LUISS

Degree Program in Data Science and Management

Course of Big Data and Smart Data Analytics

# Gender Analysis and Prediction in Computer Science Publications


Prof. Irene Finocchi

SUPERVISOR

Prof. Blerina Sinaimeri

CO-SUPERVISOR

Omar Temirgali 780481

CANDIDATE


Academic Year 2024/2025

# Abstract

This thesis measures gender representation in computer science publishing using the dblp bibliography as the sole data source. I parse the dblp XML into per-authorship records, clean and normalize the metadata, and build analysis tables that cover 7,993,261 publications. Expanding to authorships yields 27,189,298 name instances, which I reduce to 3,982,244 unique authors for inference. Research areas are assigned from titles with sentence-transformer embeddings; I evaluate three encoders and select all-MiniLM-L6-v2 for its balanced quality and runtime, then refine assignments with a venue-level mapping for high-volume and single-scope outlets. Author gender is inferred with GenderAPI using confidence scores; I combine a fast bulk pass with a targeted second pass that invokes the provider's AI setting for unresolved names, and I carry confidence through to all summaries.

Results show rapid growth of computer-science output since the 2000s, led by journal articles and conference papers. The female share rises steadily over time, from roughly one in ten authorships in the 1980s to about one in five in recent years, while the male share declines by the same amount and remains the majority. The study contributes a reproducible pipeline for dblp, a practical method for topical assignment from short titles, and an uncertainty-aware analysis of gender composition across time, venues, and research areas.

Keywords: dblp; author gender; GenderAPI; sentence transformers; publication trends; research categories;

# Acknowledgment

I am grateful to **Professor Irene Finocchi** for helping me choose this topic and for her steady support throughout the thesis. I am also grateful to the GenderAPI team for their generous support of this research. In particular, I thank **Onur Ozturk** for his timely guidance on using GenderAPI efficiently (e.g., practical code suggestions, batching strategies, and robust handling of null/unknown responses) and for providing additional API credits that made the large-scale experiments in this master's thesis feasible. Their practical advice and responsiveness materially improved the quality and scope of my analysis. Any remaining errors are my own.

# Contents

# List of Tables

# List of Figures

# Acronyms

**CPU** Central Processing Unit. 13

**CSV** Comma-separated values. 14, 15, 21

**dblp** Digital Bibliography & Library Project. 5, 8, 9, 11, 13, 14, 15, 16, 21, 22, 24, 25, 26, 27, 28, 31, 33, 34, 38, 39, 40

**DOI** Digital Object Identifier. 14

**ETL** Extract, Transform, Load. 11

**GDP** Gross Domestic Product. 13

**GPU** Graphics Processing Unit. 13

**ISBN** International Standard Book Number. 14

**LSTM** Long Short-Term Memory. 13

**ORCID** Open Researcher and Contributor ID. 12, 40

**VIAF** Virtual International Authority File. 12

**XML** Extensible Markup Language. 8, 9, 11, 14, 39

# Chapter 1

# Introduction

Gender representation in computer science research has been examined for many years, yet findings often remain fragmented across venues, subfields, and time periods. This thesis offers a comprehensive view built on the Digital Bibliography & Library Project (dblp) computer science bibliography as the sole data source. I construct an end-to-end pipeline that parses the dblp Extensible Markup Language (XML) into per-authorship records, standardizes and cleans the metadata, assigns topical categories to publications using sentence-transformer embeddings refined with venue-level curation, and infers author gender using probability scores from gender prediction tool. The analysis spans publication growth, venue and category composition, and country patterns that are interpreted as indicators of name origin. Throughout the study I use probability-aware summaries and targeted sensitivity checks so that conclusions remain stable under reasonable decision rules. The overall aim is to provide a transparent and reproducible account of how participation is distributed over time, across venues, and across research areas in computer science publishing.

This work contributes on three fronts. First, it delivers a scalable and documented data engineering workflow that turns a large bibliographic XML into analysis ready tables without sacrificing provenance or clarity. Second, it combines modern text embeddings with a practical layer of venue curation, which improves topical precision on short titles while remaining reproducible. Third, it applies probabilistic gender inference at scale and reports results in ways that respect uncertainty, which strengthens the interpretation of field and venue differences. Ethical considerations guide the design and the reporting. Gender is treated as an inferred attribute from names that is suitable for aggregate analysis only, and country labels supplied by the service are descriptively used as a signal of name origin.

## 1.1 Research Objectives

The main objectives are to predict gender from author names, classify publication titles into research categories, and analyze gender patterns throughout the corpus.

1. **Predict gender from author names.**
   Use tools on the unique author list, normalize confidence scores, and integrate results into the per authorship data.
2. **Classify titles into research categories.** Assign categories using sentence transformer embeddings refined with venue-level curation for high volume or single scope venues.
3. **Analyze gender patterns.**
   Measure composition over time and compare between venues, research categories, and country of name labels; relate venue results to their dominant categories.
4. **Quantify publication dynamics.**
   Describe long run trends by year and type of document to provide context for gender analysis.
5. **Validate modeling choices.**
   Examine the speed and quality trade-off between selected embedding models and check robustness using probability weighted and thresholded summaries.
6. **Ensure reproducibility.**
   Document parsing, cleaning, inference, categorization, and aggregation so that results can be replicated and extended.

## 1.2 Structure of the thesis

The thesis opens with Chapter 1, which introduces the problem, motivates the use of dblp as a comprehensive source of computer science publications, and states the research objectives and contributions. Chapter 2 reviews the relevant literature on computer science publishing, previous evidence on gender representation, methods for name-based gender inference, approaches to topic assignment from short texts, and ethical guidance for aggregate reporting. Chapter 3 details the data and methods. It explains how dblp XML is parsed and cleaned, how authorship tables are constructed, how categories are assigned using sentence similarity in combination with venue curation, and how gender is inferred with associated confidence. It also records the uncertainty handling and the reproducibility choices that support the analysis.

Chapter 4 reports on the empirical results. It begins with publication trends by year and by document type, then presents gender composition in general and by year, venue, and research category, and finally summarizes country patterns using the service's country labels as indicators of origin of the name. Where informative, results are related across these views, for example, by linking venue patterns to their dominant categories. Chapter 5 discusses the conclusion of the work.

# Chapter 2

# Literature Review

Recent infrastructure work shows that converting the dblp XML into a relational database greatly improves the feasibility of large scale scientometric analysis. Berenguer (2024) describes an Extract, Transform, Load (ETL) pipeline that preprocesses the raw XML, ingests it into PostgreSQL with XPath, and packages the workflow in Docker for reproducibility and monthly updates. The resulting schema separates publications, publication groups, researchers, and authorships, which enables richer queries than the flat XML and reveals practical issues that matter for this thesis: inconsistent page formats, distinctions among books, proceedings volumes, and individual papers, encoding problems, and performance trade offs at full dblp scale. I follow the same ethos of reproducible data engineering and extend the analysis layer to include uncertainty aware gender inference and topic assignment from titles. In this way the prior work serves as both a methodological reference for robust data preparation and evidence that careful normalization is essential for credible downstream analytics.

Cobo-Serrano et al. (2024) examine gendered authorship patterns in the Journal of Information Science for 2015–2020 and provide a focused example of journal-level bibliometrics with explicit gender identification. The study retrieves full texts from the publisher site, standardizes author name variants, determines gender through web verification and direct email when needed, and stores records in an Access database with fields for authorship type, affiliation, and geography. The corpus comprises 326 articles and 697 authors; overall, 69.58% of identified authors are male and 30.42% are female, with multiple authorship far more common than single authorship. Collaboration tends to involve two or three authors, and most contributors are affiliated with universities. The authors also report 152 signer cases that remained unidentified in multi-authored papers, which they acknowledge as a limitation. As a journal-specific case study, the work reinforces two themes relevant to this thesis: persistent gender imbalance and

the importance of careful identity resolution before analysis. It complements large-scale datasets by showing what can be learned when manual validation is applied in depth to a single venue.

Boté-Vericad et al. (2025) propose a systematic method to identify author gender in bibliometric datasets from Web of Science and Scopus, combining manual verification with scripted queries to Wikidata, Virtual International Authority File (VIAF), Open Researcher and Contributor ID (ORCID), Scopus, and GenderAPI. Using a sample of 187 authors, they report strong agreement for manual coding (Krippendorff's alpha 0.848) and show that manual checks identify 50.8% of authors in Wikidata and 37.43% in VIAF, while automated scripts identify 42.24% and 31.01% respectively. The study documents strengths and limits of each source, notes that automated services tend to miss culturally ambiguous names and non-binary identities, and argues for linked open data, reproducible scripts, and clear uncertainty handling in gender studies of authorship. This evidence backs up my thesis's mixed approach, which combines automated inference with focused follow-up for cases that remain unresolved and reports results with a clear focus on data provenance and confidence.

VanHelene et al. (2024) compare three name-to-gender tools on a gold-standard set of 32,968 clinical trial authors and report high overall accuracy for the commercial services, with 96.6% for Genderize and 96.1% for Gender API when no country is supplied. Accuracy varies by region, with markedly lower performance on several East and Southeast Asian name sets. The study shows that adding country information changes the error profile, slightly improving Gender API and slightly reducing Genderize, while increasing the share of no-prediction cases for some settings. Formatting choices matter. Genderize performs best when two-part given names are concatenated, and it fails when they are separated by spaces, whereas Gender API performs best with a space delimiter and handles diacritics more reliably. Reported confidence aligns with empirical accuracy, with strong correlations and low Brier scores, and the authors note cost differences that favor Genderize at the time of study. The paper recommends careful preprocessing, explicit handling of uncertainty, and periodic recalibration as datasets and naming practices evolve. This evidence supports the use of probability-aware workflows and clear decision rules when applying automated gender inference at scale.

In the study, a large scale evaluation of name to gender inference compares six existing tools and several simple and hybrid baselines across multiple labeled datasets of names, including ACL, CMU, DIME, Facebook, Florida voter files, and SSA (Krstovski et al., 2023). The study reports accuracy, precision, recall, and F1 by dataset and in weighted aggregate, and also quantifies algorithmic bias with a gender bias error metric. Three services consistently perform

well in the aggregate (a character Long Short-Term Memory (LSTM) "chicksexer," NamSor, and Gender API), while a hybrid that first applies a maximum likelihood model trained on combined data and then uses the best external model for uncertain cases attains the highest overall accuracy and F1. Error analysis shows that mistakes are more common for longer names, for names with non-English characters, and for character sequences that the models rarely observe. The results argue for careful preprocessing, explicit probability use, and simple ensemble strategies when scaling gender inference to large bibliographic corpora.

A comparative study of sentence transformer models shows how model choice affects semantic search over scientific corpora (Galli et al., 2024). Using a dataset of 6,110 abstracts from a published systematic review and 24 manually verified target papers, four pretrained encoders—all-MiniLM-L6-v2, all-MiniLM-L12-v2, all-mpnet-base-v2, and all-distilroberta-v1 were used to embed texts and rank articles by cosine similarity to the review's focused questions. The best retrieval came from all-mpnet-base-v2, which surfaced all targets within the top few hundred to seven hundred results, while the lighter MiniLM variants delivered slightly lower similarity scores but far shorter encoding times. Reported runtimes on local Central Processing Unit (CPU) hardware and on Google Colab (CPU and T4 Graphics Processing Unit (GPU)) confirm substantial speedups for the smaller models and for GPU execution, with MiniLM producing 384-dimensional vectors and mpnet 768-dimensional vectors. The study concludes that transformer embeddings can meaningfully shrink the manual screening effort in evidence synthesis, and that an accuracy–latency trade-off favors MiniLM for fast iterations and mpnet when maximum recall is needed; an observation that aligns with my use of all-MiniLM-L6-v2 for large-scale title categorization and all-mpnet-base-v2 as a quality reference.

A cross-national study of Elango and Oh (2022) uses Scopus data accessed via SCImago to rank the top publishing countries and to relate output to economic and disciplinary profiles. It reports that 24 countries held at least one percent of global output in 2018, with the United States and China each surpassing 600 000 documents that year. The analysis tracks change across 1998, 2008, and 2018 using Compound Annual Growth Rate and Activity Index, and shows strong correlations between publication rank, Gross Domestic Product (GDP) rank, and Nature Index rank. Several countries increased their global share substantially over the period, and higher income groups dominate overall output. Notably, only a small subset of nations exceeded the world average activity in computer science. This macro view provides useful context for my venue and category analyses by showing how national capacity and disciplinary focus shape the landscape into which dblp publications and author populations fit.

# Chapter 3

# Gender Prediction and Analysis

This chapter documents the core methodological workflow. I parse the dblp XML corpus, clean and normalize the records, and prepare analysis-ready tables. I then assign research categories using sentence transformer models and infer author gender with GenderAPI. Together, these steps produce a consistent, reproducible dataset for the analyses that follow.

## 3.1 DBLP Database

The dblp computer science bibliography is a comprehensive, open bibliographic information service covering major journals and conference proceedings in computer science (dblp.org). Originating at the University of Trier, dblp is now curated and maintained by Schloss Dagstuhl – Leibniz Center for Informatics. Its sustained editorial curation and broad venue coverage make it a widely used primary source for bibliometric and scientometric studies in computer science. In this thesis, dblp serves as the sole data source for publication and authorship metadata.

The raw corpus comes as a single dblp XML file in which each bibliographic record is a top-level element. The principal element types are: **article** (journal or magazine papers), **incollection** (chapters in books/encyclopedias), **book** (complete monographs, typically with editors/International Standard Book Number (ISBN)/publisher), **inproceedings** (individual papers within conference proceedings), **proceedings** (the editors' volume for a conference), and academic theses (**phdthesis**, **mastersthesis**). dblp also includes **www** entries (e.g., personal or project webpages), which are not scholarly outputs; therefore **www** was excluded from our analysis. We parsed the XML and converted it to a normalized Comma-separated values (CSV) table, retaining standard fields across types (e.g., title, year, journal_or_booktitle, pages, and Digital Object Identifier (DOI)).

After parsing and converting steps, our final corpus contains 8,054,191 publication records. This scale is consistent with dblp's own milestone announcement that the bibliography has surpassed 8 million publications (blog post dated July 25, 2025), confirming that our extraction aligns with the current size of the database.

## 3.2 Data Cleaning and Preparation

After generating the final CSV, we loaded the corpus into a Pandas DataFrame for analysis and gender prediction. The section documents each column and our missing-data policy: records with non-essential gaps are retained; absent venue, pages, and DOI values are standardized to the sentinel "unknown"; and only records without an author list are excluded from author-level analyses.

*type* and *title* are complete (0 missing). *year* is essentially complete (3 rows; 0.00004%). In data, some works appear more than once, for example, encyclopedia chapters or reference book entries may have parallel incollection records, one with an explicit author list and another without. A concrete case is "Radiometric Camera Calibration" (see Figure 3.1), which appears twice: once in Computer Vision, A Reference Guide without authors, and once in the Wiley Encyclopedia of Computer Science and Engineering with the author listed. Because our analyses are author-centric (gender inference requires names) and to prevent double counting, we dropped all records with an empty authors field. This removed 59,909 rows (≈0.74%) from the raw corpus of 8,054,191 publications, yielding 7,994,282 records for the analytic dataset. This filter primarily excludes editor-volume placeholders and metadata-only entries, improving consistency without materially altering coverage.

| type | title | authors | year | journal_or_booktitle | pages | doi |
|---|---|---|---|---|---|---|
| incollection | Radiometric Camera Calibration. | NaN | 2014 | Computer Vision, A Reference Guide | 658 | https://doi.org/10.1007/978-0-387-31439-6_100172 |
| incollection | Radiometric Camera Calibration. | Leonard G. C. Hamey | 2008 | Wiley Encyclopedia of Computer Science and Engineering | NaN | https://doi.org/10.1002/9780470050118.ecse590 |

Figure 3.1: Example publications from the DataFrame, containing duplicate title name, one with and one without authors' names.

*journal_or_booktitle* is missing for 168,464 records (2.09%), largely theses (*phdthesis*, *mastersthesis*) and other types where the venue is captured elsewhere (e.g. school/publisher). *pages* is absent for 1,337,232 records (16.60%), often due to article-number schemes or online-first items. *doi* is not available for 1,318,738 records (16.37%), particularly in older entries or cases with only an *doi* link. For analyses that require valid venues or DOIs, we either exclude the

”unknown” category or report it explicitly; *pages* are not used in numeric computations.

We detected 1,021 exact duplicates (rows identical across all columns) and removed them, retaining a single canonical instance per record. This affects only ≈0.013% of the corpus and prevents double counting in all aggregates. When we checked potential duplicates using the subset of ’*title*‘,‘*authors*‘,‘*year*‘, it was 320,432 rows of publications. Manual spot checks showed these records typically differ by type and/or journal_or_booktitle. Since these are distinct bibliographic manifestations, they were keeped in the DataFrame.

Before performing gender prediction task, we constructed a canonical author list to avoid redundant predictions where the original DataFrame contains 7,993,261 publications now. Expanding these records into per-authorship rows yields 27,189,298 name instances, which would be computationally inefficient for API-based predictive model that will be discussed in the following section. We obtained 3,982,244 unique authors — a number consistent with dblp’s public statistics. Gender was then predicted once per unique author and merged back to the per-authorship table, reducing the number of API calls by approximately 85% while preserving complete coverage in downstream analyses.

## 3.3 Assigning Category to Publication

To enrich the gender-based analysis with a topical lens, I assign each publication to one of a consolidated set of computer-science categories. I evaluated two approaches: zero-shot text classification and sentence-similarity matching. Zero-shot models proved unstable on short titles and domain-specific terminology, so I do not pursue them further. The sentence-similarity approach is therefore the primary method.

Because no single authoritative taxonomy spans the breadth needed here, I compiled a curated list of ten high-level categories by harvesting and harmonizing labels from four reputable sources: arXiv (Computer Science), the IEEE Computer Society publications portal, PeerJ Subjects: Computer Science, and Cambridge University Press (Computer Science). Overlapping or synonymous terms were merged into a coherent table (see Appendix A, Figure A.1). This taxonomy underpins the similarity matching from publication titles and, when helpful, venue names.

Preprocessing and representation. Titles are lower-cased, punctuation-normalized, and stripped of non-informative tokens while preserving acronyms and diacritics. For each category, I concatenate its name, description, and subjects into a single descriptor. Titles and category descrip-

tors are then embedded into dense vectors with pretrained Sentence Transformers. I compute cosine similarity between a title embedding and all category embeddings and select the highest-scoring category. I retain the raw similarity score as a confidence proxy for later sensitivity checks.

| Model Name | Performance Sentence Embeddings | Performance Semantic Search | Avg. Performance | Speed | Model Size |
|---|---|---|---|---|---|
| all-mpnet-base-v2 | 69.57 | 57.02 | 63.30 | 2800 | 420 MB |
| multi-qa-mpnet-base-dot-v1 | 66.76 | 57.60 | 62.18 | 2800 | 420 MB |
| all-distilroberta-v1 | 68.73 | 50.94 | 59.84 | 4000 | 290 MB |
| all-MiniLM-L12-v2 | 68.70 | 50.82 | 59.76 | 7500 | 120 MB |
| multi-qa-distilbert-cos-v1 | 65.98 | 52.83 | 59.41 | 4000 | 250 MB |
| all-MiniLM-L6-v2 | 68.06 | 49.54 | 58.80 | 14200 | 80 MB |
| multi-qa-MiniLM-L6-cos-v1 | 64.33 | 51.83 | 58.08 | 14200 | 80 MB |

Table 3.1: A snippet of selected Sentence Transformer models (full table here on sbert.net). The models are evaluated on their ability to embed sentences (Performance Sentence Embeddings) and to embed search queries and paragraphs (Performance Semantic Search).

In line with public benchmarks from sbert.net (Table 3.1), I use three encoders to characterize the accuracy–latency trade-off: all-mpnet-base-v2 (strongest retrieval quality), all-MiniLM-L6-v2 (about five times faster with competitive quality), and all-MiniLM-L12-v2 (intermediate). In my environment (NVIDIA GeForce RTX 3050 Ti, batch size 1028), end-to-end assignment finished in 01:15:02 for all-MiniLM-L6-v2, 01:42:31 for all-MiniLM-L12-v2, and 03:18:52 for all-mpnet-base-v2. The primary model is all-MiniLM-L6-v2, chosen for its balanced quality and runtime; mpnet serves as a quality reference. I store per-title similarities from all three models, which allows agreement checks and simple tie-breaks (e.g., if two models concur and the third is close).

Automated assignment is complemented by a deterministic venue→category mapping for the most prolific and clearly scoped outlets. If a venue is single-field by editorial design, the curated label overrides the model output. Generalist venues retain the model assignment. This hybrid strategy reduces cross-category leakage and improves stability while remaining reproducible.

I conduct manual spot checks on samples drawn across venues and years, inspect low-confidence tails, and compare distributions across the three encoders. Analyses that follow report category results as assigned and, where informative, include brief sensitivity comments using the stored similarity scores. The combination of text similarity and venue curation yields consistent and reproducible topical assignments for downstream gender analysis.

Title-only classification is inherently sparse and can be sensitive to abbreviations, transliteration, and venue-specific jargon. The curated taxonomy is broad by construction, and fine-

grained subfields are aggregated into the ten categories for tractability. These choices are documented so that future work can refine the taxonomy or add abstract-level text where available.

The chart (see Figure 3.2) compares category assignments produced by three sentence similarity models. Bars show the number of publications per category, with percent shares annotated. Overall, the rankings are broadly consistent — Artificial Intelligence & Machine Learning and Theory of Computation & Algorithms are among the largest categories but each model shifts mass differently.

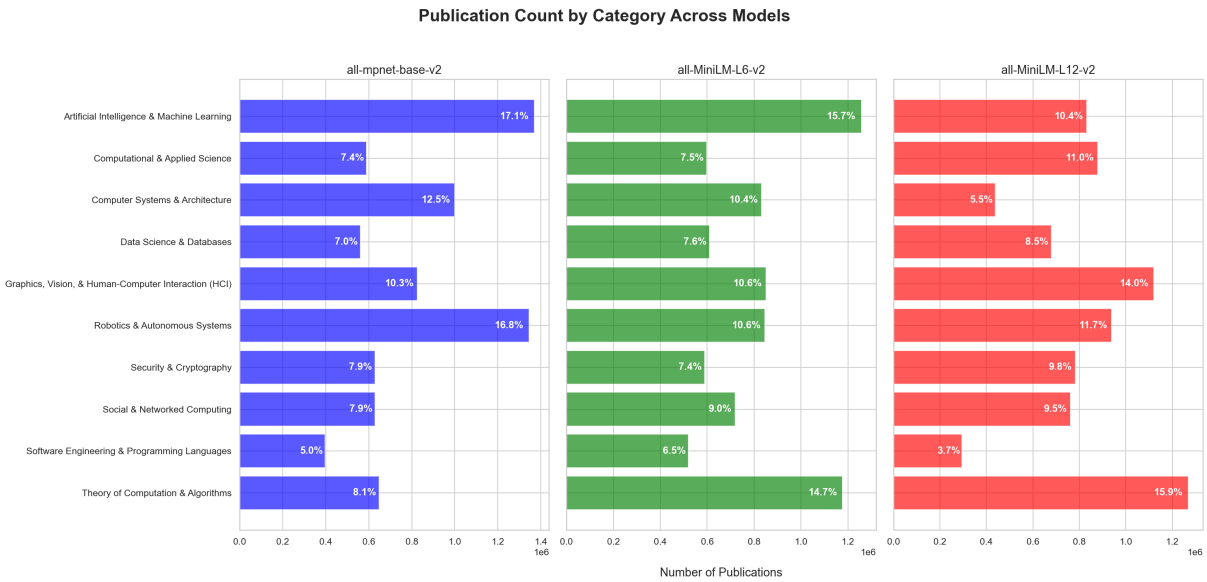


Figure 3.2: Bar charts showing publication counts by category across three models.

Figure 3.3 demonstrates distributions where the x-axis is the similarity score (title ↔ category/description cosine), and the y-axis is the number of publications. All three models produce a single bell-shaped distribution with heavy overlap, peaking around 0.18–0.25. With any fixed similarity threshold, mpnet would classify the most items; L12 the fewest; L6-v2 is a balanced middle ground. Combined with its much faster runtime, this supports the idea of using all-MiniLM-L6-v2 as the primary model in the thesis.

After all these observations I selected all-MiniLM-L6-v2 as the primary model because it provides a balanced category mix (neither overconcentrated nor overly diffuse) and the best speed–quality trade-off in our setting. On an NVIDIA GeForce RTX 3050 Ti (batch size 1028), wall-clock times were:

- all-MiniLM-L6-v2: 01:15:02 (1 hour and 15 minutes and 2 seconds)
- all-MiniLM-L12-v2: 01:42:31 (1 hour and 42 minutes and 31 seconds)
- all-mpnet-base-v2: 03:18:52 (1 hour and 18 minutes and 52 seconds)

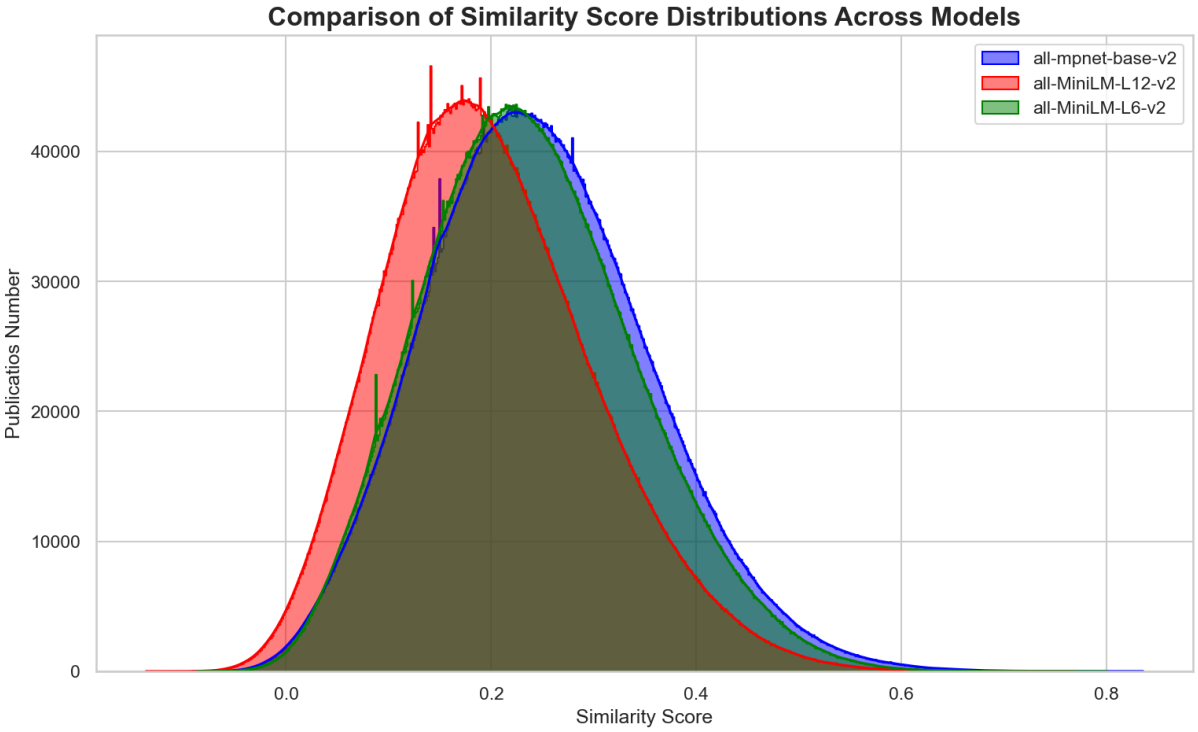


Figure 3.3: Distributions of similarity scores across three models.

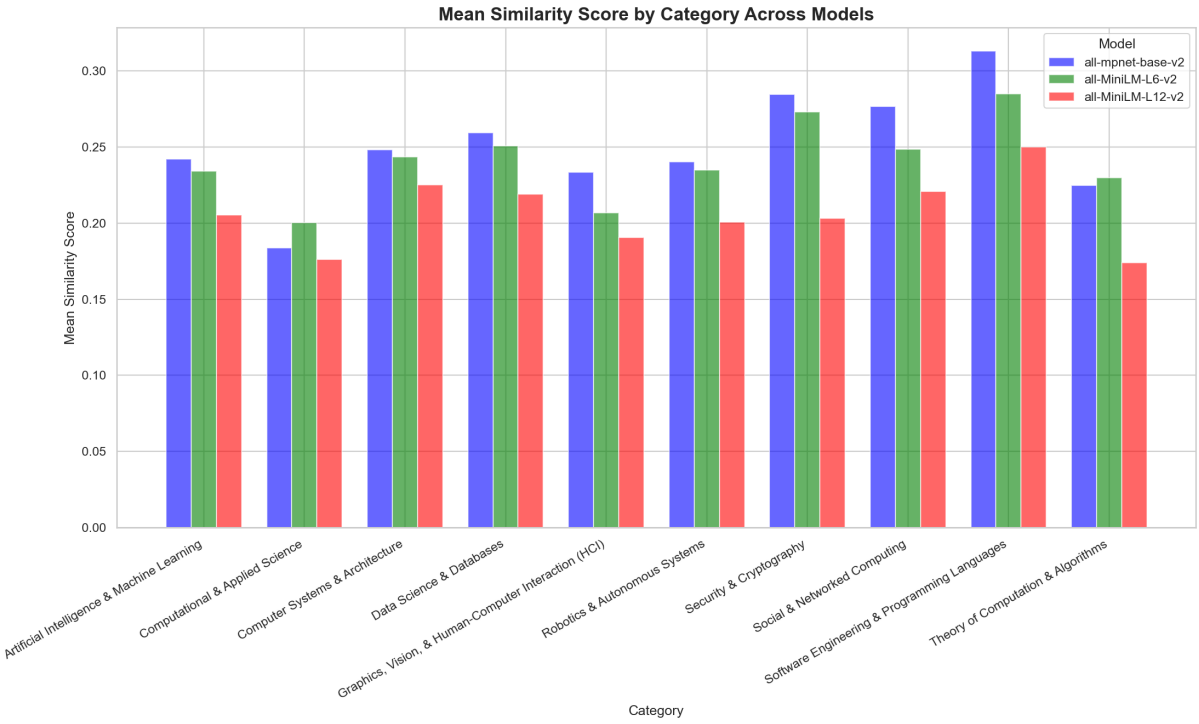


Figure 3.4: Bar charts of mean similarity scores by category across three models.

Thus, L6-v2 was approximately 2.6 times faster than mpnet while producing distributions that align with domain expectations, making it the pragmatic default for the rest of the analysis.

To improve topical precision beyond automated title matching, I added a venue level curation step. First, I identified the 80 most popular venues and, among them, those with more than

10,000 publications. For each high volume venue, I reviewed its editorial scope (single field vs. all fields) and built a deterministic venue category mapping. Venues with a clearly bounded scope were hard-assigned to that category (e.g., ”Computer Vision, A Reference Guide” to Graphics, Vision, & Human-Computer Interaction (HCI)), while generalist venues retained the sentence-similarity assignment. I then overwrote the model’s labels with the curated venue labels where applicable.

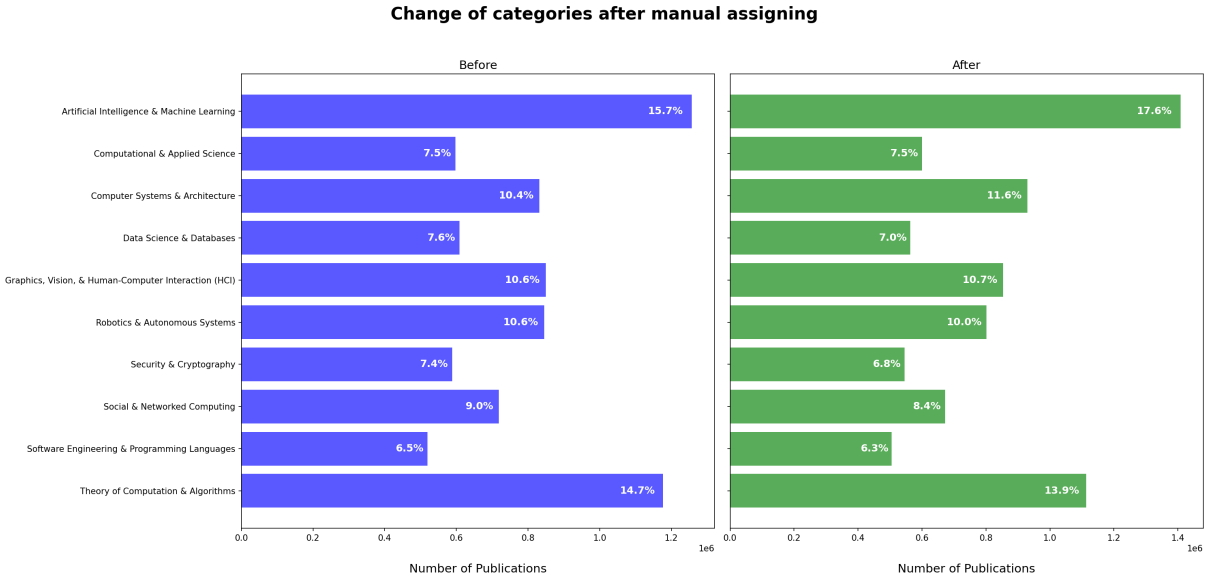


Figure 3.5: Bar charts showing change of categories after mapping.

As shown in Figure 3.5, this produced moderate but systematic shifts (e.g., increases in Artificial Intelligence & Machine Learning and Computer Systems & Architecture, with a corresponding reduction in Theory of Computation & Algorithms). This hybrid procedure model predictions refined by venue rules, reduced cross-category leakage and improved alignment with domain knowledge, while remaining reproducible via the explicit venue category table.

## 3.4 GenderAPI

GenderAPI (https://www.genderapi.io), developed by Ozan Soft, is a service that infers a likely gender from textual identifiers such as a first name, full name, email address, or username. The tool serves simple users, researchers, data analysts, and developers by enriching datasets with gender information, which can be pivotal for demographic analysis and other data-driven studies. GenderAPI’s use of artificial intelligence and a large name database forms the basis of its functionality. The system’s AI-powered models examine the input against its vast name repository and statistical patterns when a query is submitted. In addition to its own proprietary

tools, the service uses an ensemble methodology, utilizing several cutting-edge AI technologies at once, such as large language models such as ChatGPT, Grok, LLaMA, and DeepSeek. The highly accurate gender determination that is produced by this parallel analysis method is given back to the user along with a confidence score that shows the probability of the result. The service supports non-Latin scripts and a wide range of languages (e.g., Chinese, Japanese Hindi, Arabic), making it suitable for global datasets.

I used GenderAPI to predict genders for authors in the dblp corpus. Queries were sent with full names (no truncation to given names), and the returned probability was normalized to [0, 1] for analysis. All results are treated as probabilistic, meaning the aggregate statistics incorporate each prediction's confidence score, gender label, and country of origin.

## 3.5 Gender Prediction

I performed gender prediction on unique authors DataFrame that was discussed in previous section (see Section 3.2) using GenderAPI's (see Section 3.4) bulk endpoint and a checkpointed, append-only workflow designed to be resumable and fault-tolerant. The workflow follows as:

1. **Resume logic / checkpointing.**
   Before starting, I load any previously saved results from the output CSV (reading only author_id to save memory) and build a processed_ids set. This lets the script resume exactly where it left off after interruptions, without re-querying authors already processed.
2. **Chunked bulk requests.**
   I split the unique authors DataFrame into chunks of 100 rows (the payload for one bulk request). For each chunk I filter out already processed IDs, so only new authors are sent to the API. Each request uses full names and includes the author_id as an id field so results can be matched back deterministically.
3. **Result handling and persistence.**
   When a response arrives, I convert the names list into a DataFrame, select the relevant fields id, gender, probability, country rename id → author_id, and cast author_id to int64 for consistent merging later. I append these rows to the CSV (writing the header only if the file does not yet exist) and immediately update processed_ids. This streaming write keeps memory use low and ensures progress is saved continuously.
4. **Robustness.**
   The loop prints progress for each chunk (processed count and totals). If the API returns

no results for a chunk, I log that and continue. Any exceptions are caught and logged; the chunk is skipped so the overall job does not crash, and the script can be rerun to retry only missing authors thanks to the checkpoint.

5. **Outcome.**
   The procedure yields an incrementally built results file with one row per processed author (author_id, gender, probability, country). Because the process is idempotent with respect to author_id, it avoids duplicate API calls and prevents double counting. Downstream, I normalize probability to [0,1] and keep unknown predictions as a separate category for uncertainty-aware analysis.

After the bulk run, the results file still had gaps. 163,648 missing gender and 165,335 missing country. A large share of the unresolved cases are records where dblp lists the given name only as initials, for example, *"A. V. Olifer," "A. A. Servedio," "X. Wu," "Z. M. Ma."* Such abbreviations (single-letter or multi-initial forenames) provide too little lexical signal for name-based inference, especially across languages and transliterations. Consequently, GenderAPI returns null for these entries. To target these "hard" cases, I executed a follow-up pass that queries one author at a time with the vendor's heavier ensemble setting askToAI=True (see Listing 3.1).

```
result = api.get_gender_by_name(author_name, askToAI=True)
```

Listing 3.1: A snippet of a single request to API forcing usage of AI

The second AI assisted pass used single name queries with askToAI=True for the unresolved cases and finished in 143,615.39 s (≈ 39 h 53 m 35 s), averaging about 0.9 s per author. Because this approach is much slower than bulk queries, I applied it only to the difficult subset: bulk for speed, single askToAI for recall. During this pass two records failed with HTTP 504 Gateway Timeout errors ("Bakhtiyorjon Bakirovich Akbaraliev" and "A. K. Bedyal"), which I resolved by manual verification and entry. After completion, 29,937 authors still lacked a confident label; I left these untouched and coded their gender as "unknown", which remains a separate category in all aggregate analyses.

As a result most predictions are extremely confident. In Figure 3.6, the density of probabilities p≥0.5) piles up near 1.0 for both male and female labels, with only a thin tail in the 0.5–0.9 range. The male and female curves almost sit on top of each other, and the dashed "overall" line follows the same shape, which suggests the model behaves similarly across genders. In practice, this means the bulk of our gender assignments are made with very high certainty; the

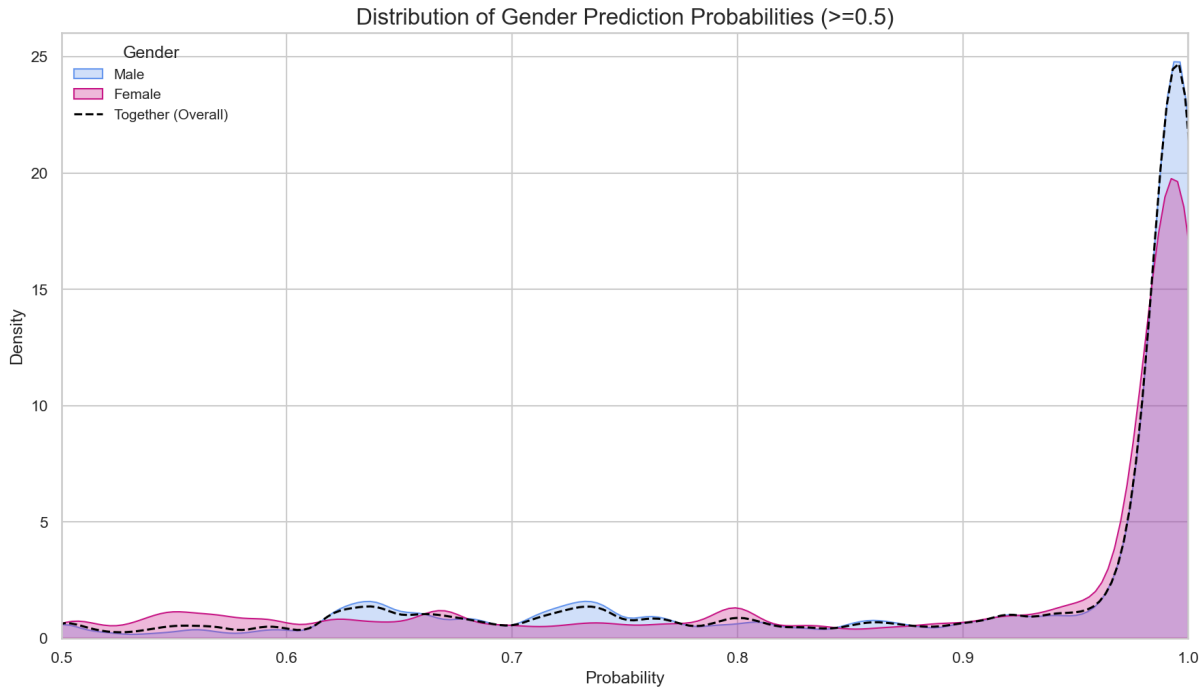


Figure 3.6: Distribution of gender-prediction probabilities ($p \geq 0.5$) by assigned gender.

few mid-range cases exist but are comparatively rare and should be treated with more caution in downstream analyses.

# Chapter 4

# Results and Analysis

This chapter presents the main empirical results from the dblp corpus. I first analyze publications over time and by document type to establish context. I then examine authors' gender composition overall and by year, venue, and research category, with focused views of the most prolific outlets. Country patterns are summarized using the service's country labels as a proxy for name origin and related back to venues and categories when informative. Category labels come from the sentence-similarity method refined with venue curation, and gender inference comes from GenderAPI with confidence scores; I report both probability-weighted and thresholded estimates to check robustness. The chapter closes with a brief synthesis of the main patterns and their implications for representation across computer-science research.

## 4.1 Publication Analysis

This section summarizes the longitudinal output of computer science publications recorded in dblp.

The corpus begins in 1936, but counts prior to the mid-1980s are extremely small. For readability I therefore plot this range of years 1984–2026 (see Figure 4.1). From 1984 through the late 1990s the series rises steadily from a few tens of thousands of records per year to approximately 100 thousand. Growth accelerates in the 2000s and 2010s, reflecting the expansion of conference proceedings and digital publishing, with annual totals surpassing 400k in the early 2020s and reaching a local maximum in 2024 in this extraction. The apparent collapse at 2026 is not substantive: dblp occasionally lists forthcoming proceedings and volumes under their nominal publication year, and coverage for the most recent year(s) is incomplete at the time of data collection. Consequently, values at the right edge should be interpreted as provisional.

Unless otherwise stated, the counts include all dblp publication types used in this study (journal articles, conference papers, books, chapters, theses; excluding www entries) and are aggregated by the year field after the data cleaning steps described earlier.

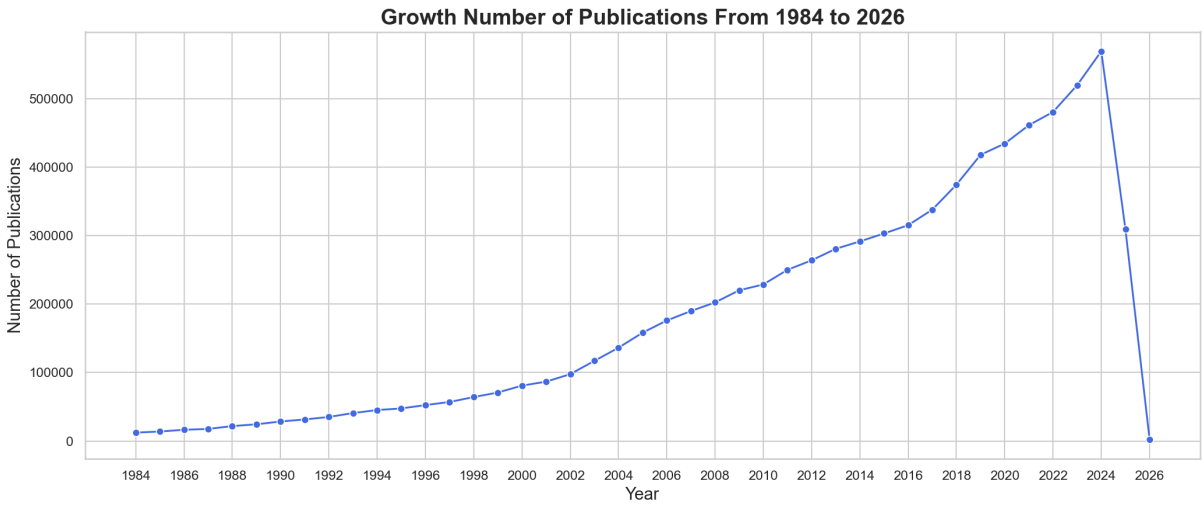


Figure 4.1: Number of computer science publication over the years in dblp

Here (see Figure 4.2) the yearly totals are split by publication type in dblp (1984–2026). I did not include master's theses because they have very low (27) numbers. Two patterns dominate:

- **Articles** and **inproceedings** account for almost all growth. **Articles** rise steadily through the 1990s, then accelerate sharply after 2005, exceeding a few hundred thousand per year by the early 2020s. **Inproceedings** follow a similar trajectory strong expansion from the late 1990s to 2020—reflecting the conference-centric culture of computer science and the proliferation of journals, special issues, and digital publication workflows.
- **Books**, **proceedings volumes**, **incollections**, and **theses** remain comparatively small. CS publishes relatively few monographs (hence a flat "book" series). **Proceeding** in dblp refer to the edited volume record (one entry per conference year), not the individual papers—so counts are naturally low and stable. **Incollection** (book or encyclopedia chapters) shows only modest numbers with occasional bumps when large reference works appear. PhD **theses** grow gradually then taper, consistent with dblp's partial coverage of theses rather than comprehensive indexing.

Overall, the figure confirms that the long-run expansion of dblp is driven primarily by journal articles and conference papers, with other types contributing only marginally to total volume.

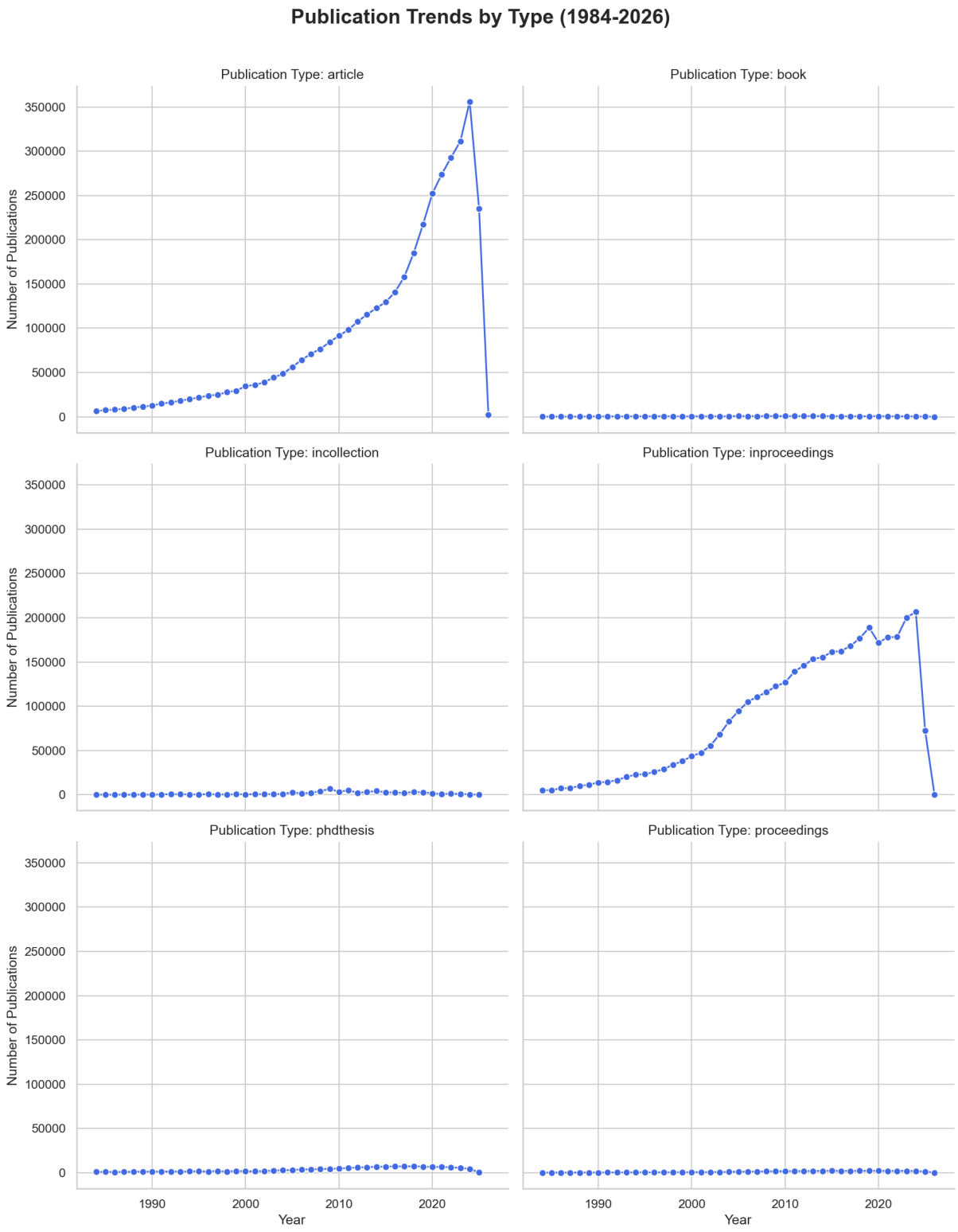


Figure 4.2: Growth of dblp records by type: articles and inproceedings dominate (1984–2026).

## 4.2 Author Gender

This section presents an overall analysis of the unique authors in the dblp corpus. The pie chart (see Figure 4.3) summarizes the distribution of inferred author gender in the dblp corpus (unique author list used for prediction). The population is highly male-skewed: roughly three quarters of authors are classified as male, about one quarter as female, and <1% remain unknown. The small “unknown” slice reflects the limited number of cases for which a confident assignment could not be made.

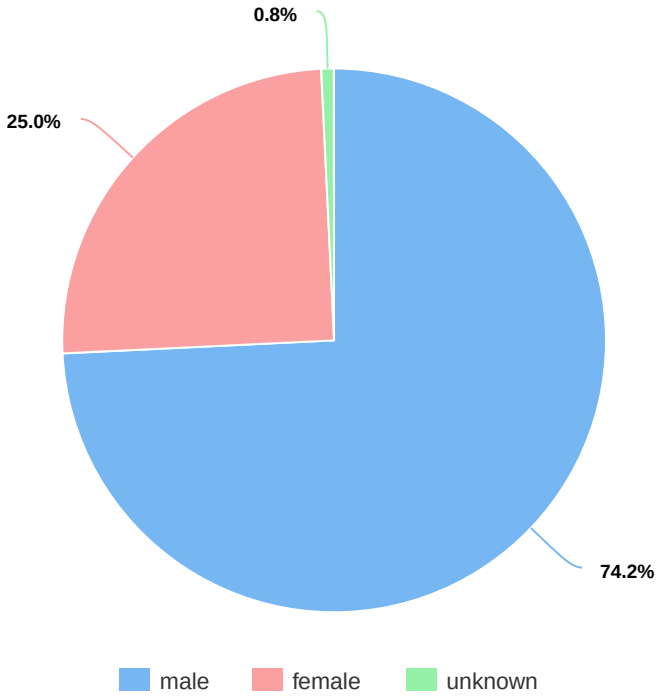


Figure 4.3: Overall inferred author gender distribution in the dblp corpus

The choropleth displays (see Figure 4.4) the top 20 countries by authors’ predicted country labels returned by GenderAPI responses. These labels should be read as a name origin proxy, not as verified nationality, residence, or institutional affiliation. In practice, a researcher with a Japanese given name working in Italy would still be labeled ”JP”. Within the Americas, the United States dominates ( about 1.3M labeled authors), followed by Brazil (about 294k) and Mexico (about 105k). The prominence of Mexico is analytically interesting: in the scientometric literature Mexico does not consistently rank among the top producers of computer science publications. This discrepancy likely reflects onomastic conflation across Spanish/Portuguese naming conventions common to Latin America and the diaspora, which can bias a name country classifier toward ”MX” even when the true affiliation lies elsewhere. Put differently, the country

label captures how the name "sounds" statistically, not where the author works or publishes.

Across Asia, large signals appear for Malaysia, India, China, Indonesia, and Japan, with additional contributions from Turkey, Thailand, Iran, Korea, Saudi Arabia, and Vietnam. Several of these patterns are consistent with shared linguistic and religious naming traditions. For example, Arabic-derived given names are common across Malaysia and Indonesia, which can blur country boundaries in a name-based model. Romanization further complicates attribution: pinyin for Chinese, Revised Romanization for Korean, and varying transliteration conventions for Russian/Arabic can cluster distinct national origins into a single country label.

Within Europe, higher counts are visible in Russia, Germany, Italy, Belgium, France, and Greece. Darker shading in Western/Central Europe is broadly consistent with long-standing participation in CS publishing and dense research networks. At the same time, diaspora effects (e.g., Indian or Chinese names in U.S./EU institutions) mean that name-origin and affiliation geography will not coincide one-to-one as I said before.

Hereafter I discuss results continent by continent. For each continent I include the most frequent countries in the dblp author unique list; the panels are therefore representative, not exhaustive (e.g., Europe may omit lower-frequency countries).

The figure Figure 4.5 shows kernel-density distributions of GenderAPI prediction probabilities (restricted to scores $\geq 0.5$) for the United States (US), Brazil (BR), and Mexico (MX), broken out by predicted gender with a pooled curve (black dashed). All three countries exhibit a strong spike near 1.0, indicating that the large majority of assignments are high confidence.

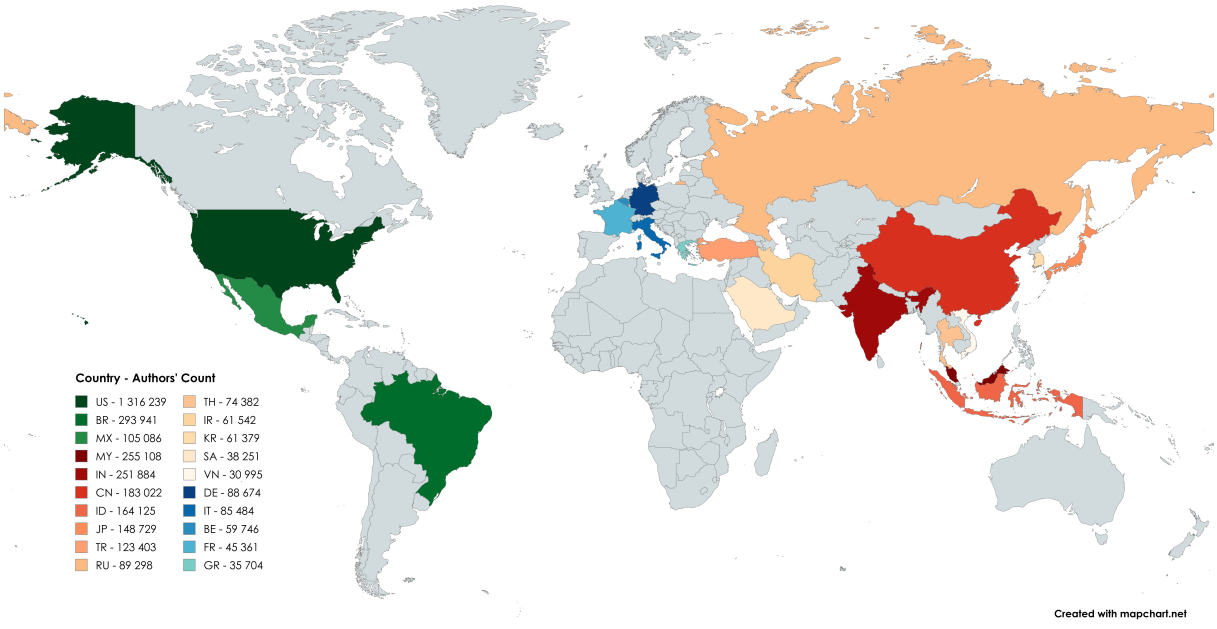


Figure 4.4: Top 20 countries by authors' predicted country labels (GenderAPI).

The US panel shows a slightly wider shoulder around 0.65–0.85, suggesting more ambiguous cases (e.g., initials, cross-lingual names). Brazil is the tightest around 1.0, while Mexico shows a small secondary bump near 0.8, consistent with the earlier caveat that Spanish/Portuguese naming shared across Latin America can introduce ambiguity. Overall, the distributions support using probability-weighted summaries: low-confidence tails are present but minor relative to the mass at high confidence.

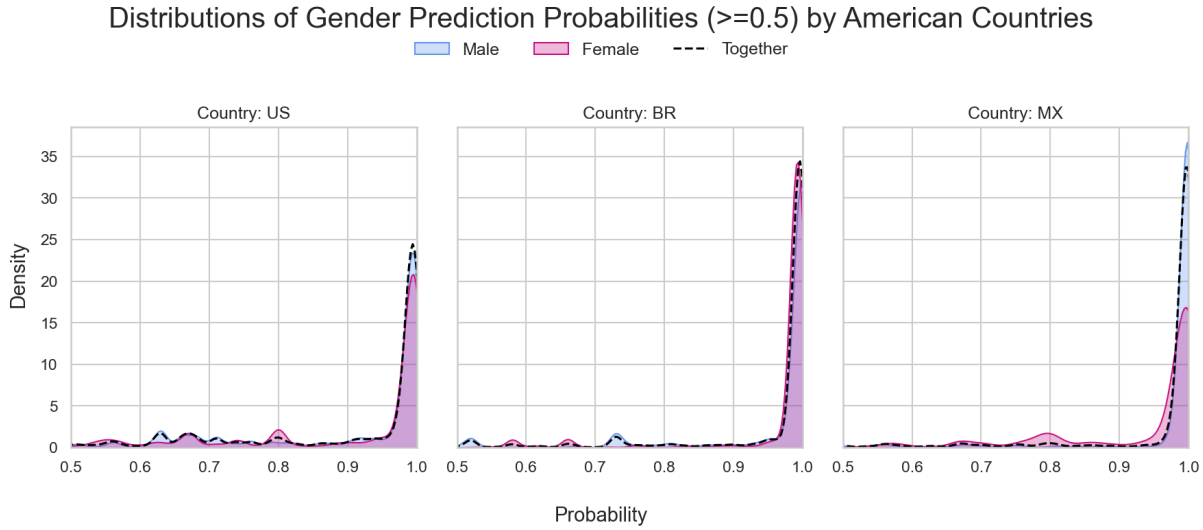


Figure 4.5: Distributions of gender–prediction probabilities (scores $\geq$ 0.5) for the Americas: United States, Brazil, and Mexico.

Asian countries are especially interesting to examine because name formation and transliteration make gender and country inference harder. Figure Figure 4.6 shows the distributions of GenderAPI prediction probabilities (scores $\geq 0.5$) for the most frequent Asian labels in the author list: MY (Malaysia), IN (India), CN (China), ID (Indonesia), JP (Japan), RU (Russia), TH (Thailand), IR (Iran), and KR (South Korea). Several panels have a sharp peak near 1.0, especially India, Iran, Russia, and Japan, which suggests that once rendered in Latin script many names carry very distinctive patterns. China and Korea display broader shoulders around 0.6–0.85, consistent with ambiguity introduced by Romanization and shared syllabic structures. Malaysia and Indonesia also have wider spreads, which fits the mix of Malay and Arabic naming traditions across the region.

Thailand is the most unusual case. Thai surnames are legally required to be unique to a family and are often long, so truly Thai names tend to yield very confident scores (for example, Nachatchapong Kaewsompong 0.98). In our data, however, the TH label also includes non-Thai names such as Tao Jiang 0.79, Qizhi Fang 0.77, Jinny McGill 0.94, and Ping Deng 0001 0.50; the "0001" suffix is a DBLP disambiguation marker rather than part of the name. This spillover

illustrates a general limitation of country inference from names: the label reflects likely name origin, not citizenship, residence, or institutional affiliation. This says that not only ”TH” label has this issue, there might be another country labels that can have names from another countries.

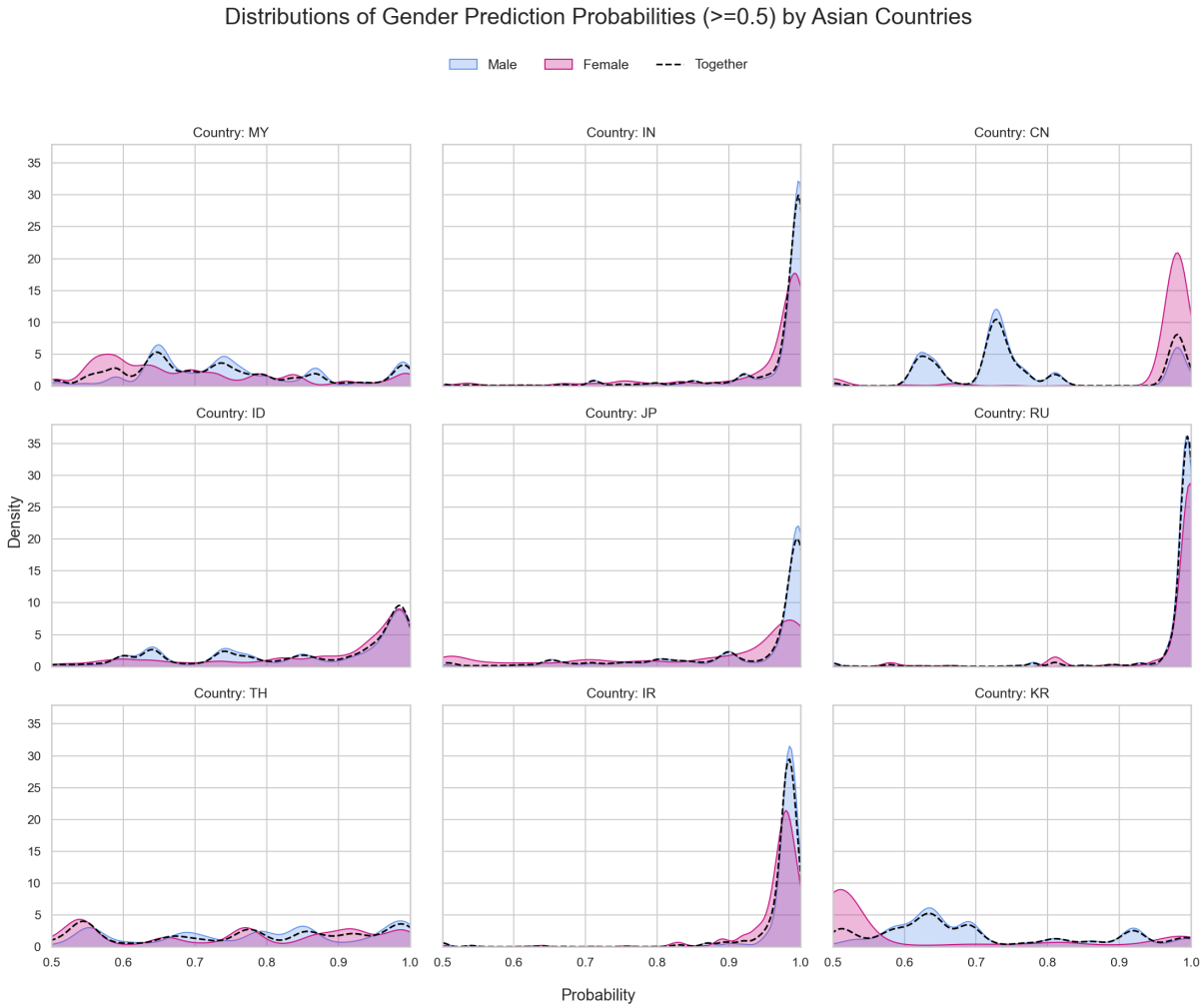


Figure 4.6: Distributions of gender prediction probabilities (scores $\geq 0.5$) for selected Asian countries.

Coming to European countries (see Figure 4.7), the distributions are tightly concentrated near 1.0 in every panel, which indicates very high confidence for most assignments. Germany, Italy, Belgium, France, Greece, Poland, the Netherlands, and Romania show a narrow spike close to one with only a faint shoulder between 0.6 and 0.9. Finland has a slightly broader tail around 0.90 to 0.97, which may reflect shorter or less strongly gender marked given names, as well as some spillover from non-Finnish names in the dataset. Small mid-probability bumps in several panels likely arise from initial-only records and cross-linguistic or immigrant names. As throughout, the country tag signals likely name origin rather than nationality or affiliation. Generally, European names are the most easiest to investigate, because of their structurally consistent and strongly gender-marked naming conventions.

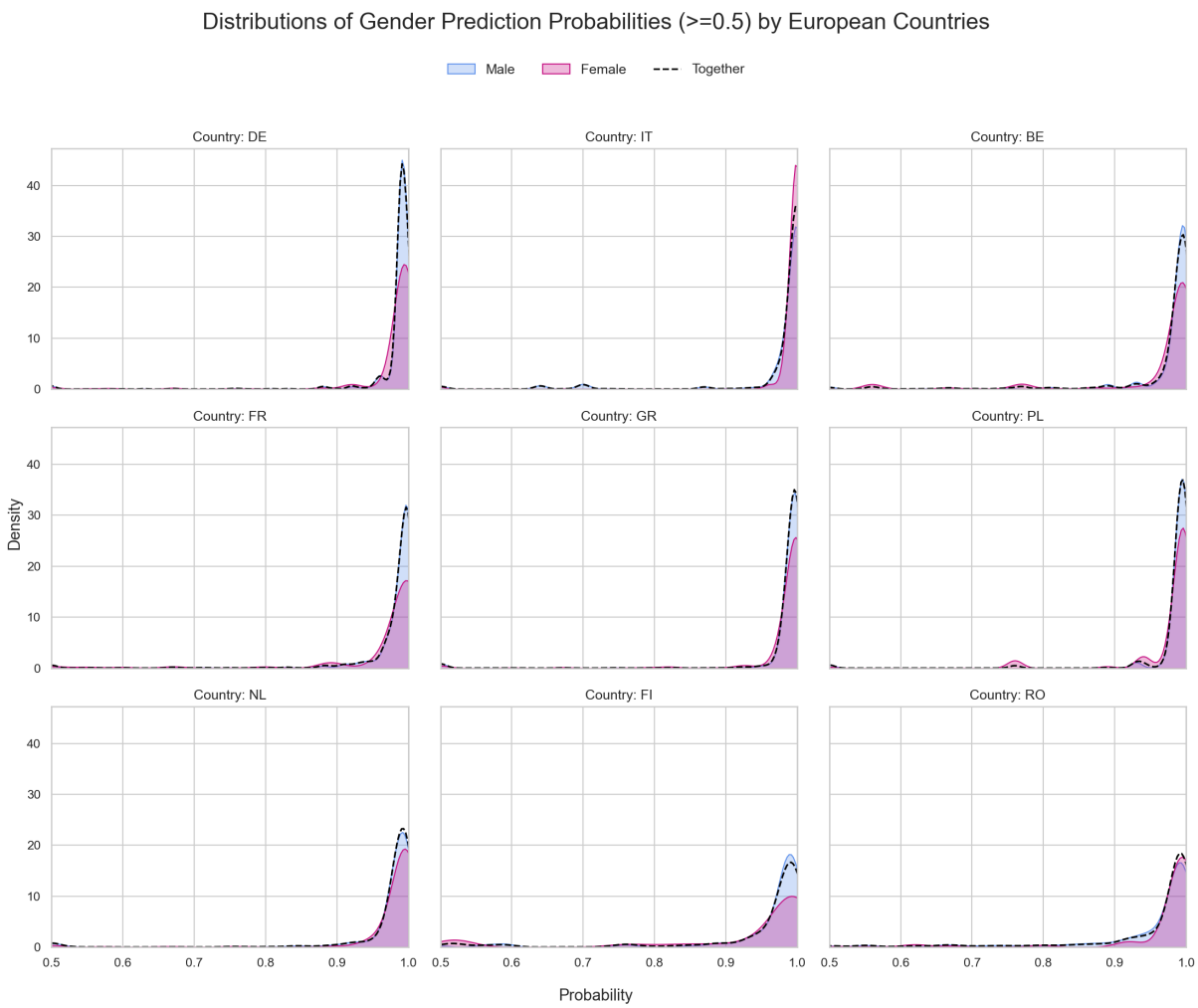


Figure 4.7: Distributions of gender prediction probabilities (scores $\geq 0.5$) for selected European countries.

Overall, the country findings reflect patterns of name origin, not verified nationality or affiliation. Most probability distributions are concentrated near 1.0, which indicates confident assignments at scale. Where the curves are wider, common causes include initials in place of given names, transliteration and romanization differences, shared naming conventions across languages, and diaspora effects. To preserve transparency, I retain an unknown category, use probability-weighted summaries, and treat cross-country contrasts with caution.

GenderAPI is designed primarily for gender inference. The accompanying country label should be viewed as metadata inferred from name–country frequency statistics or as context that can improve gender prediction through optional country filters. It is not a verified attribute of the person. In this thesis the country field is used descriptively, and geographic claims should be corroborated with affiliation data in future work.

The final Figure 4.8 in this section shows 100 percent stacked bars for each year, with the share of male and female authors in the dblp corpus. Two features stand out. First, the very

early years contain few records, so the female share fluctuates widely and sometimes appears higher than in later decades simply because the denominator is small. Second, once annual volume grows, the pattern becomes stable and shows a gradual rise in the female share over time. From the 1980s onward the female share increases from roughly ten percent to about twenty to twenty-two percent in the early 2020s, while the male share declines by the same amount but remains the clear majority. The small gray segment reflects cases without a confident assignment and is most visible in the early decades; it becomes negligible in recent years as coverage improves. Values at the rightmost edge should be read with caution because the most recent year may be incomplete. It is also reasonable to assume that GenderAPI struggled more with older entries, where names are often written with initials, metadata are sparse, and transliteration is inconsistent.

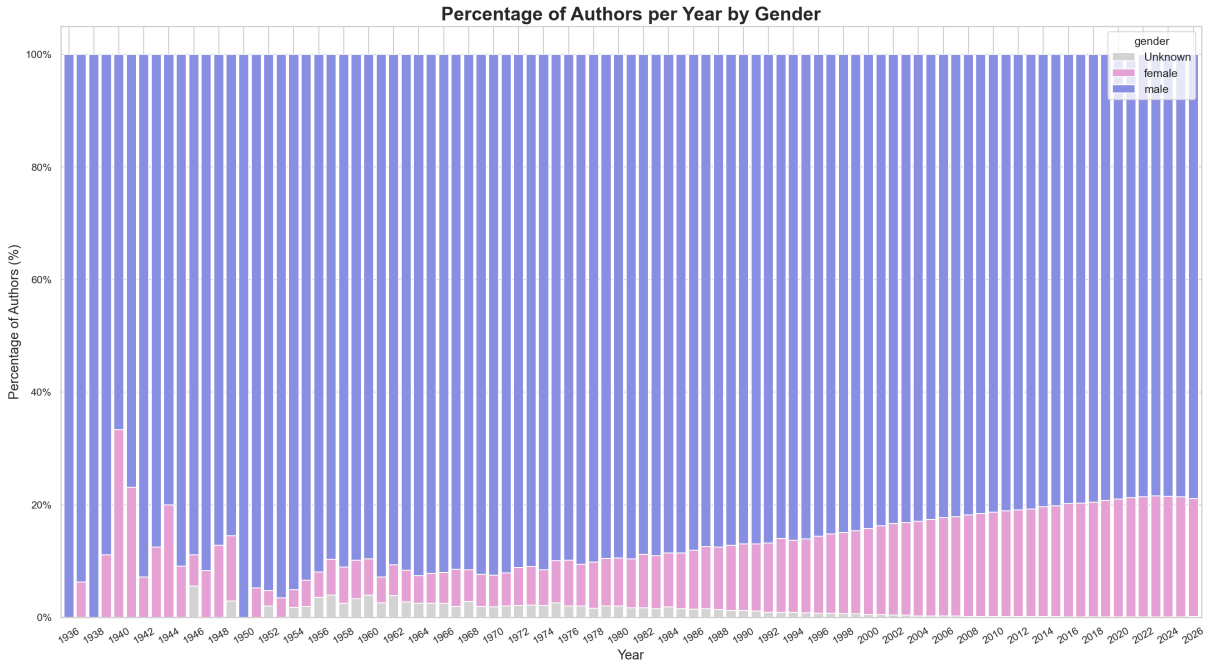


Figure 4.8: Distributions of gender prediction probabilities (scores $\geq 0.5$) for selected European countries.

## 4.3 Gender Distribution by Venue and Assigned Category

This section discusses gender distribution by venue and by assigned research category. The goal is to see how much women and men contribute across the most active outlets and, by extension, across major areas of computer science.

The Figure 4.9 reports the percentage of authorships within the ten most prolific venues in the corpus. Several patterns emerge. Journals in sensing and Earth observation show the highest female shares: Remote Sens. at about 22.7% and Sensors at about 21.8%, with IGARSS close by at 21.5%. Broad outlets such as IEEE Access and CoRR sit in the middle with female shares near 19.3% and 18.1%. Research-intensive conference venues in robotics and signal processing show the lowest female shares: ICRA at about 13.1%, IROS at about 13.3%, and ICASSP at about 15.8%. Computer vision and machine learning are in the mid range with CVPR at 17.8% and NeurIPS at 16.8%.

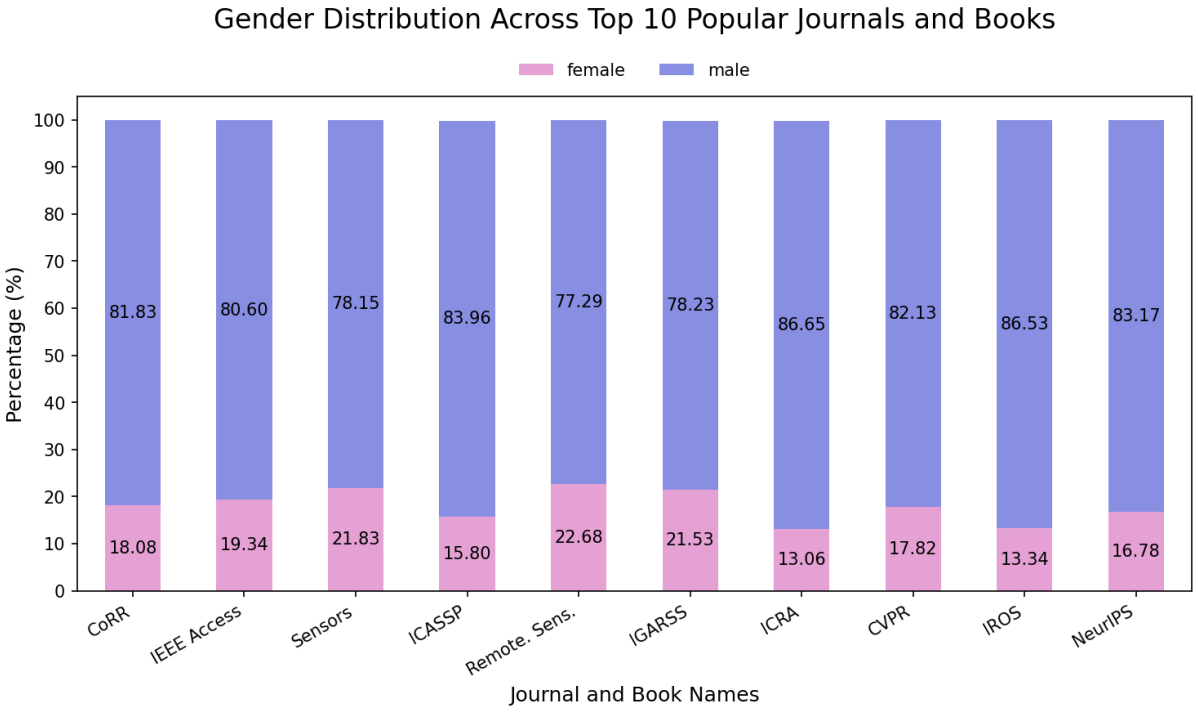


Figure 4.9: Percentage of authorships by gender in the 10 most prolific dblp venues (journals and conferences).

These differences are consistent with long-standing field compositions and with the editorial scope of the venues. Sensing and geospatial outlets draw from communities with a stronger representation of women in adjacent disciplines, while robotics and certain engineering-heavy areas remain more male dominated. CoRR functions as a general preprint repository, so its composition tracks the overall average. The chart summarizes authorships rather than unique

people, which is appropriate for measuring contribution to venue output. In the following pages I relate these venue-level results to the topical categories derived from the sentence-similarity and venue-curation pipeline and show that the same contrasts appear at the category level as well.

The Figure 4.10 tracks the annual percentage of female and male authorships within the ten most prolific venues. Each panel shows a separate outlet, so we can compare how gender composition evolves within a venue rather than across the entire corpus. Across all panels the same broad pattern appears: as annual volume grows, the female share rises gradually while the male share declines by the same amount. Early bars in several panels show larger swings because the number of papers in those first years is small; once venues mature the trajectories become smoother.

In CoRR the female share starts in the low teens and climbs toward roughly one fifth by the most recent years. IEEE Access shows a similar upward drift from the teens into the vicinity of twenty percent. These outlets cast a wide net across computer science, and their time series largely mirror the field-level trend.

Sensors, Remote Sensing, and IGARSS exhibit some of the highest female shares among the top venues. Sensors and Remote Sensing begin around the high teens and move into the low to mid twenties. IGARSS follows a comparable pattern with a steady incline across the window. Year to year variability is present in the early stages, but the trajectory is clearly increasing.

ICASSP maintains one of the lowest female shares for most of the period, typically in the low to mid teens, although the last few years show a modest rise. ICRA and IROS display a similar profile. Both begin near ten percent and move upward into the mid teens by the end of the period. These venues remain the most male dominated among the group, which aligns with discipline-specific patterns reported in prior work.

CVPR shows a long series with a visible increase from single digits in the earliest years of the panel to the high teens more recently. NeurIPS appears later in the window and shows female shares in the mid teens with a slight upward slope. These levels are below those of Sensors and Remote Sensing but above those of the robotics conferences.

Gaps at the beginning of some panels reflect the fact that a venue either did not yet exist or was not yet covered by dblp. Small gray slivers correspond to cases without a confident assignment and are more visible in the earliest years. Percentages are computed on authorships rather than unique people, which is appropriate for measuring contribution to venue output. Taken together, the panels show a consistent, gradual improvement in female representation across

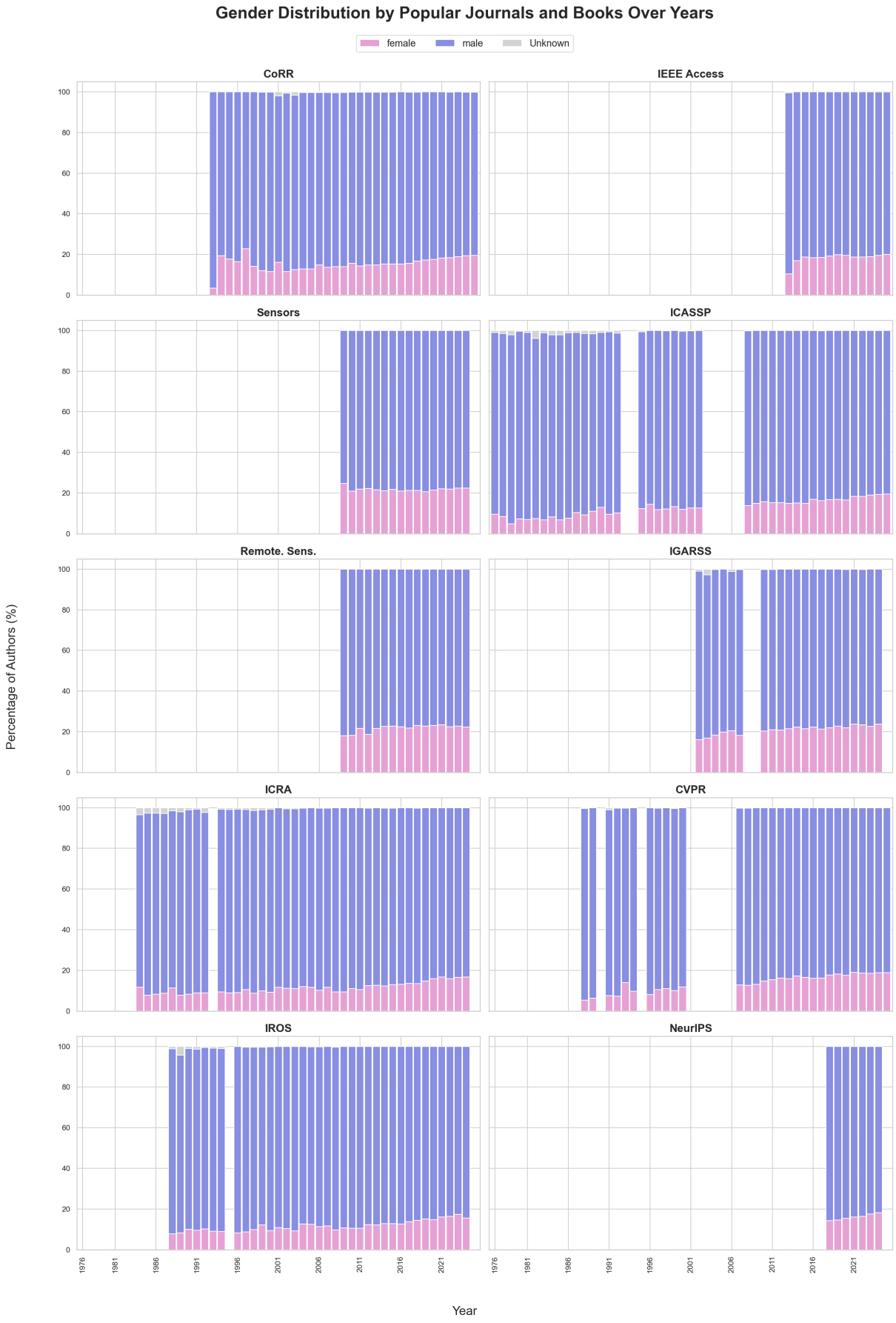


Figure 4.10: Annual gender composition of authorships in the ten most prolific dblp venues.

leading journals and conferences, with substantial variation by field and venue that persists to the present.

Now let us move to the assigned categories. The following Figure 4.11 shows total authorships per category (bar length) and the share of women and men within each bar. Across the corpus, Artificial Intelligence and Machine Learning accounts for the largest volume with about 5.35 million authorships, followed by Computer Systems and Architecture (3.37 million) and Graphics, Vision, and HCI (3.22 million). Mid-sized groups include Theory of Computation and Algorithms (2.93 million), Robotics and Autonomous Systems (2.88 million), and Computational and Applied Science (2.17 million). Social and Networked Computing, Data Science and Databases, Security and Cryptography, and Software Engineering and Programming Languages range from roughly 1.5 to 2.0 million.

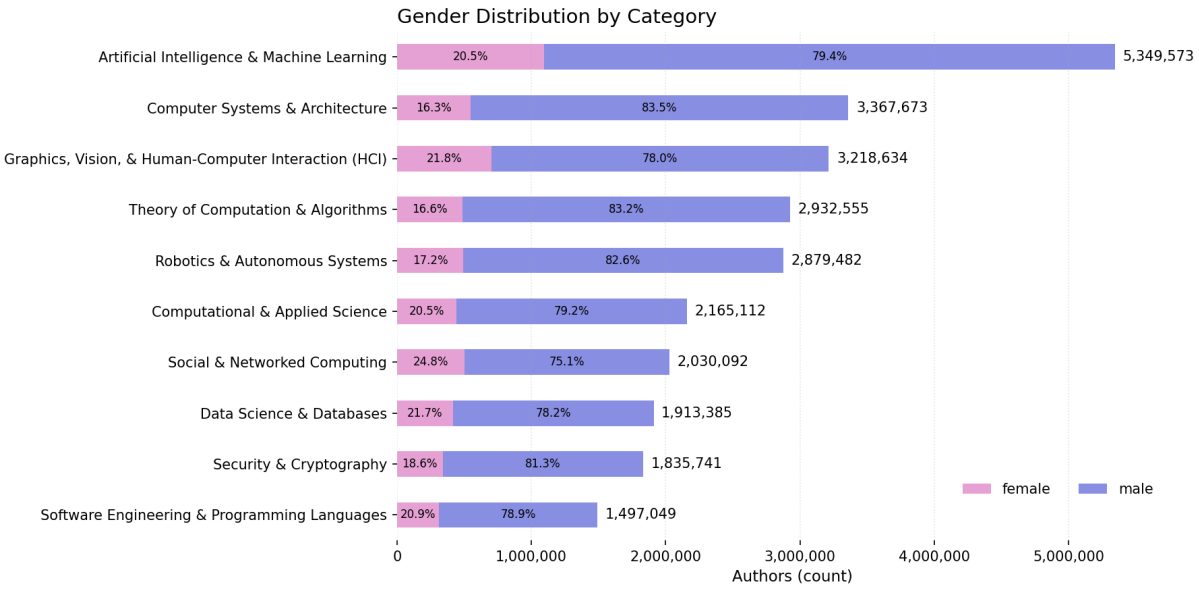


Figure 4.11: Gender distribution by assigned research category.

Female representation varies by field within a relatively narrow band of about ten percentage points. The highest shares appear in Social and Networked Computing (about 24.8%) and Graphics, Vision, and HCI (about 21.8%), with AI and ML, Data Science and Databases, and Software Engineering and Programming Languages clustered around 20–21%. Lower shares are found in Security and Cryptography (about 18.6%), Robotics and Autonomous Systems (about 17.2%), Theory of Computation and Algorithms (about 16.6%), and Computer Systems and Architecture (about 16.3%). These contrasts mirror long-standing differences across subfields: areas oriented to people, data, and applications tend to have higher female participation, while core theory, systems, and robotics remain more male dominated. Counts are authorships rather than unique individuals, which is appropriate for measuring contribution to category out-

put. Category labels come from the title-based sentence-similarity method refined with venue curation described earlier.

Let's also discuss about annual percentage of female and male authorships within each assigned research category (see Figure 4.12). Reading across the panels, the pattern is broadly consistent: as publication volume grows, the female share increases gradually while the male share declines by the same amount. The rise is most pronounced in categories oriented to people, data, and applications. Social and Networked Computing and Graphics, Vision, and HCI move from the mid-teens in the early years into the low-to-mid twenties in recent years. AI and ML, Data Science and Databases, and Software Engineering and Programming Languages show steady gains that place them around one fifth by the end of the period. Core infrastructure and mathematically intensive areas increase more slowly: Computer Systems and Architecture, Security and Cryptography, Robotics and Autonomous Systems, and Theory of Computation and Algorithms tend to remain in the mid-teens even in the most recent years. Early bars in several panels fluctuate because the number of publications is small; once the series matures, trajectories smooth out and the upward trend becomes clearer. Percentages are computed on authorships rather than unique people, which is appropriate for assessing contribution to category output.

Across venues and categories the patterns are largely consistent, with differences of degree rather than direction. Venue results mirror the composition of their dominant subfields: robotics conferences (ICRA, IROS) sit at the lower end of female representation, just as the Robotics & Autonomous Systems category does; computer vision venues (CVPR) land in the middle, matching Graphics, Vision, and HCI; and broad outlets (CoRR, IEEE Access) track the overall average much like cross-cutting categories such as AI & ML or Software Engineering & Programming Languages. Categories smooth over venue-specific idiosyncrasies, so their female shares cluster in a tighter band (roughly 16–25 percent), whereas individual venues show a bit more spread (about 13–23 percent) due to editorial scope, community practices, and history. Both views show the same long-run movement: gradual gains in the female share over time, with persistent gaps between theory, systems, robotics, and security on one side and more application- and people-oriented areas on the other. In short, category-level distributions provide a stable summary of field differences, while venue-level distributions reveal local variation around those field norms.

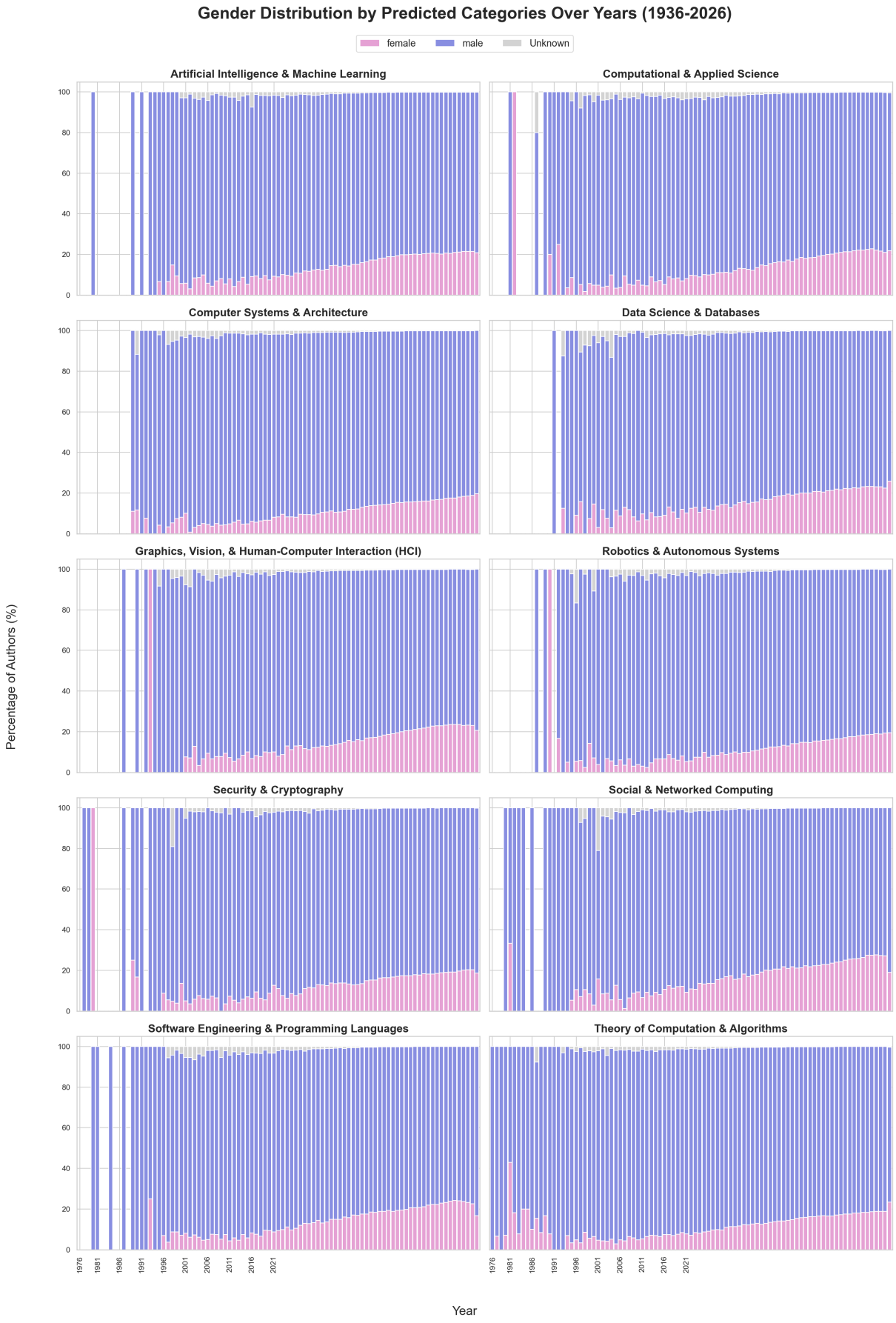


Figure 4.12: Annual gender composition by assigned research category in the dblp corpus.

# Chapter 5

# Conclusion

This thesis set out to measure gender representation in computer science publishing using the dblp bibliography as the sole source. I built a complete workflow that parses the dblp XML into per-authorship records, cleans and normalizes the data, assigns categories to titles with sentence embeddings refined by venue curation, and infers author gender with probability scores from GenderAPI. The final analytic corpus covers 7,993,261 publications, expands to 27,189,298 authorships, and reduces to 3,982,244 unique authors for inference. The analysis traces publication growth, compares venues and research areas, and summarizes country patterns as indicators of name origin.

Publication volume grows sharply from the 2000s onward, with articles and in-proceedings accounting for most of the expansion, while books, edited proceedings volumes, chapters, and theses remain comparatively small. Gender composition changes gradually through time. The female share rises from roughly one in ten authorships in the 1980s to about one in five in recent years, and the male share declines by a similar amount while remaining the majority. Venue results mirror their dominant subfields. Sensing and earth observation journals report higher female shares than robotics and parts of signal processing, while computer vision and machine learning sit between these extremes. Category results show the same structure. Social and networked computing and graphics, vision, and HCI have the highest female shares, whereas systems, theory, robotics, and security are lower. Country labels from the service describe the geographic distribution of names rather than nationality or affiliation. The maps show wide global coverage together with expected ambiguities where naming traditions overlap or transliteration varies.

Gender is inferred from names and is suitable for aggregate analysis only. Initial-only records, rare names, and transliteration can reduce confidence and are more common in older

metadata. Country labels from the service reflect name origin statistics, not residence or affiliation. dblp coverage is broad but not uniform across time and venue, and counts for the newest year can be incomplete at the time of extraction. These factors introduce uncertainty. I mitigate them by keeping probabilities, reporting sensitivity checks, documenting decision rules, and focusing on group-level patterns rather than person-level claims.

The results provide a coherent picture of where participation concentrates in computer science publishing and how it has changed. Venues and categories that engage human factors, data, and applications appear to have moved faster toward balance. Areas anchored in theory, systems, robotics, and security change more slowly. These patterns can inform venue policies, mentoring programs, and field-specific monitoring. The pipeline itself offers a reusable template for institutions that wish to track representation using public bibliographic data.

Several extensions are natural. First, link authors to affiliations or ORCID records to connect name origin to institutional geography. Second, enrich topic assignment with abstracts and full texts where available, and expand the taxonomy to finer subfields. Third, validate inference with larger gold-standard samples that include non-binary identities and multiple writing systems, and recalibrate thresholds as services and datasets evolve. Fourth, study authorship order and collaboration networks to relate participation to roles and influence. Finally, package the workflow as a public release with dashboards that support regular updates.

In closing, the thesis shows that a carefully engineered pipeline, modern text representations, and probability-aware inference can produce a clear and reproducible view of gender representation across time, venues, and research areas in the dblp universe.

# Bibliography


Raquel Berenguer. Enabling the analysis of computer science research via a relational database from dblp, 2024. URL https://hdl.handle.net/10609/150472.

Juan-José Boté-Vericad, Miquel Centelles, and Núria Ferran-Ferrer. Determining gender in academic authorship: a comprehensive and methodical approach. *Digital Library Perspectives*, 41(2):346–366, 03 2025. ISSN 2059-5816. doi: 10.1108/DLP-05-2024-0080. URL https://doi.org/10.1108/DLP-05-2024-0080.

Silvia Cobo-Serrano, Rosario Arquero-Avilés, and Gonzalo Marco-Cuenca. Journal of information science: A gender-based bibliometric study (2015–2020). *Journal of Information Science*, 50(1):116–128, 2024. doi: 10.1177/01655515221081346. URL https://doi.org/10.1177/01655515221081346.

B Elango and Dong-Geun Oh. Scientific productivity of leading countries. *International Journal of Information Science and Management (IJISM)*, 20(2):127–143, 2022. ISSN 2008-8302. URL https://ijism.isc.ac/article_698383.html.

Carlo Galli, Nikolaos Donos, and Elena Calciolari. Performance of 4 pre-trained sentence transformer models in the semantic query of a systematic review dataset on peri-implantitis. *Information (Basel)*, 15(2):68, January 2024.

https://www.genderapi.io. GenderAPI | Name Gender Checker – Determines Gender from Name with AI — genderapi.io. https://www.genderapi.io/.

Kriste Krstovski, Yao Lu, and Ye Xu. Inferring gender from name: a large scale performance evaluation study, 2023. URL https://arxiv.org/abs/2308.12381.

Alexander D. VanHelene, Ishaani Khatri, C. Beau Hilton, Sanjay Mishra, Ece D. Gamsiz Uzun, and Jeremy L. Warner. Inferring gender from first names: Comparing the accuracy of genderize, gender api, and the gender r package on authors of diverse nationality. *PLOS*

*Digital Health*, 3(10):1–15, 10 2024. doi: 10.1371/journal.pdig.0000456. URL https://doi.org/10.1371/journal.pdig.0000456.

# Appendix A

# Tables

| Category | Description | Subjects |
|---|---|---|
| Artificial Intelligence & Machine Learning | This field is dedicated to creating systems that can perform tasks that normally require human intelligence. At its core, it explores how computers can learn from data to make predictions, decisions, and classifications. This ranges from Machine Learning algorithms that identify patterns in massive datasets, to Natural Language Processing that allows machines to understand and generate human language, to Knowledge Representation which aims to formally model the world so a machine can reason about it. The ultimate goal is to build systems that can not only automate tasks but also perceive, reason, and adapt in complex environments, driving innovations from self-driving cars to medical diagnoses. | Artificial Intelligence, Machine Learning, Natural Language Processing (Computation and Language), Neural and Evolutionary Computing, Knowledge Representation, Pattern Recognition |
| Theory of Computation & Algorithms | This is the mathematical foundation of computer science. It tackles fundamental questions about what problems can be solved by computers (computability theory) and how much time and memory are required to solve them (complexity theory). Researchers in this area design and analyze algorithms—the step-by-step recipes for solving problems—to ensure they are correct and efficient. Using tools from logic and discrete mathematics, this field establishes the absolute limits of computation and provides the theoretical toolkit that all other areas of computer science rely on to build efficient and reliable software. | Algorithms and Data Structures, Computational Complexity, Formal Languages and Automata Theory, Logic in Computer Science, Discrete Mathematics, Game Theory |
| Computer Systems & Architecture | This category focuses on the design and operation of the physical and low-level software components of a computer. It spans from the design of microprocessors and memory hierarchies (Hardware Architecture) to the creation of the Operating System, the fundamental software that manages all hardware resources and allows other programs to run. It also includes the principles of Distributed and Parallel Computing, which explores how to connect many computers into a single, powerful system. This field is constantly pushing the boundaries of speed, efficiency, and scale, from the smallest embedded devices to massive supercomputers. | Hardware Architecture, Operating Systems, Distributed, Parallel, and Cluster Computing, Networking and Internet Architecture, Performance, Emerging Technologies like Quantum Computing |
| Software Engineering & Programming Languages | This field applies engineering principles to the complex process of building and maintaining software. It addresses the challenge of creating reliable, efficient, and scalable software through systematic processes for design, development, testing, and deployment. A critical component is the study of Programming Languages, which are the formal tools used to give instructions to a computer. Researchers in this area not only create new languages and paradigms (like object-oriented or functional programming) but also develop the methodologies and tools that allow teams of developers to collaborate on large-scale software projects that can be sustained for many years. | Software Engineering, Programming Languages, Software Development, Mathematical Software |
| Security & Cryptography | This area is focused on defending computer systems and data against attack, damage, or unauthorized access. Cryptography provides the mathematical tools for ensuring confidentiality (preventing eavesdropping), integrity (preventing tampering), and authenticity (verifying identity). Security as a whole encompasses a broader scope, including network security, system security (protecting against malware and intrusions), and creating dependable systems that are resilient to both malicious attacks and accidental failures. In an increasingly connected world, this field is essential for protecting everything from personal privacy to critical national infrastructure. | Cryptography and Security, Privacy, Dependable and Secure Computing |

| | | |
|---|---|---|
| Security & Cryptography | This area is focused on defending computer systems and data against attack, damage, or unauthorized access. Cryptography provides the mathematical tools for ensuring confidentiality (preventing eavesdropping), integrity (preventing tampering), and authenticity (verifying identity). Security as a whole encompasses a broader scope, including network security, system security (protecting against malware and intrusions), and creating dependable systems that are resilient to both malicious attacks and accidental failures. In an increasingly connected world, this field is essential for protecting everything from personal privacy to critical national infrastructure. | Cryptography and Security, Privacy, Dependable and Secure Computing |
| Data Science & Databases | This category revolves around the entire lifecycle of data: how to store, manage, retrieve, and ultimately extract valuable insights from it. Databases provide the core technology for efficiently storing and querying vast amounts of structured information. Building on that, Data Science and Data Mining use statistical methods and machine learning algorithms to discover hidden patterns, trends, and knowledge from both structured and unstructured data. This field powers everything from modern e-commerce recommendation engines and business intelligence dashboards to scientific discoveries driven by large-scale data analysis. | Databases, Data Mining, Information Retrieval, Big Data, Digital Libraries |
| Graphics, Vision, & Human-Computer Interaction (HCI) | This field governs the crucial interface between humans and computers. It is composed of three closely related parts: Computer Graphics (how computers generate and manipulate images, from rendering photorealistic scenes in movies and video games to creating scientific visualizations), Computer Vision (the inverse challenge of how computers can understand and interpret the content of images and videos, enabling tasks like facial recognition and autonomous navigation), and Human-Computer Interaction (HCI) (the study and design of how people interact with technology, focusing on creating interfaces that are intuitive, effective, and enjoyable to use). | Computer Graphics, Computer Vision, Human-Computer Interaction, Multimedia, Sound, Affective Computing |
| Robotics & Autonomous Systems | This field brings computation into the physical world by creating machines that can sense, act upon, and interact with their environment. Robotics is a deeply interdisciplinary field that combines computer science (for perception, planning, and control) with mechanical and electrical engineering (for building the physical body, sensors, and actuators). It also includes the study of Multiagent Systems, where teams of robots or software agents coordinate to achieve a common goal. The challenge lies in building systems that can operate reliably and intelligently in the unpredictable physical world, from manufacturing assembly lines to planetary exploration rovers. | Robotics, Multiagent Systems, Systems and Control, Sensors and Signal Processing |
| Computational & Applied Science | This area uses high-performance computing to solve complex problems in other scientific and engineering disciplines, often called the "third pillar" of science alongside theory and experimentation. By creating sophisticated mathematical models and simulations, researchers can study phenomena that are too large, too small, too fast, or too dangerous to investigate in a laboratory. This includes simulating galaxy collisions in astrophysics, folding proteins in bioinformatics to discover new drugs, modeling climate change, and designing new materials or financial instruments. | Computational Engineering, Finance, and Science; Computational Biology and Bioinformatics; Scientific Computing and Simulation; Symbolic Computation |
| Social & Networked Computing | This field studies the intersection of computing technology and human social behavior. It analyzes how social and information networks, like the internet and social media platforms, shape communication, influence, and the spread of ideas. Researchers in this area investigate the structure of online communities, the dynamics of collective behavior, and the ethical implications of algorithms that govern our digital lives. It combines techniques from computer science, network science, and sociology to understand and build the digital society we increasingly inhabit. | Computers and Society, Social and Information Networks, World Wide Web and Web Science, Mobile and Ubiquitous Computing |

Figure A.1: Table of computer science categories.