman, T., B.V. Looy, and K. Debackere. 2015. Does involvement in patenting jeopardize one's academic footprint? An analysis of patent-paper pairs in biotechnology. *The New Data Frontier 44*(9): 1702–1713. https://doi.org/10.1016/j.respol.2015.06.005 .

Marx, M. and A. Fuegi. 2020. Reliance on science: Worldwide front-page patent citations to scientific articles. *Strategic Management Journal 41*(9): 1572–1594. https://doi.org/10.1002/smj.3145 .

Marx, M. and A. Fuegi. 2022. Reliance on science by inventors: Hybrid extraction of in-text patent-to-article citations. *Journal of Economics & Management Strategy 31*(2): 369–392. https://doi.org/10.1111/jems.12455 .

Marx, M. and E. Scharfmann 2024. Does Patenting Promote the Progress of Science? Technical report, Working Paper.

Murray, F. and S. Stern. 2007. Do formal intellectual property rights hinder the free flow of scientific knowledge?: An empirical test of the anti-commons hypothesis. *Academic Science and Entrepreneurship: Dual engines of growth 63*(4): 648–687. https://doi.org/10.1016/j.jebo.2006.05.017 .

Nachar, N. 2008. The Mann-Whitney U: A test for assessing whether two independent samples come from the same distribution. *Tutorials in quantitative Methods for Psychology 4*(1): 13–20 .

Nguyen, V.T. and R. Carraz. 2025. "Exploring academic patent-paper pairs: a new methodology for analyzing Japan's research landscape". *Scientometrics 130*(3): 1329–1356. https://doi.org/10.1007/s11192-025-05275-5 .

Raiteri, E. and B. Buettner 2024. Unveiling Hidden Connections Between Science and Innovation A Novel Approach to Patent-Paper Pairs. In *EPIP Conference*, EPIP Conference, Pisa.

Wang, Y., L. Pei, J. Sun, and L. Kang. 2025. Trace on both sides: a two-step text mining method to identify academic inventors' patent–paper pairs. *Scientometrics 130*(2): 833–860. https://doi.org/10.1007/s11192-024-05207-9 .