Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Beyond Citations: Explainable Machine Learning for Identifying Early-Career Researchers from Bibliometric Data

Avijit Gayen
Techno India University
Somyajit Chakraborty
University College Cork
Debashis Chakraborty
Techno India University
Angshuman Jana
angshuman@iiitg.ac.in

Indian Institute of Information Technology Guwahati

Research Article

Keywords: Early Career Researchers, Machine Learning, Bibliometric Analysis, Collaboration Networks, Supervised Classification

Posted Date: November 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11309801/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Beyond Citations: Explainable Machine Learning for Identifying Early-Career Researchers from Bibliometric Data

Avijit Gayen[1,2], Somyajit Chakraborty[3], Debashis Chakraborty[3], Angshuman Jana[1*]

[1*]Dept. of CSE, Indian Institute of Information Technology Guwahati, Bongora, Guwahati, 781015, Assam, India.
[2]Department of CSE, Techno India University West Bengal, Salt Lake, Kolkata, 700091, West Bengal, India.
[3]Dept. of Computer Science, University College Cork, Cork, Ireland.
[3]Department of CSE, Maulana Abul Kalam Azad University of Technology, Nadia, West Bengal, India.

*Corresponding author(s). E-mail(s): angshuman@iiitg.ac.in;
Contributing authors: avijit.gayen@iiitg.ac.in, avijit.g@technoindiaeducation.com; SomyajitChakraborty@ucc.ie; debashischakrabarti969@gmail.com;

Abstract

Identifying ECRs (early career stage researchers) is crucial for developing academic collaborations and supporting emerging scholars, but in the context of large bibliographic datasets it is a challenging task. In this paper, a machine learning framework for automated identification of ECRs is proposed, with a particular focus on distinguishing between early career stage researchers and more established researchers. Citation based and non-citation based features are leveraged in the creation of a training set, which consists of publication and collaboration related features extracted from the AMiner dataset with additional annotations to represent ECR researchers. We evaluate different feature selection and extraction techniques as well as several supervised classification algorithms including Random Forest, Logistic Regression, Support Vector Machines, XGBoost and Gradient Boosting, finding that the best performing models (XGBoost and Gradient Boosting) achieve 99% accuracy in ECR classification. The application of explainable AI methodology provides further insights into the contribution of

different features to the classification task, revealing that the most informative non-citation features are related to trajectory based analysis. Moreover, the analysis shows that non-citation features have greater discriminatory power for ECR identification compared to citation features, illustrated in the OpenAlex case study. The proposed approach enables more accurate recommendation of potential collaborators as well as more informed policy design related to the needs of the ECR community.

Keywords: Early Career Researchers, Machine Learning, Bibliometric Analysis, Collaboration Networks, Supervised Classification

1 Introduction

Scientific investigation and advancement play a crucial part in raising a nation's growth, ultimately contributing to its economic growth [1, 2]. In the contemporary research landscape, research activities are expanding rapidly not just in technology and science but in many other fields as well. Such collaborations facilitate the resolution of complex problems through the collective expertise of researchers. They improve research quality [3], generate substantial knowledge, and foster innovation and productivity [4, 5]. As a result of this growth in research and development, bibliographic metadata has increased exponentially. Within this expanding research ecosystem, Early Career Researchers (ECRs) represent an important population whose development and participation contribute to the continuity and advancement of scientific research. Accurate identification of ECRs is important for research evaluation, academic career studies, talent development, and the design of effective research support and science policy interventions. However, identifying ECRs reliably at scale is challenging because researchers at different stages of their careers may exhibit overlapping bibliometric characteristics, while large bibliographic datasets contain substantial variation in publication and citation patterns.

Although the involvement of ECRs in cooperative study is indispensable, identifying an effective collaboration to respond to a possible study problem remains a significant challenge [6]. This difficulty is further compounded by the ever-expanding bibliographic information. Although this problem is prevalent among scholars, of all career stages, ECRs may find it particularly difficult to select an ideal academic collaborator because of their restricted domain expertise and inability to contact domain specialists. As a result, they may either form ineffective collaborations [7, 8] or miss out on valuable opportunities, leading to a scarcity of meaningful collaborations in the early stages of their research careers. In addition to this main obstacle, Additionally, ECRs face a number of significant challenges that impede their ability to advance professionally. Due to insufficient attention toward ECRs, they often confront fierce competition from reputable researchers or research teams for securing project financial support [9]. Moreover, their contributions to identified research problems may be overshadowed by recognized scholars, limiting the recognition of their work [10].

The literature lacks a clear and standardized definition of Early Career Researchers (ECRs), which can lead to difficulties in accurately identifying them within large bibliographic datasets. This ambiguity can lead to inaccurate classification of researchers, affecting downstream analytical and policy-related decisions. The literature lacks a clear and standardized definition of Early Career Researchers (ECRs), which can lead to difficulties in accurately identifying them within large bibliographic datasets. This ambiguity can lead to inaccurate classification of researchers, affecting downstream analytical and policy-related decisions. Because the literature does not provide a universally accepted temporal boundary for identifying Early Career Researchers (ECRs), this study uses a publication-based temporal window as an operational criterion for identifying early-career research trajectories. The seven year window used in our primary analysis is therefore a bibliometric reference point to be reasserted as a reproducible finding rather than a defining boundary for academic seniority or research maturity. In a more general sense, the initial period of approximate a decade represents an important developmental phase in which researchers establish publication patterns, collaborations, research identity, and scholarly visibility. To evaluate the sensitivity of this temporal operationalization, alternative five year and ten year windows are also considered.

It is important to distinguish career-stage identification from research seniority. Researchers may acquire research autonomy at different rates, and elapsed publication time alone cannot directly capture independent funding, project leadership, supervision, or other institutional responsibilities. Therefore, the ECR classification in this study should be interpreted as a bibliometrically observable career stage classification rather than a comprehensive measure of research maturity or academic seniority. Previous studies have shown that such improper collaborations tend to be less stable over time [11]. Additionally, the absence of a precise ECR definition can cause misallocation of research funding, negatively impacting the overall growth of research and development. Identifying ECRs using existing methodologies presents two significant challenges: (a) the lack of a well defined criterion that distinctly separates ECRs from other researchers, and (b) the complexity of identifying ECRs in large-scale bibliographic data. For example, certain authors may have a comparatively low h-index, yet they may not be in the early stage of their careers. This discrepancy could arise due to factors such as working in domains with fewer citations, conducting research in small, close knit groups, or working independently circumstances that naturally result in fewer citations.

Thus, there is a strong need to develop an automated system capable of accurately identifying ECRs from vast bibliographic datasets. Such a system would not only facilitate precise identification of ECRs across diverse research domains but also improve large-scale researcher identification and analysis frameworks, ensuring that ECRs are matched with 'beneficial' and 'sustainable' research partnerships that can significantly advance their careers. In this study, we draw insights from the work of Gayen et al. [12], which analyzes publication patterns and developmental trajectories of early-career researchers, providing useful context for understanding their evolution in bibliographic datasets.Building on these observations, we propose a machine learning-based classification model to identify ECRs from large bibliographic datasets. Since scientific

performance is often evaluated through citation based metrics, and ECRs typically accumulate fewer citations in the starting phases of their careers, to increase the precision of ECR identification, our method combines citation based and non-citation based information. This evaluation raises two fundamental research questions—

- **RQ1**: *What are the contributions of cited features in the identification of ECR?*
- **RQ2**: *Are the non-citation based features alone able to separate the extensive academic dataset from the ECR?*

In this study, we propose a machine learning-based framework for identifying early career researchers using large scale bibliographic data. The approach integrates both citation based and non-citation based features to improve classification accuracy. Detailed methodology, including dataset processing, feature engineering, and model evaluation, is presented in Section 3.

The methodological contribution is not the use of any individual machine learning classifier in isolation. Rather, the contribution lies in formulating ECR identification as a scalable classification problem and constructing a heterogeneous predictor space that combines citation based impact indicators with non-citation based productivity, temporal, collaboration, and contextual information. The systematic comparison of complementary classifier families, feature-selection strategies, feature-family ablation, and explainability analysis enables us to determine which dimensions of scholarly trajectories actually contribute to ECR identification.

Our Contributions: The following is a summary of this paper's merits:

1. We examine the details of the Aminer dataset that reveals the academic demographics and bibliometric analysis. This analysis highlights the background of the dataset used in our proposed model development and underscores the role of various bibliometric data that help us with the feature selection for our proposed model.
2. We perform the feature modeling and further feature selection as well as the feature extraction method before the deployment of the proposed ML model.
3. We employ various ML models to identify the ECR from the bibliographic dataset. The performance of the proposed model is validated by the LOOCV method.
4. To answer the interpretability of the selected features in our proposed model, we adopted the LIME and SHAP models of explainable AI.

The remainder of this paper is organized as follows: Section 2 reviews related work on early-career researchers and collaboration analysis. Section 3 describes the proposed pipeline, including the datasets, feature modeling, annotation strategy, and the supervised learning and explainability methods used for ECR identification. Section 4 reports experimental results, including annotation validation and model performance (accuracy, precision, recall, F1-score, and AUC–ROC) under three settings: (a) all features, (b) reduced features via feature selection, and (c) reduced features via feature extraction; it also summarizes the XAI findings used to address the research questions. Section 5 presents a case study on the OpenAlex CS Dataset. Finally, Section 6 concludes the paper and discusses limitations and future work. Throughout, we use

the terms "researchers" and "authors" interchangeably; likewise, although "collaboration" can be broader, in this study it refers specifically to co-authorship relationships observed in bibliographic records.

2 Related Works

We list the contributing works pertaining to the ECR in this section. First, we discuss recent trends in collaboration advancements and the contributions made to scientific collaborative research activity. We also describe how contemporary AI technology have contributed to recent advancements in scientific collaboration. We also describe the ECR-related works and their function in research. We also draw attention to the issues and difficulties facing ECRs in the field of research. We also emphasize the difficulties in identifying ECR. Lastly, we highlight the work's scope while pointing out the literature's research gap.

2.0.1 Early Career Stage Researchers (ECRs)

Definition and Conceptualization: Early Career Researchers (ECRs) are generally understood as individuals in the initial phase of their academic careers, typically including doctoral candidates, postdoctoral researchers, or scholars within the first few years after obtaining their Ph.D. [13–15]. However, the literature highlights that ECR status cannot be defined solely based on time since degree completion. Instead, it is shaped by institutional roles, career transitions and disciplinary norms. For example,[16] emphasizes that ECR identity is dynamic and evolves through mobility, institutional expectations, and changing academic responsibilities.

Structural Struggles and Career Instability:

ECRs operate within highly competitive academic environments characterized by employment precarity and uncertain career progress. Prior studies indicate that early stage researchers frequently face short term contracts, limited access to tenure-track positions, and significant career uncertainty [17, 18]. This "risk career" condition often leads to both occupational and intellectual insecurity, discouraging long term academic commitment.

Productivity Pressure and Systemic Inequality:

The "publish or perish" culture imposes substantial pressure on ECRs to demonstrate high research productivity despite limited resources [19, 20]. Moreover, the academic system is influenced by cumulative advantage mechanisms, where established researchers benefit disproportionately from visibility, funding, and recognition, thereby placing ECRs at a structural disadvantage [21]. As a result, ECRs often struggle to gain recognition for their contributions.

Funding and Publication Challenges:

Limited access to research funding is a critical barrier for ECRs, as funding opportunities are often biased toward senior researchers with established track records [17]. This constraint restricts their ability to initiate independent research and build collaborations. Additionally, ECRs face challenges in the publication process, including difficulties in publishing in high impact venues and issues related to authorship recognition and peer review practices [22].

Workload, Identity, and Well being:

ECRs are required to balance multiple responsibilities, including research, teaching, and administrative duties, which often leads to time constraints and reduced research focus [13]. At the same time, they must develop a stable academic identity while navigating institutional expectations and limited mentorship [23]. These pressures, combined with job insecurity, contribute to increased levels of stress, anxiety, and burnout among ECRs [24, 25].

Summary and Research Gap:

Overall, the literature demonstrates that ECRs are defined not only by their early career stage but also by the structural constraints and systemic challenges they encounter. Despite their critical role in sustaining scientific innovation, there remains a lack of scalable and data driven approaches to systematically identify and support ECRs within large bibliographic datasets. This gap motivates the present study, which proposes a machine learning based framework for ECR identification.

2.0.2 Bibliometric Predictors for Researcher Characterization

Recent scientometric research shows that researchers can be captured by heterogeneous bibliometric information beyond just citation based indicators. Momeni et al. [26] distinguish between prior impact based features like publications, citations, and h-index, and non-prior impact based features associated with authors, co-authorship, publications, and venues. Their analysis in different career phases shows that non-prior impact based features can yield better predictive information for younger researchers, particularly for longer prediction horizons. These results motivate the combined use of impact related and non-impact related indicators with young career researchers.

Similarly, Zhu et al. [27] show the value of supervised machine learning and feature selection methods for finding informative bibliometric characteristics, but their goal is measuring academic influence not career stage. Frandsen and Nicolaisen [28] further show that ECR identification can depend substantially on the choice of publication-based operational definition, highlighting the importance of carefully selecting measurable indicators when operationalizing ECR status.

These studies provide theoretical and methodological motivation for our predictor design. However, existing work primarily focuses on researcher impact, publication-based ECR definitions, or future citation performance rather than the direct classification of ECRs from large bibliographic datasets. Therefore, we extend this literature by organizing heterogeneous bibliometric predictors according to the dimensions of scholarly impact, productivity, research trajectory, and academic context, and by systematically evaluating their contribution to ECR identification.

2.0.3 Difficulties and Issues with ECR

It is evident that ECR exists in every subject of research, and it is an indisputable fact that ECR must be included in the solution of important issues. However, in their field of interest as well as in cross domain research, ECR encounters a number of difficulties and problems. Previous research [29, 30] examined early stage research employment frequently involve a significant amount of teaching. As a result, ECR struggles with time management, research focus, and other challenges. The authors of

another work [31] emphasize that the identity that the ECR helps them establish is crucial to their professional growth. Research project funding is crucial to the development of scientific endeavors. Horta et al. noted in a significant study [32] that ECR's performance is influenced by financing and recognition. For the transdisciplinary domains, such as ocean and coastal sustainability, the ECR's walls folded several times. The authors of a new study [33] highlighted the main obstacles to ECR in the long term stability of the ocean and the land as well as strategies for overcoming them.

The "risk career" that ECRs have may seldom locate a dependable, secure academic career because of the existing higher education policy, according to authors in some earlier works [34]. According to a significant study [35], post-PhD researchers' experiences are impacted by career uncertainty. According to the study's findings, post-PhD researchers encounter two types of professional doubt during their job experiences: occupational and intellectual uncertainty. Potentially talented researchers are drawn to other employment options by the uncertainties surrounding research careers. In their analysis of the review procedures used by traditional peer reviewed journals, Mcdowell et al. found that the ECR has a "ghost" review writing process that has been carried out without acknowledgment [36]. This situation highlights both the misuse of ECR and the compromise of the review process, which lowers the caliber of research. Drosou et al. also point out in another work [37] that ECR always has trouble getting their study outputs published in reputable journals.

This growing trend of cooperative study has compelled the researcher community to look for possible collaborations. Even if searching through this constantly updated bibliographic collection is arduous, the ECR's predicament is made more difficult by the dearth of information about domain specialists and the emergence of sophisticated technique collaborations for career growth. The majority of effort to date has been on suggesting the greatest number of collaborations without considering the sustainable nature of cooperation or identifying beneficial collaborators (BCs), even though the emergence of numerous AI driven collaboration recommendation systems has improved the situation. In [38], Kong et al. noted that BCs are able to offer academic guidance in addition to working with researchers. They could make quick progress with the assistance of BCs. According to other studies, a long-term partnership significantly boosts a researcher's output and influence [11]. There are not many works that concentrate on ECRs among the many that deal with collaborative advice.

According to the literature now in publication, there is a significant need to concentrate on "early-stage" researchers in order to advance research. Surprisingly, none of the works made an effort to supply a significant solution to the difficulties pertaining to the ECR as their primary focus, even though the literature concurs with the current concerns of the ECR. It is difficult to find ECRs using the previous definition in the extensive bibliographic dataset. Therefore, a formal methodology to identify ECR is highly necessary in order to fix the problems surrounding them. In this study, we provide a machine learning based methodology using the electronically maintained citation record to recognize the scholars at their "early stage".

3 Proposed Method

In this part, we first give a thorough explanation of the dataset that was used for this study. We also go over how the dataset was cleaned using data prepossessing and how the dataset was appropriately annotated for supervised learning. We go on to discuss feature modeling, which covers feature selection and feature extraction in general. We go into detail about the several citation based and non-citation based variables that the model uses to classify writers into ECR and non-ECR types. Below is a description of the many machine learning classifiers we used in our investigation to learn the bibliographic data set. To determine the influence of the features in our suggested model, we use an explainable AI model. In figure 1, we also include a schematic diagram that illustrates the procedure to carry out the approach.

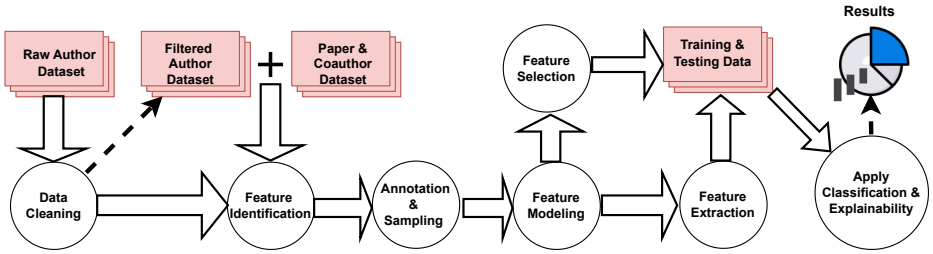


Fig. 1: An illustration of the flowchart's schematic used in the suggested study is displayed in the figure.

Our methodology for Early Career Researcher (ECR) identification follows a systematic pipeline combining data collection, preprocessing, feature engineering, and machine learning classification. The process starts with acquiring a raw author dataset containing researchers' metadata, including publications, affiliations, and collaborations. After data cleaning to eliminate inconsistencies and duplicates, a refined paper and co-author dataset is generated. Then the data is preprocessed, split into training and testing sets, and normalized for machine learning applications. Feature engineering identifies key attributes such as publication trends, collaboration networks, and citation metrics, capturing an author's academic trajectory. Labeled data is obtained through expert annotation, and sampling techniques ensure balanced class distribution. Machine learning models, including Random Forest, SVM, and Gradient Boosting, are trained and optimized for classification. Explainability techniques like SHAP and LIME provide model interpretability, while performance is evaluated using F1-score, AUC-ROC, recall, accuracy, and precision. This structured framework offers a reliable, data-driven approach for ECR identification by leveraging socio-technical network analysis and machine learning.

3.1 Methodology Overview

To develop an ECR identification model using bibliographic data, we utilize the AMiner dataset [39]. The raw data is first preprocessed through cleaning and handling of missing values. Based on prior studies [12], we construct a feature set incorporating both citation based and non-citation based attributes that capture collaboration patterns and publication trends.

To enable supervised learning, the dataset is annotated using structured strategies, and annotation reliability is validated using Cohen's Kappa measure. Since the dataset is inherently imbalanced due to the dominance of low-profile authors, a sampling strategy is applied to construct a balanced dataset of 10,000 instances. The dataset is then divided into training and testing sets in a 7:3 ratio.

We initially identify six citation based and six non-citation based features. Feature selection and feature extraction techniques are applied to reduce dimensionality and improve computational efficiency. Multiple supervised machine learning classifiers, including XGBoost, Gradient Boosting, SVM, Decision Trees, Naive Bayes, and Logistic Regression, are used for model training.

The model is evaluated under three scenarios: (a) using all features, (b) using selected features, and (c) using extracted features. Finally, explainability techniques such as LIME and SHAP are employed to analyze the contribution of both citation based and non-citation based features in ECR identification.

3.2 Operational Definition of Drop-out Researchers

Within the bibliographic dataset, drop out researchers are identified based on sustained publication discontinuity after an initial active period. Consistent with prior Scientometrics research on publication discontinuation patterns, this classification is based on the absence of subsequent indexed publications within the observation window.

This designation reflects dataset observed inactivity and should not be interpreted as confirmed permanent exit from academia or research careers.

3.3 The dataset

We employ the AMiner dataset for our evaluation. [39]. K. Scott Mader maintains this free and open source dataset, which was last updated on December 12, 2018, and is licensed under the Creative Commons license (CC BY-NC-SA 4.0). The dataset includes bibliographic data for publications published between 1936 and 2014. We used the bibliographic information from the paper publication as well as specific information about the authors and their relationships with coauthors in our research. Three distinct files containing this data are arranged according to the characteristics listed in table 1.

In table 2 and 3, we represent the spinet of the Aminer paper and author dataset. In the paper dataset does not contain the author index, thus there are some disambiguation problems related to the authors' names in the dataset. We carefully removed those ambiguous records from our observation. In table 4, we represent the general statistics of the raw and filtered data of the Aminer dataset.

Paper		Author	
Attributes	Attribute Definition	Attributes	Attribute Definition
index	paper ID	index	author ID
*	paper title	n	author name
@	authors	a	affiliations
o	affiliations	pc	publication count
t	published year	cn	number of citations
c	publication venue	hi	h-index
%	the ID of references	pi	p-index(equal A-index)
!	abstract	upi	p-index(unequal A-index)
-	-	t	area of interest

Coauthor	
Attributes	Attribute Definition
index	Author ID
index	Author ID
count	frequency of co-authorship

Table 1: Definitions of dataset attributes

Table 2: Basic format of AMiner Paper dataset [39]

```
#index13
#*The VLSI Complexity of Selected Graph Problems.
#@Joseph Já Já
#oThe Pennsylvania State Univ., University Park, PA
#t1984
#cJournal of the ACM (JACM)
#%1068477
#!NaN
```

Table 3: Basic format of AMiner Author dataset [39]

```
#index 1488277
#n Juanzi Li
#a Tsinghua University;
Department of Computer Science & Technology,
Tsinghua, University, Beijing, China 100084
#pc 70
#cn 370
#hi 9
#pi 76.3254
#upi 73.7573
#t semantic web;social network;Semantic Annotation; ontology caching;
semantic information;knowledge base
```

3.3.1 Data cleaning

For our investigation, the AMiner dataset [39] is being used. K. Scott Mader maintains the dataset, which is open-source and free. It is distributed under a Creative Commons license (CC BY-NC-SA 4.0) and was last modified on December 12, 2018. Citations data for publications published between 1936 and 2014 is included in the dataset. We

Table 4: The Arnet Miner dataset (publication tenure 1936—2014) includes both raw and filtered general data.

Arnet Miner dataset	Raw	Filtered
Count valid papers	20,92,356	20,54,704
Amount of papers without author details	37505	-
Number of papers without publication year	147	-
Number of researchers	17,12,433	17,12,431
Number of Authors who lack proper details	2	-
Avg. amount of papers per author	-	2.887
Avg. number of authors per paper	-	3.262
Total number of unique coauthorship relation	-	42,58,946

employed comprehensive information on authors and their relationships with coauthors, as well as bibliographic data from the original publication, in our research. Three distinct files, arranged based on the characteristics indicated in the table 1, contain this information. Since most of them have either very few or no follow up papers, we classify them as *'discontinued'* researchers. We have excluded *'discontinued'* researchers from our study because they have the least impact on the co-authorship network. Following preprocessing, we present the dataset's general information in table 4.

3.3.2 Annotation

Since the AMiner dataset does not provide an explicit ECR/non-ECR label, a manual expert based annotation procedure was used to construct the reference labels required for supervised learning. The annotation was performed on selected author level bibliographic records rather than generating the labels solely through an automated seven-year rule. The available bibliographic information for each selected researcher was examined, and each researcher was assigned to one of two classes: ECR or non-ECR.

The temporal career stage criterion adopted in this study was used as an operational reference for identifying early career trajectories. It was not treated as a direct measure of academic seniority or research maturity. In particular, publication age alone cannot capture research autonomy, independent funding, project leadership, supervision, or other institutional responsibilities. These aspects are therefore recognized as limitations of a bibliometric only operationalization.

Because annotation of the complete AMiner population would be time consuming and impractical, approximately 10,000 author records were selected for annotation and model development. The original dataset exhibited substantial class imbalance, with approximately 90% of the researchers belonging to the ECR category at the time of annotation. Directly using this distribution for supervised classification could cause the classifier to favor the majority class. Therefore, a statistical sampling procedure was applied to obtain a more balanced representation of ECR and non-ECR researchers for model development.

The resulting annotated dataset contained 10,000 author instances, with 57.3% labeled as ECR and the remaining 42.7% labeled as non-ECR. The annotated dataset was subsequently divided into training and testing subsets using a 70:30 ratio. The sampling procedure was used specifically for supervised model development; the

Table 5: Expert annotation procedure used for ECR label construction.

Stage	Procedure
Input data	Author level bibliographic records obtained from the AMiner dataset, including publication history, citations, affiliations, and co-authorship information.
Career stage assessment	Researchers were assessed using the operational career stage criterion adopted in the study, with expert judgment used for annotation rather than treating the temporal criterion as a direct measure of academic seniority.
Label assignment	Each selected author record was assigned to one of two classes: ECR or non-ECR.
Expert agreement	The consistency of the expert annotations was assessed using Cohen's Kappa. The obtained value was $\kappa = 0.92$.
Class imbalance handling	Because the original dataset exhibited substantial class imbalance, statistical sampling was applied to obtain a more suitable class distribution for supervised model development.
Model development dataset	Approximately 10,000 annotated records were retained for model development and divided into training and testing subsets using a 70:30 ratio.

broader AMiner dataset was retained for the bibliographic analyses conducted outside model training.

The reliability of the annotation was assessed using Cohen's Kappa, which measures agreement between annotators while accounting for agreement occurring by chance. The obtained value of $\kappa = 0.92$ indicates a high level of agreement and supports the consistency of the annotation procedure.

Table 5 summarizes the expert driven annotation workflow used to construct the ECR/non-ECR labels.

3.3.3 Academic Demographics and Bibliometric Analysis

This section examines the three main statistics of our dataset: a) the distribution of researchers by country and institute; b) the distribution of publications by country and institute; and c) the type of publishing location (e.g., workshop, journal, conference, etc.). These statistics show the statistics of the scholars whose citation metrics were used throughout the study in this paper.

The descriptive analysis of the dataset reveals that a small number of countries (e.g., USA, China, and European nations) contribute a significant proportion of global research output, which is consistent with prior literature. Detailed distributions of authors and publications across countries and institutions are provided in Appendix A.

Figure 2 sorts publications according to their category, giving a summary of the most popular ways to share research in our collection. The most important finding is that conference proceedings outnumber journal articles as the most common publication type, with workshops, books, and other forms accounting for comparatively lesser amounts. In fields where conferences are essential for exchanging early research results and encouraging collaborations, this tendency is especially noticeable. Drawing from this discovery, we further propose that the dataset's reliance on conference papers might be a reflection of the kind of academic disciplines that emphasize rapid idea transmission. It might also point to difficulties in obtaining opportunities to publish

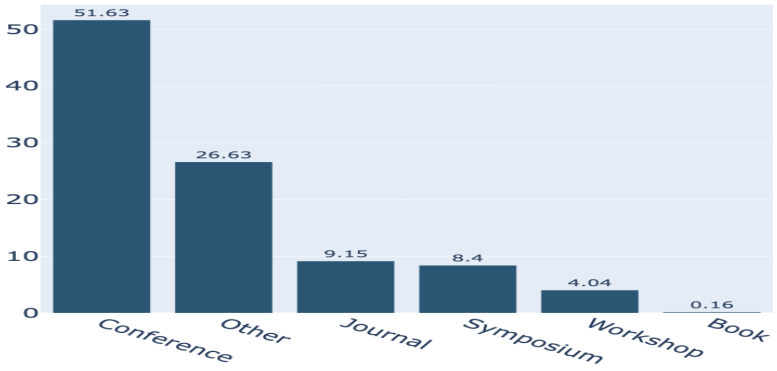


Fig. 2: The figure is a bar chart that displays the various publication types, including workshop," "conference," "journal," "symposium," "book," and "others." In our dataset, we have labeled publications as "others" in the bar chart if we cannot find any of the aforementioned types. The graph plots the publication categories along the x-axis, and the proportion of each category publication is displayed along the y-axis.

in journals, which sometimes call for additional time, money, and mentoring especially for researchers in their early careers.

3.4 Predictor Selection and Theoretical Motivation

The predictor set was designed to capture complementary dimensions of a researcher's scholarly trajectory rather than relying on a single bibliometric indicator. The selection was motivated by the conceptualization of ECR status as a career stage characteristic and by prior scientometric research showing that researcher impact and development can be characterized using both impact-based and non-impact-based information. [26, 28] Accordingly, the predictors used in this study are organized into two broad families: citation based indicators and non-citation based trajectory indicators.

Citation based predictors capture accumulated scholarly visibility and recognition. The citation count, h-index, p-index, up-index, citation growth rate, and highly cited publication ratio represent complementary aspects of citation based impact. These indicators are useful because established researchers generally have a longer period over which citations and impact metrics can accumulate. However, citation based measures alone may disadvantage younger researchers because they have had less time to accumulate citations. Momeni et al. [26] similarly distinguish prior impact based features from other researcher characteristics and show that non-impact based information can become particularly informative for younger researchers. Therefore, citation based indicators are included as one component of the predictor space rather than being treated as sufficient evidence of career stage.

Non-citation based predictors capture observable aspects of research activity and career trajectory that are not directly determined by accumulated citation counts. Publication Count represents research productivity, while Research Publication Tenure

and Publication Gaps describe the duration and continuity of scholarly activity. Collaborator Count captures the breadth of co-authorship activity, whereas Institutional Affiliations and Research Areas of Interest provide information about the researcher's academic context. These variables therefore provide complementary information about researcher development that may not be reflected adequately by accumulated citation measures.

The use of heterogeneous predictor families is therefore theoretically motivated rather than arbitrary. citation based indicators capture scholarly visibility and accumulated impact, whereas non-citation based indicators capture productivity, temporal trajectory, collaboration activity, and academic context. Their joint evaluation allows us to test whether ECR identification can benefit from information beyond conventional citation based measures.

The predictor families and their theoretical motivation are summarized in Table 6.

3.5 Feature Modeling

The different characteristics used for the classification are explained in this part, along with the pre-processing technique that was employed in the features prior to the application of the classifier for machine learning. Our model has both features that rely on citation and those that do not. The features used are listed, and each feature is briefly described in table 7. The dataset's citation based features do not need to be preprocessed before being fed into the classifier, but all noncitation based variables, aside from publication count, do. In our study, feature extraction and selection were crucial steps in developing a model capable of accurately distinguishing between ECR and non-ECR categories. These processes allowed us to emphasize the most informative features while minimizing noise and computational complexity.

3.5.1 Mathematical Framework

In this section, we represent mathematically each feature used in our work. It includes both 6 different citation based in addition to 6 distinct non-citation based characteristics along with their abbreviation as referred in table 7. In the subsequent section, we talk about those features in details and represent them in mathematical formulation.

Citation Count(C):

Citation count calculates the frequency with which other researchers have referenced an author's work in their papers. To calculate it, we consider all citations received from each of his or her publications cumulatively during his research tenure. This is a widely used, straightforward metric to assess how much a researcher's work has affected and influenced their field.

$$C = \sum_{i=1}^{P} C_i \tag{1}$$

where:

- C_i is the quantity of citations that the ith publication has received.
- P is the amount of publications overall.

Table 6: Theoretical motivation for the predictor families used for ECR identification.

Family	Predictor	Dimension captured	Relevance to ECR identification
citation based	Citation Count	Scholarly visibility	Captures accumulated recognition of a researcher's publications; useful for distinguishing accumulated research impact, but potentially influenced by career duration.
citation based	h-index	Productivity and citation impact	Combines publication productivity and citation accumulation; provides an established measure of scholarly impact, but may disadvantage researchers with shorter publication histories.
citation based	p-index	citation based popularity	Captures citation related prominence within the research community and complements conventional citation counts.
citation based	up-index	Contribution-adjusted impact	Extends citation based popularity by accounting for unequal authorship contribution.
citation based	Citation Growth Rate	Recent citation momentum	Captures the rate at which scholarly recognition develops rather than only accumulated citation volume.
citation based	Top 10% Highly Cited Paper Ratio	Exceptional visibility	Captures the proportion of highly cited publications and complements aggregate citation measures.
Non-citation based	Publication Count	Research productivity	Captures the amount of scholarly output and provides information complementary to citation accumulation.
Non-citation based	Institutional Affiliations	Academic context	Captures institutional context and affiliation patterns that may provide information about a researcher's academic trajectory.
Non-citation based	Research Areas of Interest	Disciplinary context	Accounts for the research domains in which an author is active and helps contextualize differences in publication and citation practices.
Non-citation based	Collaborator Count	Co-authorship activity	Captures the breadth of an author's co-authorship activity and provides information not directly represented by citation metrics.
Non-citation based	Publication Gaps	Research continuity	Captures discontinuities in publication activity that cannot be represented by total publication count alone.
Non-citation based	Research Publication Tenure	Career trajectory	Captures the duration of observable publication activity and therefore provides direct temporal information about research experience.

h-index (hi):

The h-index is a widely used measure to estimate the impact and productivity of a researcher's publications. It is defined as the maximum value of h such that the researcher has at least h papers, each of which has been cited at least h times.

$$hi = \max\{h \mid \text{at least } h \text{ papers have } \geq h \text{ citations}\} \tag{2}$$

p-index (pi):

The p-index is computed under the assumption of equal authorship contribution and measures the popularity of a researcher within their community.

Table 7: Features(with their abbreviation) used in Machine Learning Classifiers.

Features (citation based)	**Description**
Citation Count(C)	It counts how many times other scholars have referenced an author's work in their publications. This is a widely used and straightforward indicator to assess how much a researcher's work has affected and influenced their field.
h-index(hi)	A well-liked citation based indicator for evaluating the output and impact of a researcher's publications is the h-index. Both the number of publications and the number of citations the articles have got are taken into account.
p-index (pi)	The p-index calculates a researcher's level of popularity in their field of study. Assuming that each author's contribution to that publication is equal, When the A-index (authorship index) is equal, the value in the 'pi' field (p-index with equal A-index) is displayed.
up-index (upi)	When the A-index is unequal—that is, when the author's contributions to their publications vary—the p-index value is represented by the 'upi' (p-index with unequal A-index).
Citation Growth Rate(G_c)	This indicates how many citations an author has received on average annually since their initial publication, indicating the impact and relevance of their work over time.
Top 10% Highly Cited Paper Ratio(R_{10})	Proportion of an author's publications that fall into the top 10% of most-cited papers, reflecting the influence and recognition of their work.
Features (Non-citation based)	**Description**
Publication Count(P)	The total number of research papers written by a researcher is represented by the number of publications each author has published.
Institutional Affiliations(I_{aff})	refers to the establishments or groups that the writer is connected to. Semicolons in the dataset allow researchers to have more than one affiliation.
Research Areas of Interest(R_{int})	Indicates the research fields or areas of expertise for each author. It offers details on the particular areas or topics that writers concentrate on when writing their works.
Collaborator Count(C_{collab})	indicates how many different co-authors a specific author has worked with during the course of their scholarly career.
Publication Gaps (G_p)	The total number of years during which an author has no publications within their research tenure.
Research Publication Tenure(T)	gives information about an author's seniority and experience by referring to the duration of their active research activity.

$$pi = \left(\frac{\sum_{i=1}^{P} C_i}{P} \right)^{\frac{1}{2}} \tag{3}$$

where:

- C_i is the number of citations received by the ith publication.
- P is the total number of publications.

up-index (upi):

When authors have varying contributions across publications, the p-index is adjusted by weighting authorship contributions via the A-index.

$$upi = \left(\frac{\sum_{i=1}^{P} \frac{C_i}{A_i}}{P} \right)^{\frac{1}{2}} \tag{4}$$

where:

- A_i is the number of authors for the ith paper.

Citation Growth Rate(G_c):

The citation growth rate G_c measures the average number of citations a researcher receives per year.

$$G_c = \frac{C}{T + \epsilon} \tag{5}$$

where:

- C is the total number of citations.
- T is the researcher's tenure (years since first publication).
- ϵ is a small constant to avoid division by zero.

Top 10% Highly Cited Paper Ratio(R_{10}):

This metric quantifies the proportion of an author's publications that fall within the top 10% of the most cited papers. To compute this feature we have considered only the set of publications in our dataset.

$$R_{10} = \frac{P_{10}}{P} \tag{6}$$

where:

- P_{10} is the number of publications in the top 10% of citations.
- P is the total number of publications.

3.5.2 Not citation based Characteristics

Publication Count(P):

The total number of publications authored by a researcher is given by:

$$P = \sum_{i=1}^{T} P_i \tag{7}$$

where:

- P_i is the number of publications in i-th year by the author.

Collaborator Count(C_{collab}):

The number of unique co-authors C_{collab} for a researcher is determined from the co-authorship network.

$$C_{collab} = \sum_{i=1}^{P} C_{collab-i} \tag{8}$$

where:

- $C_{collab-i}$ is the number of new co-authors in i-th publication by the author.

Publication Gaps (G_p):

A researcher's publication gap G_p is computed as the number of years where they did not publish any paper.

$$G_p = T - |\{y_i \mid y_i \in Y\}| \tag{9}$$

where:

- Y is the set of years in which the researcher published.

Research Publication Tenure(T):

The difference between an author's earliest and most recent publication years is their publication tenure T_p.

$$T_p = y_{\max} - y_{\min} + 1 \tag{10}$$

where:

- $y_{\max}$ is the year of the latest publication.
- $y_{\min}$ is the year of the first publication.

3.5.3 Method of Feature Estimation & Vectorization

In this section, we majorly discuss the methodology adopted to compute the various features that has been derived from the dataset and further used in our proposed model. Further, we discuss the various vectorization methods to represent the several categorical features in the dataset.

To estimate a researcher's Collaborator Count (C_{collab}), we identify their distinct coauthors from the collaboration connections stored in the "Coauthor" dataset. The Publication Gaps (G_p) represent the total number of years in which the researcher has no published work, derived from the paper information dataset. Meanwhile, the Research Publication Tenure(T) is calculated by calculating the interval between the year of their first publication that was noted and the most recent one that is accessible in the dataset. The year of the first publication is taken as the minimum publication year associated with the researcher. To compute the Top 10% Highly Cited Paper Ratio (R_{10}), we first extract the citation count (C) for all papers in the dataset. Next, we determine the 90^{th} percentile citation threshold to identify the top 10% of

publications. The researcher's publications that fall within this highly cited category are then counted, and their proportion is calculated accordingly.

The `CountVectorizer` class from the `sklearn.feature_extraction.text` module was used to convert the features Institutional Affiliations (I_{aff}) and Research Areas of Interest (R_{int}) into numerical representations. Each feature had its own instance of `CountVectorizer`, and the textual data was transformed into a numerical format using the `fit_transform` function. This approach tokenizes the text, learns the vocabulary, vectorizes it, counts the instances of each token, and displays the text as a matrix. `CountVectorizer`, which tokenizes the text and turns it into numerical representations by counting the occurrences of each token, was used to vectorize these features.

Next, to reduce the dimension of the features that would improve the computational complexity of our proposed model, we have adopted two methods: 1) feature extraction and 2) feature selection.

3.5.4 Feature Extraction

The feature extraction involved transforming the original dataset into a more meaningful form by focusing on the most significant features. To implement the feature selection technique, we employed three prominent methods as follows:

Principal Component Analysis (PCA): We chose PCA as our primary feature extraction technique because it effectively captures the majority of the variance within the dataset. By transforming the original features into a smaller set of uncorrelated components, PCA enabled us to reduce dimensionality, noise, and redundancy. This reduction not only speed up model training but also minimized the risk of overfitting. It focuses on the most informative aspects of the data that significantly enhance the accuracy and efficiency of our classification models. This key feature of PCA makes it the most beneficial technique in our analysis.

Independent Component Analysis (ICA): We further tested ICA to identify statistically independent components within the dataset. This technique uncovered hidden factors crucial for distinguishing between ECR and non-ECR categories, providing insights that other methods might have overlooked.

t-Distributed Stochastic Neighbor Embedding (t-SNE): Although this method was primarily employed for data visualization, t-SNE was valuable for us in mapping high-dimensional data into lower dimensional space. While it was less effective for improving classification accuracy, t-SNE gave us valuable insights into the separability of the categories and aided in understanding the data structure.

When we apply feature extraction techniques such as Principal Component Analysis (PCA), Independent Component Analysis (ICA), and t-Distributed Stochastic Neighbour Embedding (t-SNE), the original dataset is transformed into a new space with new feature axes called components. These components serve different purposes depending on the technique used. The component 1 in figure 3(b) captures the highest variance in PCA or the most independent information in ICA. It is a linear combination of multiple original features. It provides the most critical structure of the dataset and is the primary factor distinguishing data points. The component 2 is orthogonal (uncorrelated) to Component 1, capturing the next highest variance. It ensures



(a) PCA (b) ICA (c) t-SNE

Fig. 3: Feature-space visualizations produced by three feature extraction techniques: (a) PCA (Figure 3(a)), (b) ICA (Figure 3(b)), and (c) t-SNE (Figure 3(c)).

that no redundant information is carried from Component 1. It adds a new dimension of variation, helping to improve classification and data separation. On the other hand in *t*-SNE, component 1 and 2 do not correspond to actual features or statistical properties like PCA/ICA. They are simply visualization axes that help in plotting high-dimensional data into two-dimensional space. The purpose of these components is to help in identifying clusters and understanding the separability of different categories (e.g., ECR vs. non-ECR).

The figure 3(a) visualizes the PCA-transformed feature space, showcasing a clear separation between Early Career Researchers (ECRs) and non-ECRs. This demonstrates PCA's effectiveness in dimensionality reduction and clustering of the data. The figure 3(b) illustrates the ICA transformed feature space, emphasizing the separation achieved between ECRs and non-ECRs through Independent Component Analysis (ICA). This highlights ICA's role in extracting independent features for classification. The figure 3(c) depicts the t-SNE-transformed feature space, highlighting the ability of t-SNE to identify nonlinear clusters and separate ECRs from non-ECRs effectively.

When feature extraction techniques such as Principal Component Analysis (PCA), Independent Component Analysis (ICA), and t-Distributed Stochastic Neighbour Embedding (t-SNE) are applied, the original dataset is transformed into a new space defined by components. In PCA, Component 1 captures the maximum variance in the data, while Component 2 captures the next highest variance and is orthogonal to Component 1. In ICA, these components represent statistically independent directions in the data.

In contrast, for t-SNE, Component 1 and Component 2 are not interpretable as statistical features but serve as visualization axes that preserve the local structure of high-dimensional data in a two-dimensional space. These components help visualize clustering patterns and separability between classes, such as ECR and non-ECR groups.

Feature extraction techniques reveal that the original 12 features reduced dimension to 8, 8 and 2 for PCA, ICA and *t*-SNE respectively. Though the comparatively lower dimension of *t*-SNE is more better from the perspective of computational complexity, the method ignore the statistical behavior of the dataset. Thus, Principal Component Analysis (PCA) emerged as the most effective feature extraction method, significantly enhancing classification accuracy by transforming the dataset into a more insightful and manageable form. PCA effectively captures the majority of variance by converting data into a smaller set of uncorrelated components, enabling a more focused analysis. By reducing dimensionality, it accelerates model training and mitigates overfitting by eliminating noise and redundant features. Furthermore, by concentrating on the most informative features, PCA enhances classification performance, making it an optimal choice for distinguishing between ECR and non-ECR.

3.5.5 Feature Selection

Although the feature extraction method efficiently reduces the dimension that leverages the reduction of computational complexity retaining the performance of the proposed model, it loses the individuality of the features in the extracted feature dimension. Consequently, it was impossible to explain how different citation based and non-citation based variables contributed to the identification of ECR in the suggested model. Therefore, we used a feature selection strategy in the suggested model to decrease the feature dimension while maintaining the uniqueness of the citation and non-citation based characteristics. We conducted two standard feature selection techniques, i.e., a)Recursive Feature Elimination(RFE) and b)SelectKBest, to reduce the feature dimension by retaining only the most crucial features.

We calculate the Pearson correlation coefficient for each pair of characteristics in order to quantify the interdependence between the different citation based and non-citation based features utilized in the suggested model. We represent the same in figure 4 in the form of a heatmap. The correlation heatmap of the 12 features before the feature selection process reveals notable patterns in how various citation based and non-citation based metrics relate to each other. A strong positive correlation exists among Citation Count (C), p-index (pi), up-index (upi), h-index (hi), and Publication Count (P), indicating these metrics often co-vary and reflect overlapping dimensions of scholarly productivity and impact. Additionally, h-index shows a moderately strong relationship with Collaborator Count (C_{collab}), suggesting that researchers with higher impact tend to collaborate more. Research Tenure (T) and Publication Gaps (G_p) are highly correlated, implying that longer research careers often include inactive publication periods. In contrast, Research Areas of Interest (R_{int}), Institutional Affiliations (l_{aff}), and Citation Growth Rate (G_c) exhibit low or negligible correlations with most other features, highlighting their distinct nature. Particularly, G_c maintains near zero correlation across the board, reinforcing its unique role in capturing recent momentum in citation trends, independent of overall volume or historical productivity.

The observed dependencies further support the need to treat the predictor set as a collection of complementary dimensions rather than as independent indicators. Citation Count, h-index, p-index, up-index, and Publication Count exhibit strong positive associations, indicating that these variables capture partially overlapping aspects of

scholarly productivity and impact. Similarly, Research Publication Tenure and Publication Gaps are strongly related because longer observable research histories provide more opportunity for inactive publication periods. In contrast, Research Areas of Interest, Institutional Affiliations, and Citation Growth Rate exhibit comparatively weaker associations with the other predictors, suggesting that they provide additional information beyond aggregate productivity and impact. These dependencies motivate the subsequent use of feature-selection methods to remove redundant information while retaining complementary predictors.

These patterns justify the need for a careful feature selection process to eliminate redundancy and retain complementary indicators for effective ECR identification. This experiment not only helps us to understand the dependencies among the features but also reciprocates the selection of the most contributory technique of feature selection methods, i.e., RFE and SelectKBest. The correlation heatmap reveals key relationships between various features, shedding light on underlying dependencies within the scientific collaboration network which is required to understand for the model preparation.

Recursive Feature Elimination (RFE): We applied RFE using a Random Forest classifier, which evaluates feature importance at each iteration. RFE provides an iterative approach to feature selection, systematically eliminating the least significant features. This ensures that the final model is both efficient and focused on the most relevant features. When paired with a robust classifier like Random Forest, RFE not only selects the best features but also considers the importance of feature interactions, leading to better generalization and predictive accuracy.

SelectKBest: We also used SelectKBest to rank features. This method helped us to isolate statistically significant attributes, which improved model performance by discarding less informative variables. In our study, we employed two different criteria for feature selection using SelectKBest: a) Mutual Information: Measures the dependency between features and the target variable. b) Chi-Square Test: Assesses the statistical relationship between categorical features and the target class. Using these methods, we selected the top K features that had the highest significance in predicting ECR classification outcomes.

In figure 5, we present the result of the RFE and SelectKBest method. It shows that in both of the methods reduce the dimension to 9 distinct features. Both RFE and SelectKBest methods highlight the importance of features beyond mere citation count for identifying early-career researchers. The RFE plot reveals that publication gap (G_p), research tenure (T) and collaborator count (C_{collab}) are among the top-ranked features, emphasizing their significance in characterizing research trajectories. Interestingly, h-index (hi) and authorship weighted indices (pi, upi) also secure strong positions, reinforcing the role of balanced productivity and contribution in the selection process. Similarly, SelectKBest aligns with RFE by ranking G_p, T, and top 10% citation ratio (R_{10}) highly, while also showing notable preference for features like collaborator count (C_{collab}). The consistent prominence of G_c in both methods underlines its critical role in capturing the evolving impact of researchers' work over time especially valuable in distinguishing emerging researchers with upward citation trends.

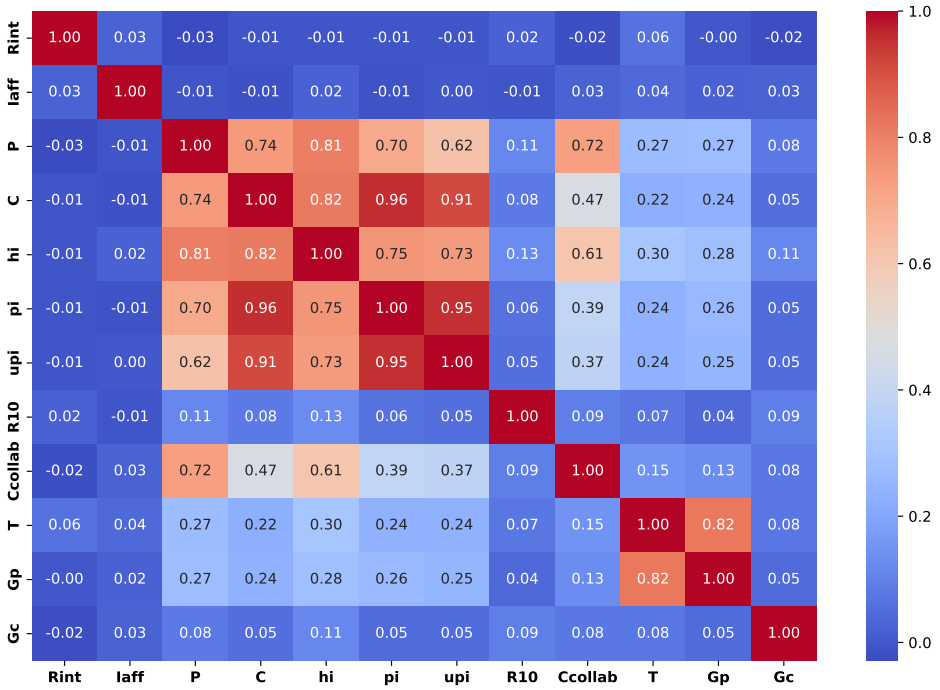


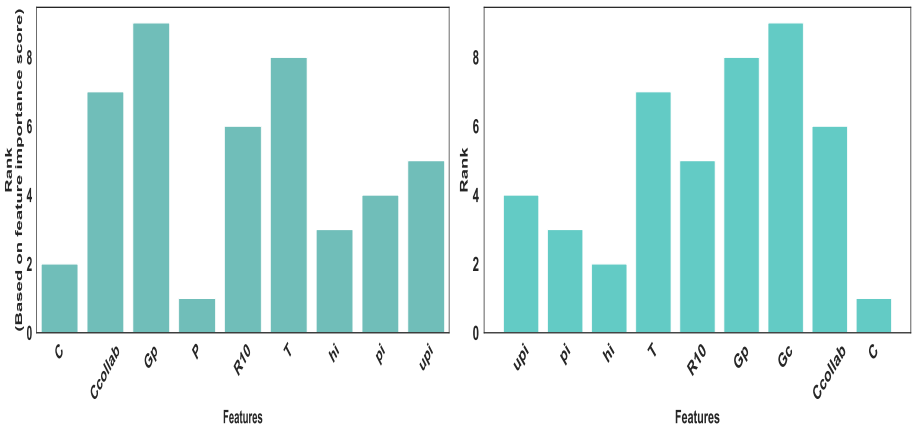
Fig. 4: Figure represents the correlation heatmap of the features used in our model. It include all the features before the feature selection.

To justify the usage of selected features along with the specific machine learning model for further explainability of the model, we have applied those features (identified by SelectKBest) with Random forest model to compare the performance of RFE and selectKBest. The result has been shown in the table 8. It shows that RFE achieved higher accuracy compared to SelectKBest. RFE also had a better Recall and F1-score, making it a superior choice for feature selection.

Method	Accuracy	Precision	Recall	F1-Score
RFE	0.9871	0.9844	0.9947	0.9895
SelectKBest	0.9839	0.9843	0.9895	0.9869

Table 8: Performance of Random Forest with RFE and SelectKBest.

We further plot the pairwise correlation of the 9 selected features(as selected using RFE method) shown in figure 6. Key relationships between the citation based and non-citation based metrics employed in the suggested model are highlighted in the correlation heatmap of the 9 characteristics that were chosen. Strong positive correlations



(a) RFE (b) SelectKBest

Fig. 5: Figure shows the comparison of the selected features in two different kind of feature selection techniques i.e., a) RFE, and b)SelectKBest shown in figure figure 5(a), and 5(b) respectively. In fig. 5(a), we plot feature importance score based rank of the selected features along y-axis and in fig. 5(b), we plot Rank of the feature along y-axis. In both of the figure along x-axis, we represent the selected features for the ECR classifier.

are observed between Citation Count (C), p-index (pi), and up-index (upi), indicating these features reflect similar aspects of research impact. The h-index (*hi*) also shows a strong association with Citation Count (C) and Publication Count (P), reinforcing its role as a combined measure of productivity and influence. Research Tenure (T) and Publication Gaps (G_p) are highly correlated, suggesting longer careers often include inactive periods. Collaborator Count (C_{collab}) maintains moderate correlation with several productivity indicators. Notably, Citation Growth Rate (G_c) shows minimal correlation with any other feature, suggesting it captures a distinct aspect of an author's academic trajectory.

The correlation analysis confirms that the selected features provide both overlapping and complementary information for ECR identification. While some citation based features are closely related, their combined use strengthens the model's ability to capture research impact. The inclusion of non-citation based features supports the model's broader understanding of early career patterns. Importantly, Citation Growth Rate (G_c) plays a unique role in the model. Its independence from other features allows it to highlight authors who may have a lower total citation count but are experiencing a rapid rise in citations. This makes G_c especially useful for identifying active and emerging researchers, improving the model's ability to distinguish ECRs with recent momentum. This observation also corroborates the rank of the G_c in RFE method based feature selection technique as shown in figure 5.

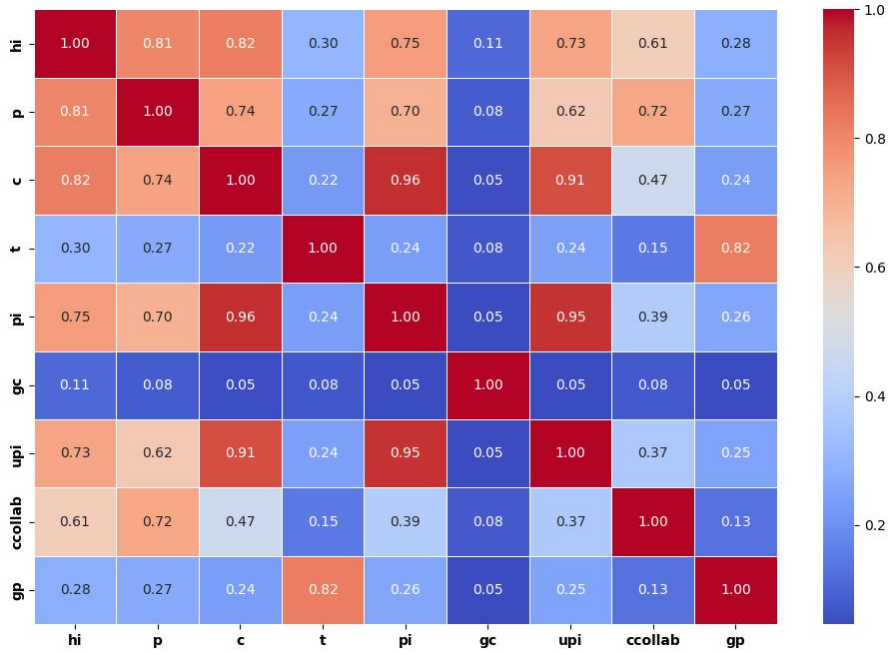


Fig. 6: Figure represents the correlation heatmap of the selected features(produced by RFE method) used in our model.

3.6 Machine Learning Model Selection

The machine learning algorithms were selected to represent complementary modeling assumptions for structured author level bibliographic classification. The objective was not to exhaustively evaluate every available classifier, but to compare representative approaches from different model families, including linear, probabilistic, instance based, tree based, bagging, and boosting methods. This design allows the classification problem to be evaluated under different assumptions about the relationships between bibliometric predictors and ECR status.

Logistic Regression was included as an interpretable linear baseline. It provides a reference for determining whether ECR status can be separated using approximately linear relationships between the predictors and the target. Decision Tree was included as a simple nonlinear model capable of representing threshold-based relationships among publication, citation, and trajectory variables.

Random Forest was selected as a bagging based ensemble because it can represent nonlinear relationships and interactions among bibliometric variables while reducing the variance associated with an individual decision tree. Gradient Boosting and XGBoost were included as boosting based ensemble methods because they sequentially improve weak learners and can capture complex nonlinear relationships and interactions in structured tabular data.

Support Vector Machine was included as a margin based classifier, providing a modeling assumption that differs from tree based approaches. K-Nearest Neighbors was included as an instance based method that evaluates the local similarity structure of researchers in the feature space. Naive Bayes was included as a probabilistic baseline based on a substantially different conditional-independence assumption. AdaBoost was included as an adaptive boosting method that emphasizes observations that are more difficult to classify.

The resulting model set therefore covers several complementary classification families rather than relying exclusively on a single class of algorithms. This comparison is particularly relevant because the relationship between bibliometric predictors and ECR status may involve both approximately linear and nonlinear interactions.

Deep neural networks were not included in the primary classifier comparison. The present study represents each researcher using structured, aggregated author level bibliometric variables rather than raw publication text, images, or directly modeled temporal sequences. Moreover, a central objective of the study is to provide interpretable feature level evidence through SHAP and LIME. Classical supervised learning therefore provides an appropriate primary benchmark for the present feature representation. Deep learning approaches, including graph neural networks and sequence-based architectures, remain potential extensions for future work when richer temporal or network representations are modeled directly. Related deep learning work in scientometrics has primarily addressed citation prediction rather than the career stage classification problem considered here [40].

Table 9 summarizes the rationale for selecting the machine learning classifiers evaluated in this study.

3.6.1 Machine Learning Algorithms

Nine distinct supervised classifiers Logistic Regression, Support Vector Machine, Random Forest, Decision Tree, K-Nearest Neighbour, Gradient Boosting, AdaBoost, XGBoost, and Naive Bayes—were used in our suggested approach. In this section, we provide a quick overview of the machine learning classifiers that we employed in our suggested model and provide justification for their utility in this particular situation.

Logistic Regression

In this instance, occurrences are divided into two groups ECR and non-ECR using logistic regression. It is designed especially to solve binary classification issues. Small data sets can be effectively used for logistic regression. Even with a small initial dataset, logistic regression can generate accurate findings with little training data. We can identify which characteristics have the biggest effects on the likelihood of being an ECR by looking at the coefficients linked to each feature.

Support Vector Machine

SVM operates efficiently even if there are more features (the qualities of the researcher) than trials. Numerous attributes, such as name, affiliations, number of publications, citations, h-index, and more, are used to characterize each author in our dataset. SVM is capable of handling these high dimensional feature spaces with ease. SVM can

Table 9: Rationale for the machine learning classifiers used for ECR identification.

Model	Model family	Methodological role
Logistic Regression	Linear	Provides an interpretable linear baseline and tests whether approximately linear relationships are sufficient for ECR classification.
Decision Tree	Tree-based	Captures nonlinear threshold relationships and provides easily interpretable decision rules.
Random Forest	Bagging ensemble	Models nonlinear relationships and feature interactions while reducing the variance of individual decision trees.
Gradient Boosting	Boosting ensemble	Captures nonlinear relationships by sequentially improving weak learners.
XGBoost	Regularized boosting	Provides a strong tabular-data benchmark capable of modeling complex feature interactions and nonlinearities.
Support Vector Machine	Margin-based	Provides a complementary classification assumption based on separating class boundaries in feature space.
K-Nearest Neighbors	Instance-based	Tests whether local similarity among researchers in the feature space provides useful information for ECR classification.
AdaBoost	Adaptive boosting	Emphasizes difficult-to-classify observations and provides a complementary boosting strategy.
Naive Bayes	Probabilistic	Provides a computationally efficient probabilistic baseline based on conditional-independence assumptions.

help us to identify which features are most important for the ECR classification by providing us with information about their significance.

Random Forest

Random Forest may be a good option for identifying ECRs (Early Career Researchers) due to its capacity to handle numerous input features and their non-linear correlations. Random Forest can help reduce overfitting and reduce the system's susceptibility to noise and data anomalies by choosing a subset of each tree's characteristics at random.

Decision Tree

The classification rules provided by a decision tree classifier are simple to comprehend and use. The classifier's decision-making process can be understood by looking at the tree structure, which shows a number of if-else situations depending on certain

characteristics. Decision trees are a useful tool for capturing intricate nonlinear connections between the objective variable and the information. In author datasets, The relationship between the target value and features (such the quantity of publications, citations, h-index, etc.) might not follow a distinct linear pattern. Decision trees generate splits according to different feature combinations and thresholds, making them useful for managing this kind of non-linearity.

K-Nearest Neighbour

K-Nearest Neighbors (KNN) was included as an instance-based classifier to evaluate whether local similarity among researchers in the feature space contains useful information for distinguishing ECRs from established researchers. Its inclusion provides a modeling assumption that differs from the parametric, tree based, and boosting approaches.

Gradient Boosting

There may be an imbalance between the number of ECR and non-ECR instances in the dataset when it comes to ECR identification. Gradient boosting is an efficient method in this regard. The ability of this classifier to capture complex nonlinear relationships between the target variable and features is well established.

AdaBoost

We used AdaBoost's feature importance measure to identify which of our features in the author's dataset are critical for finding ECRs. Understanding which elements are relevant and how they influence the ECR categorization process is helpful. The classifier can improve overall performance by learning and focusing on cases that are more likely to be misclassified by iteratively weighting the misclassified cases.

XGBoost

XGBoost was included as a regularized gradient-boosting method able to capture nonlinear relationships and interactions among heterogeneous bibliometric predictors. Its computational efficiency and regularization make it a strong benchmark for structured tabular classification.

Naive Bayes

Naive Bayes was included as a probabilistic baseline based on conditional independence assumptions. Its substantially different modeling assumptions provide a useful contrast with tree based and margin based classifiers.

3.6.2 Explainable AI (XAI)

To improve the interpretability of our ECR identification model, we used Explainable AI (XAI) methods to provide insights on both global and local importance of features. In particular, we used Shapley Additive Explanations (SHAP) and Local Interpretable Model agnostic Explanations (LIME) to better understand how features contribute to the outcome of classification.

Local and Global Interpretability with SHAP:

We applied SHAP values to assess both global and local interpretability. In particular, globally SHAP provides a unified measure of the contribution of each feature to the model across all predictions, highlighting which features consistently influence the classification decisions. On the other hand, locally, SHAP helps to explain the model's predictions by attributing the model's output to a set of input features, offering transparency for each prediction made by the model.

Local Interpretability with LIME:

In addition, we implemented LIME to analyze local explanations in more detail. LIME approximates the behavior of the complex model around a specific prediction using a simple, interpretable model. This approach allowed us to validate how the classifier behaves on a per instance basis, ensuring that the model's predictions align with domain knowledge and the expectations for Early Career Researcher identification.

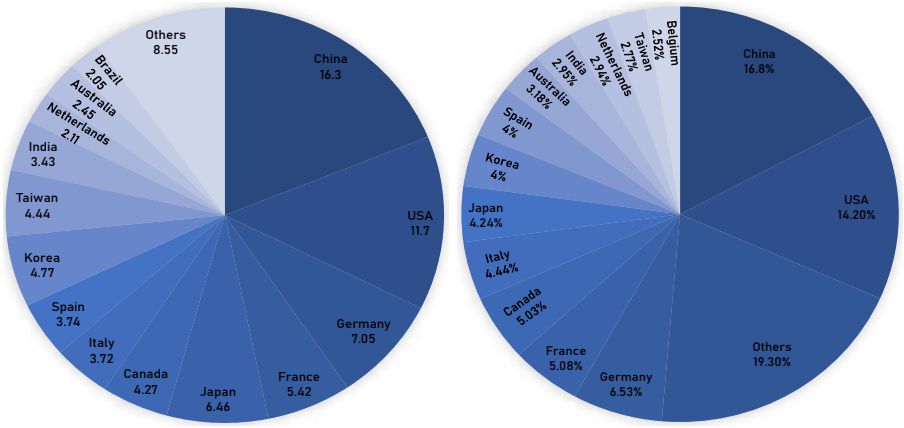
By integrating these explainability methods, we gained a clearer understanding of the influence of both citation based and non-citation based features, thereby improving trust and transparency in the proposed machine learning framework.

4 Results and Discussion

In this section, we initially present the several metrics used in our work to measure the performance of the experiment conducted. Next, we outline the detailed results and findings from the proposed ECR identification model. We evaluate models performance, interpretability, as well as validation of the model in the context of the objective of the work.

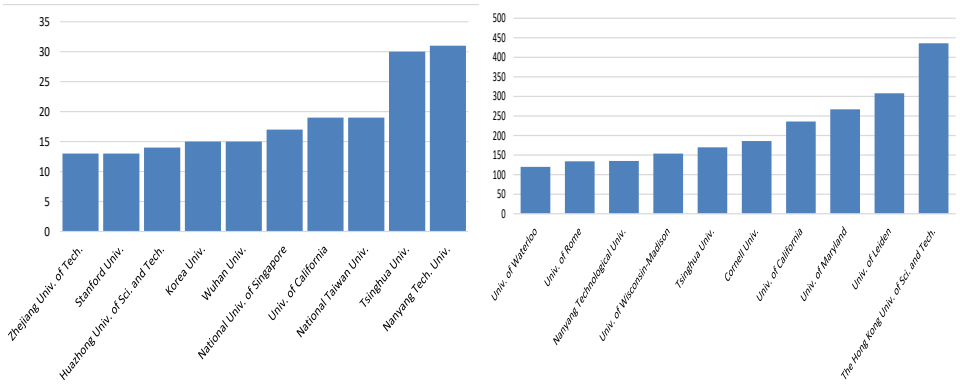
4.1 Exploratory Distribution of Publication Discontinuation

The initial analysis of the bibliographic dataset in the the discipline of Computer Science & Engineering reveals a significant prevalence of *"drop-out"* activities during the early career stage worldwide. A similar pattern has been identified by the authors in [41], further validating the presence of such occurrences within the scientific research domain. Figure 7(a) presents a pie diagram illustrating the geographic allocation of researchers who are *"drop-out"*, highlighting their proportional representation across both developed and developing regions. Early publication discontinuation may be associated with multiple structural and career related factors. However, such factors are not directly observed in the present bibliographic dataset and are therefore outside the scope of causal inference in this study.Additionally, Figure 7(b) depicts the geographic distribution of articles written by *"drop-out"* scholars, further reinforcing the insights drawn from Figure 7(a). A key observation is that while developed countries exhibit an increased total number of dropouts, they simultaneously sustain considerable publication output, indicating that researchers in these regions make substantial scholarly contributions before exiting academia. On the other hand, growing regions show less proportions in the figures, emphasizing persistent difficulties like limited funding and



(a) Author distrib. across Country (b) Author's Publication distrib. across Country

Fig. 7: Figure 7(a) shows a pie chart that shows the proportion of *'drop-out'* authors in each nation. In the pie chart, the nations that contribute $\leq 2\%$ of the total share are designated as "others." Figure 7(b) shows the percentage of publications by *'drop-out'* authors in each country as a pie chart. The countries that contribute $\leq 2\%$ of the pie chart's share are indicated as "others" in both representations.



(a) Author distrib. across Top Institute (b) Author's Publication distrib. across Top Institute

Fig. 8: Figure 8(a) shows the top 10 institutes ordered by the number of *'drop-out'* writers from our dataset in a horizontal bar chart. The y-axis shows the name of the institute, while the x-axis shows the number of authors.

Figure 8(b) shows the amount of articles the *'drop-out'* authors contributed, based on our dataset, and ranks the top 10 institutes in a horizontal bar chart. The y-axis of the chart shows the name of the institute, and the x-axis shows the number of publications that the *'drop-out'* writers contributed.

institutional backing. This analogy emphasizes the urgent requirement for global regulations aimed at reducing inconsistencies in the study environments, ensuring the higher accumulation and meaningful academic contributions.

Figures 8(a) and 8(b) depict the distribution of *"drop out"* researchers' institutes and the works that correspond to them. Figure 8(a) determines which top 10 universities have the most *drop out* researchers, revealing that main organizations, specifically from the China, the United States, and certain European countries, are prominently represented.Meanwhile, Figure 8(b) depicts the publication outcome of drop-out researchers from these institutes, revealing significant differences in their productivity level. Similar terminologies based on the publication discontinuation trajectory have previously been reported in Scientometrics studies [41]. Indeed, certain universities have abnormally low publication numbers with a high dropout rate, suggesting that variables such as institutional regulations, quality of mentorship, and study support system may greatly impact a researcher's retention and productivity [42]. On the flip side, institutions that publish more but have fewer dropouts might have a higher impact research culture and better support structures for early-career researchers. These findings call for a concerted effort needed to promote the ECRs' well-being to achieve equitable growth in science and technology.

4.2 Performance Metrics

This section represents the different metrics that have been used to measure the performance of the model. It includes metrics to measure the confidence of agreement between the expert annotators. We further explain the ML model performance metrics including the metric for cross validation.

Cohen's Kappa

Cohen's Kappa (κ) is a statistical measure that quantifies inter-rater agreement for categorical data while accounting for chance agreement. It is defined as:

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

where:

- P_o (Observed Agreement) is the proportion of times the raters agree.
- P_e (Expected Agreement) is the probability of agreement occurring by chance.

The value of κ lies in the range:

$$-1 \leq \kappa \leq 1$$

Confusion Matrix

The confusion matrix provided below can be used to assess how well the machine learning classifier identified ECR researchers from the bibliographic information [43]:

	Non-ECR	ECR
Non-ECR	TN	FP
ECR	FN	TP

Where,

1. **TN**= Total amount of correctly classified non-ECRs.
2. **TP**=The total number of ECRs accurately classified.
3. **FP**= ECRs that were incorrectly classified as non-ECRs.
4. **FN**= Non-ECRs that were incorrectly classified as ECRs.

Accuracy

For classification algorithms, accuracy is a frequently used evaluation parameter. It calculates the proportion of accurate predictions the algorithm made on a particular dataset. By dividing the number of accurate predictions made by the algorithm by the total number of predictions, one can compute accuracy mathematically:

$$Accuracy = \frac{TN + TP}{TP + FN + FP + TN} \tag{11}$$

Precision

One performance parameter used to assess a classification algorithm's effectiveness is precision which tells how well a classification algorithm performs, particularly when it comes to positive class predictions. The mathematical formula for precision is:

$$Precision = \frac{TP}{TP + FP} \tag{12}$$

Recall

A performance parameter called recall quantifies the model's capacity to accurately detect positive examples or the percentage of real positive examples that the algorithm correctly detects. The following is the formula to determine recall:

$$Recall = \frac{TP}{TP + FN} \tag{13}$$

F1-score

A popular metric for assessing an algorithm's performance in classification problems is the F1-score. It provides a fair evaluation of the algorithm's performance by combining precision and recall into a single value. The following formula is used to determine the F1-score, which is the harmonic mean of precision and recall:

$$F_1 = \frac{2 \times (Precision \times Recall)}{Precision + Recall} \tag{14}$$

ROC

A binary classification algorithm's performance is represented graphically by the Receiver Operating Characteristic (ROC) curve, which shows the trade-off between the true positive rate (TPR) and the false positive rate (FPR) for various classification

thresholds.

$$TPR = \frac{TP}{TP + FN} \tag{15}$$

$$FPR = \frac{TN}{TN + FP} \tag{16}$$

AUC

The performance of binary classification models is assessed in machine learning using the AUC (Area Under the Curve) measure, which is usually shown in the form of a Receiver Operating Characteristic (ROC) curve.

Leave-One-Out Cross-Validation (LOOCV)

A model evaluation method called Leave-One-Out Cross-Validation (LOOCV) in which each sample in a dataset of size n is used once as a test set, while the remaining $n - 1$ samples form the training set. This process is repeated for all n samples, and the overall performance is averaged.

Mathematically, for a dataset $D = \{(x_i, y_i)\}_{i=1}^{n}$, the LOOCV error is computed as:

$$E_{\text{LOOCV}} = \frac{1}{n} \sum_{i=1}^{n} L(y_i, f_{-i}(x_i))$$

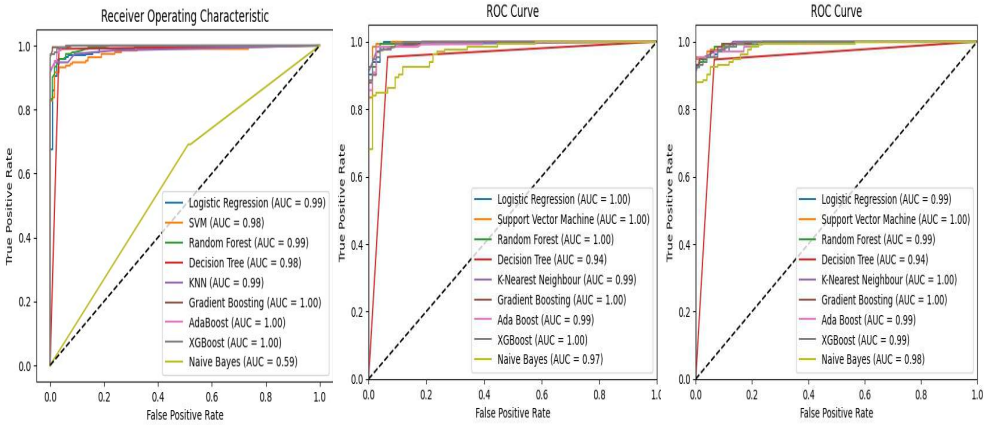
where f_{-i} is the model trained on D_{-i} (excluding (x_i, y_i)), and $L(\cdot)$ is a loss function. Although LOOCV provides an unbiased estimate of model performance, it is computationally intensive for large datasets. In our experiment, we are able to implement this method of validation as our dataset for the experiment is comparatively small in size.

4.3 Confidence of Agreement Between Expert Annotators

To evaluate the agreement between the experts who provided annotations, we decided to use Cohen's Kappa Test. This test estimates inter-rater reliability and allows us to determine whether there is a mistake in the data labeling. As a result, we found that our agreement was substantial with the score of **0.92**. Therefore, the consensus between the experts who worked on the data labeling task was quite high, which ensures the quality of our data sample.Our result confirms that the dataset annotations are highly reliable, reducing potential biases in model training. This, in turn, enhances the credibility of the proposed classification framework by ensuring that the training data accurately reflects the targeted classification criteria. This agreement assessment provides an additional validation of the expert based annotation procedure described in Section 3.3.2.

4.4 Algorithm Performance for Machine Learning

This section compares the classifier's performance metric and highlights the effectiveness of the machine learning classifier we used in our work. We initially show the performance of the nine several machine learning model with the identified 12 distinct features. Further to reduce the feature dimension, we have used two methods-



(a) ROC: All Features (b) ROC:Feature Extraction (c) ROC:Feature Selection

Fig. 9: Receiver operating characteristic (ROC) curves for the nine classifiers under three setups: (a) all features (Figure 9(a)), (b) feature extraction (Figure 9(b)), and (c) feature selection (Figure 9(c)).

a) feature extraction and b) feature selection. We show the performance of feature extraction and feature selection respectively in the following.

Classifier	Accuracy	Precision	Recall	F1-Score	AUC
Logistic Regress.	0.93	0.95	0.93	0.94	0.99
SVM	0.95	0.94	0.97	0.96	0.99
Random Forest	0.96	0.98	0.96	0.97	0.99
Decision Tree	0.99	0.99	0.99	0.99	0.99
KNN	0.92	0.93	0.93	0.93	0.98
Gradient Boosting	**0.99**	**1.00**	**0.99**	**0.99**	**1.00**
AdaBoost	0.99	0.99	0.99	0.99	1.00
XGBoost	**0.99**	**1.00**	**0.99**	**0.99**	**1.00**
Naive Bayes	0.56	0.64	0.54	0.58	0.56

Table 10: Performance of Machine Learning Model applied for classification.

As observed in Figure 9(a) and detailed in Table 10, models trained on the full feature set demonstrate strong predictive performance. Gradient Boosting, AdaBoost, and XGBoost each achieve an AUC of 1.00, with Logistic Regression, Random Forest, and KNN close behind at 0.99. SVM and Decision Tree perform well with AUCs of 0.98. However, Naïve Bayes underperforms with an AUC of 0.59, likely due to its strong independence assumptions not aligning with the data structure. These results indicate that the full feature set holds substantial discriminatory power, favoring ensemble-based models for optimal performance.

Figure 9(b) and Table 11 show that after applying feature extraction techniques(PCA method), model performance remains consistently high. Gradient Boosting, XGBoost, and SVM maintain perfect AUC scores (1.00), while KNN and

Classifier	Accuracy	Precision	Recall	F1-Score	AUC
Logistic Regress.	**0.98**	**0.97**	**0.99**	**0.98**	**0.99**
SVM	**0.98**	**0.98**	**0.98**	**0.98**	**0.99**
Random Forest	**0.97**	**0.98**	**0.97**	**0.98**	**0.99**
Decision Tree	0.94	0.96	0.95	0.95	0.94
KNN	0.97	0.98	0.96	0.97	0.99
Gradient Boosting	0.96	0.98	0.96	0.97	0.99
AdaBoost	**0.98**	**0.98**	**0.98**	**0.98**	**0.99**
XGBoost	0.95	0.97	0.95	0.96	0.99
Naive Bayes	0.88	0.88	0.94	0.91	0.96

Table 11: Performance of Machine Learning Model applied on the results of feature extraction.

AdaBoost slightly drop to 0.99. Logistic Regression also shows excellent generalization with an AUC of 1.00. The Decision Tree dips to 0.94, indicating its sensitivity to feature transformation. Notably, Naïve Bayes improves significantly (AUC = 0.97), suggesting that feature extraction helped reduce inter-feature dependencies, thereby boosting its effectiveness.

Classifier	Accuracy	Precision	Recall	F1-Score	AUC
Logistic Regress.	0.94	0.93	0.98	0.95	0.99
SVM	**0.95**	**0.94**	**0.98**	**0.96**	**0.99**
Random Forest	**0.95**	**0.97**	**0.94**	**0.96**	**0.99**
Decision Tree	0.94	0.96	0.94	0.95	0.94
KNN	**0.95**	**0.96**	**0.96**	**0.96**	**0.99**
Gradient Boosting	**0.96**	**0.96**	**0.96**	**0.96**	**0.99**
AdaBoost	0.94	0.96	0.95	0.95	0.99
XGBoost	**0.95**	**0.96**	**0.95**	**0.96**	**0.99**
Naive Bayes	0.89	0.86	0.99	0.92	0.98

Table 12: Performance of Machine Learning Model applied on the results of feature selection.

In Figure 9(c) and Table 12, results from SelectKBest feature selection similarly reflect high model performance. Gradient Boosting, XGBoost, and Logistic Regression retain AUCs of 1.00, while AdaBoost and KNN follow closely with 0.99. The Decision Tree again lags slightly behind (AUC = 0.94), and Naïve Bayes continues to show improvement with an AUC of 0.97. These findings highlight the utility of selecting the most relevant features in preserving classifier efficacy while reducing computational complexity.

Across all three setups, ensemble methods such as Gradient Boosting, AdaBoost, and XGBoost consistently outperform others, showing resilience across both full and reduced feature spaces. Logistic Regression also remains robust under all configurations. The performance gain of Naïve Bayes in Tables 11 and 12, compared to Table 10, underscores the advantage of dimensionality reduction for simpler probabilistic models. These observations support the conclusion that the combination of feature reduction techniques and ensemble classifiers offers a compelling balance between predictive accuracy and computational efficiency.

4.5 Performance of Validation Test

In this section, we represent the performance of several validation tests adopted in our work to estimate the confidence of the performance of the proposed model. To do this, we have performed a Leave One Out Cross-Validation(LOOCV) approach. Finally, we also represent the various domain-specific validation measures to estimate the performance of the proposed model in the context of the objective of the work.

4.5.1 LOOCV

We employed the Leave-One-Out Cross-Validation (LOOCV) approach to rigorously assess the model's performance on individual data points. This method yielded robust metrics across all evaluation dimensions, thus it demonstrates the model's consistency. The accuracy of the model was **0.985 (± 0.01)**, while precision was reported as **0.980 (± 0.02)**. Recall and F1-Score were recorded as **0.987 (± 0.01)** and **0.983 (± 0.01)**, respectively. These results indicate the model's ability to generalize effectively across diverse data points.

4.5.2 Correlation Analysis

We further perform the correlation analysis to highlight the significant relationships between the model's predictions and *h*-index values. We performed four different types of correlation measures, i.e., a) Pearson correlation coefficient, b) Spearman Correlation coefficient, c) Kendall's Tau, and d) distance correlation. We performed four different correlation measures to ensure a comprehensive analysis, as each metric captures different aspects of association—Pearson for linear relationships, Spearman and Kendall's Tau for monotonic and ordinal associations, and Distance Correlation for both linear and nonlinear dependencies.

The Pearson correlation coefficient was **-0.71**, indicating a strong negative linear relationship between predicted probabilities and *h*-index. The Spearman correlation coefficient was **-0.89**, reflecting a strong monotonic relationship, while Kendall's Tau was **-0.80**, signifying ordinal association. Additionally, the Distance Correlation of **0.84** captured both linear and nonlinear dependencies effectively. Overall, the strong negative correlations observed across Pearson, Spearman, and Kendall's Tau indicate that the model consistently assigns higher probabilities to entities with lower *h*-index values. This suggests a meaningful inverse relationship between the model's predictions and h-index. The high distance correlation (0.84) further confirms that this dependency is not limited to linear trends but also captures complex, nonlinear relationships. The consistency across these diverse measures strengthens confidence in the model's predictive alignment with domain-specific expectations.

4.5.3 Bibliometric Threshold Validation

In this section, we perform the analysis of citation count and publication count growth rate. Though the citation growth rate (G_c) has been defined in section 3.5, the publication growth rate(G_p) metric has been introduced in this section for validation of the proposed model.

Publication Growth Rate(G_{pub}):

The publication growth rate G_{pub} measures the average number of publications a researcher produced per year.

$$G_{pub} = \frac{P}{T + \epsilon} \tag{17}$$

where:

- P is the total number of publications.
- T is the researcher's tenure (years since first publication).
- ϵ is a small constant to avoid division by zero.

We performed the threshold validation test to ensure that researchers with high h-index or publication counts were not misclassified as ECRs as well as the Non-ECRs with moderately low publication and citation count would not be misclassified as ECR. It reflects the reliability of the model in distinguishing career stages. It had been observed that predicted ECRs exhibited higher growth rates, with a median citation growth rate of **7.1** and a publication growth rate of **1.8**. These findings align with the expectation that ECRs are in a rapid development phase of their academic careers.

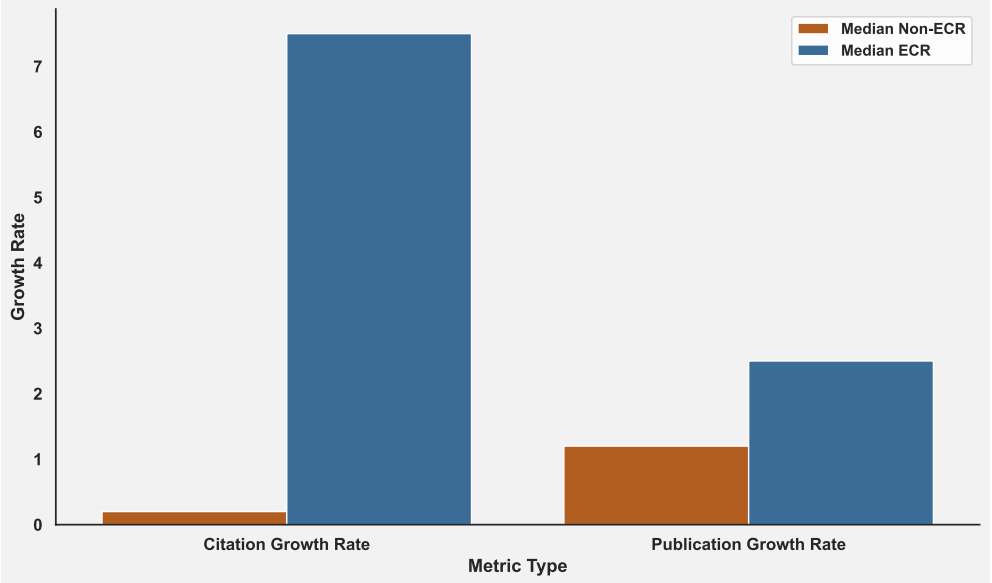


Fig. 10: Distribution of growth rates for citations (G_c) and publications (G_{pub}) comparing ECRs and non-ECRs. Predicted ECRs demonstrate higher growth rates in both metrics.

Figure 10 compares the growth rates for citations (G_c) and publications (G_{pub}) between ECRs and non-ECRs. Notably, ECRs exhibit higher growth rates in both metrics, aligning with the expectation that they are in a rapid development phase of

their academic careers. This empirical evidence supports our hypothesis that early-career researchers, despite shorter overall tenures, can quickly accumulate research output and influence relative to established researchers. This result also corroborates that though the established researchers have large number of publication and citation count, the stature of the career took long tenure to attain.

4.6 Analysis of Feature Importance

We performed explainable AI models to reveal the features importance in the identification of ECR by our proposed model. This experiment reciprocates the required output to answer the RQs that had been raised in the introduction section. In this study, we employed both Local Interpretable Model agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) to enhance the interpretability of our machine learning model for classifying Early Career stage Researchers (ECRs) and non-ECRs. Given the complex nature of our classification task, which incorporates both citation based and non-citation based features, using multiple interpretability techniques provides a more robust and comprehensive understanding of feature contributions and model behavior.

By integrating both LIME and SHAP, we ensured a dual-layered interpretability framework: LIME facilitated instance level understanding, while SHAP provided a holistic view of feature importance. This complementary approach strengthens the trustworthiness of our model, aids in identifying potential biases, and ensures that our findings align with domain knowledge in academic career progression analysis.

SHAP Model

The SHAP (SHapley Additive exPlanations) analysis provides an in-depth understanding of the contribution of individual features toward the model's predictions. The results obtained from the SHAP evaluation reveal the extent to which specific attributes influence the model output, both positively and negatively. The SHAP summary plot in Figure 11 illustrates these feature contributions. This analysis highlights the model's reliance on meaningful bibliometric indicators, ensuring its interpretability and alignment with academic metrics.

The SHAP summary plot demonstrates that certain features exhibit a stronger positive impact, while others contribute negatively. The spread of SHAP values suggests a nonlinear relationship between features and outcomes, emphasizing the complex interplay of multiple factors in determining the influence within the scientific collaboration network. Furthermore, the presence of overlapping SHAP values in certain features suggests interdependencies, which could be further explored using interaction effects analysis. This empirical evidence provides a robust foundation for modeling network dynamics in multidisciplinary domains.

The SHAP based interpretability analysis reveals critical insights into the predictive power of different features. The findings underscore the importance of tenure, citation count, and consistent research output in determining an entity's significance within the scientific collaboration network. It suggests that sustained engagement and higher citation count enhance a researcher's significance within the scientific collaboration network. Conversely, a longer publication gap negatively impacts predictions,

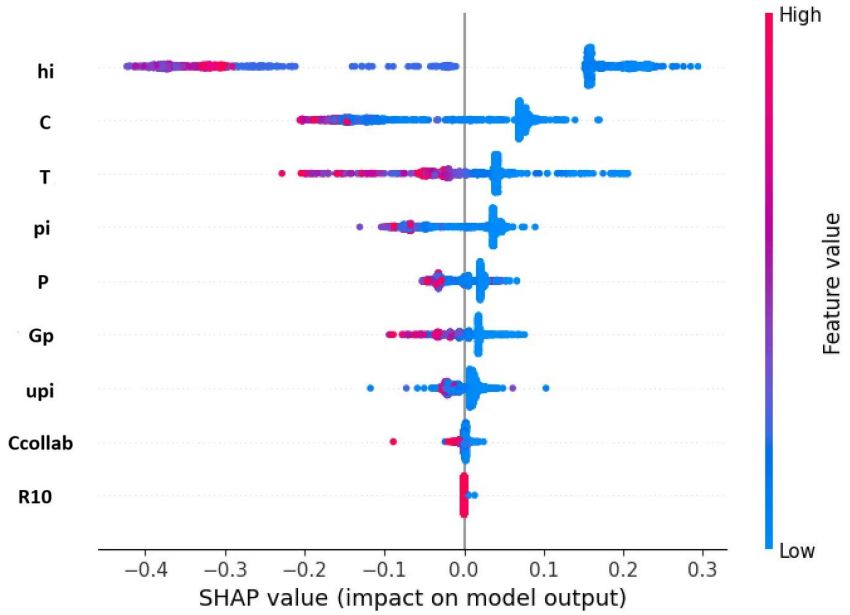


Fig. 11: Figure represents the SHAP summary plot. Along x-axis, we plot the SHAP value and along y-axis, we represent the feature used in our proposed model.

indicating that researchers with consistent research contributions are perceived as more influential. Additionally, features such as total publication count and popularity index(with unequal contribution) further reinforce the importance of sustained research activity. These observations contribute that though the citation based features are predominant in accessing and further ranking of established researchers in the scientific community, non-citation based features play a more prominent role identification of ECR from large bibliographic datasets.

LIME Model

To complement the SHAP analysis, we employed Local Interpretable Model Agnostic Explanations (LIME) to gain a more granular, instance level understanding of our classification model's behavior.While SHAP gave us a global view of the feature importance, the LIME method allows us to dig deeper into the individual test instances and analyze which exact features contribute to the model's classification decisions.

Figure 12 shows us the LIME-based feature importance analysis for multiple test instances in our data set. Every horizontal stacked bar represents a test instance and its features' contribution to the model's decision to classify this instance as ECR. The features contributing positively to the decision to classify the instance as an ECR are shown in warmer colors (red), while the ones contributing negatively are shown in cooler colors (blue). The size of the segment is proportional to the feature's importance

for the instance's classification. The legend on the right-hand side of the figure shows the correspondence between the feature names and the segments' colors.

We can see that for all instances almost always, the most important features are `Collab-count`, `hi` (h-index), and `Tenure` - this once again validates the importance of these features in our model. This means that, for example, the higher is the number of potential collaborators known to the researcher, the higher is the importance of this feature in classifying this instance as an ECR, with lower number of citations in general having negative impact. The higher is the h-index of the researcher, the more important it is for the model to classify this instance as an ECR. And finally, the longer the researcher has been active in the field, the higher is its impact on the classification.

The importance of the `Pub-gap` (the publication gap) may vary across instances - for some instances, it negatively contributes to the classification, while for others, it does not have much impact.

For example, Instance 3 was classified as ECR mainly due to a large collaboration count and recent increase in citations, while the publication gap was large for this instance as well. Meanwhile, Instance 5 was classified as non-ECR mainly due to a large publication gap, the lack of collaborations, and the absence of recent publications.

One thing that stands out in this analysis is the fact that non-citation based features impact the model's decisions in a significant way. Multiple instances show that researchers with low citation counts can still be classified as ECRs if they have good collaboration networks and are actively publishing.

By providing an in-depth perspective, the approach of LIME offers an additional layer of explainability on top of SHAP's global interpretability. The model's focus on tenure, collaboration, and citations reflects domain knowledge, thus validating that early-career identification is a multi-faceted endeavor. The instance-based nature of LIME underscores the diversity of ECR trajectories, thus highlighting the model's ability to identify emerging scholars via different career paths.

Insight of the Experiments

The results of our experiments confirm the effectiveness of the proposed model for ECR identification. A high accuracy of the model, which was demonstrated on multiple evaluation metrics, as well as its close approximation to the bibliometric indicators specific to the target category, ensure its adequacy. Furthermore, the interpretability of the obtained results, demonstrated through the SHAP and LIME explanations, allows us to gain an additional understanding of the relationship between the independent variables and the dependent one. According to the visualizations presented in Figure 12, the non-citation based features greatly impact the final prediction, thus compensating for the bias in citation-based features. In addition, the analysis of the growth rate of the number of publications and citations presented in Figure 10 supports the initial hypothesis about the higher development rate of ECRs compared to established researchers. Thus, the consideration of both types of variables in the constructed model contributes to a better understanding of the ECR concept in general and allows more accurate identification within large-scale bibliometric data sets.

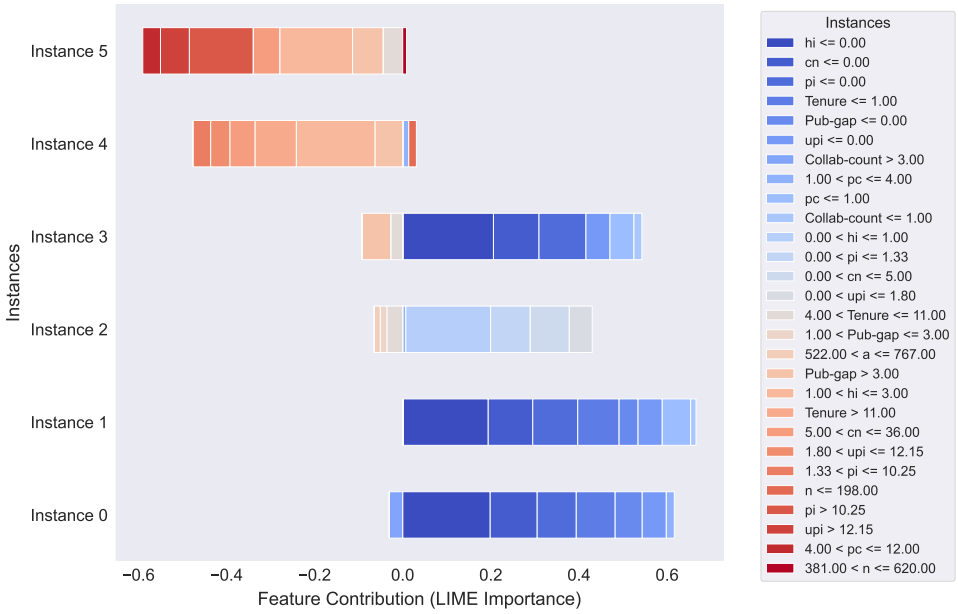


Fig. 12: LIME-based feature importance analysis for multiple test instances. Each horizontal stacked bar represents an instance, with features contributing positively (red) or negatively (blue) to the classification decision. The width of each segment indicates the feature's relative importance. The legend maps feature names to their contributions across instances.

5 Case study: OpenAlex CS Data

Building on the analyses presented in the previous sections, we note that the AMiner based study is limited in temporal scope. To address this concern, we conducted an additional case study using the continuously updated OpenAlex dataset, which is not restricted to a static database ending in 2014 [44–46]. As summarized in Supplementary Section S1.1 and Table S1, the OpenAlex sample contains 500 active computer science authors with at least two computer-science works, of whom 36.2% are labeled as early-career researchers under the seven-year definition. The sample is also heterogeneous rather than narrowly junior, with a median career age of 11 years, together with medians of 12.5 computer-science works, 62 total computer science citations, 26 unique collaborators, topic breadth across 3 subfields, and affiliation records from 3 institutions.

We use OpenAlex because it is openly accessible (supporting reproducibility and data sharing) and is continuously updated with recent scholarly records [44–46]. In contrast, Scopus requires subscription access and imposes redistribution constraints that can limit full replication. The present pipeline is database agnostic (it operates on author/work/institution trajectories) and can be executed on a current OpenAlex

Table 13: Summary statistics of the OpenAlex author sample

Statistic	Value
Number of authors	500
Share labeled as ECR under the 7-year definition	0.362
Median computer-science career age	11.0
Median computer-science work count	12.5
Median total computer-science citations	62.0
Median unique collaborator count	26.0
Median topic breadth	3.0
Median affiliation count	3.0

Table 14: Cross-validated performance for the seven-year ECR definition (kept for reviewer response)

Feature family	Model	AUROC	Average precision	Balanced accuracy	F1
Baseline	Logistic regression	0.815	0.814	0.695	0.628
Citation-only	Extra trees	0.865	0.748	0.761	0.696
Combined	XGBoost	0.992	0.987	0.939	0.926
Non-citation-only	Linear SVM	0.989	0.982	0.944	0.924

snapshot. Therefore, the empirical results are not tied to a static dataset ending in 2014; rather, they reflect what is observable in the chosen snapshot and can be refreshed as OpenAlex evolves.

Could the definition of ECR have changed over time? To address potential temporal shifts in what constitutes "early career," we evaluate multiple career age windows (5, 7, and 10 years). The model discrimination remains high and stable, indicating that conclusions do not depend on a single historical cutoff.

Table 13 summarizes the OpenAlex author sample used in our experiments. Figure 13 and Table 14 show that non-citation trajectory signals provide substantial predictive information beyond citation only indicators. Table 15 shows stability across alternative ECR windows.

The stability of the results across the five, seven, and ten year career windows further indicates that the classification behavior is not dependent exclusively on the seven-year operationalization. The seven-year window is therefore treated as one reproducible reference setting rather than as a universal boundary between junior and senior researchers.

6 Conclusions

This paper offers a machine learning-based methodology that uses both citation based and non-citation based variables taken from the AMiner dataset to identify Early Career stage Researchers (ECRs). On the sampled model development

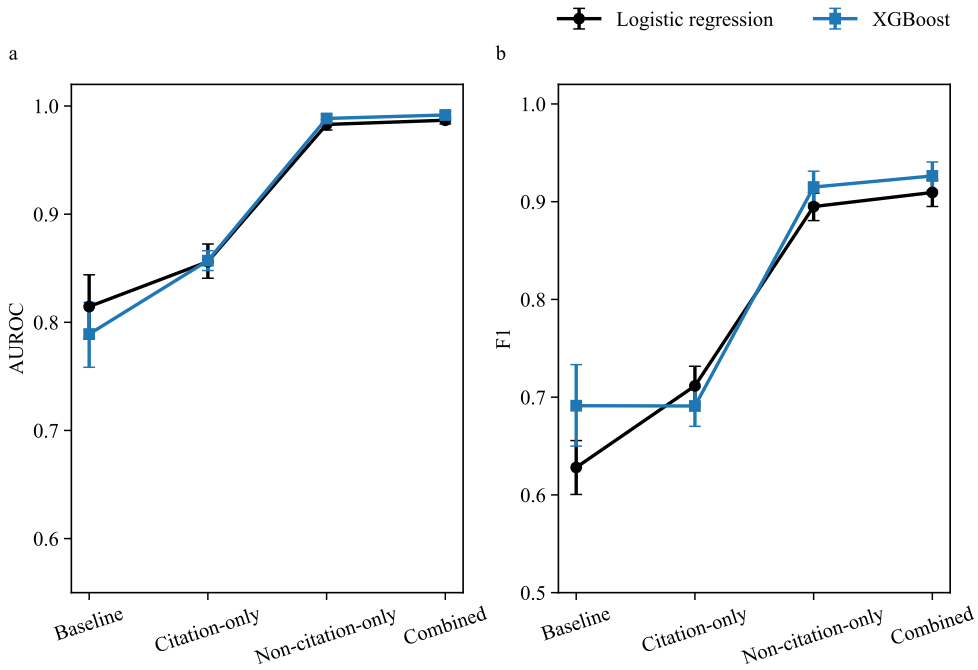


Fig. 13: Ablation across feature families (kept for reviewer response). Non-citation trajectory features dominate citation-only indicators under the seven-year definition.

Table 15: Robustness of the combined model across ECR windows (kept for reviewer response)

Window	AUROC	Average precision	Balanced accuracy	F1	ECR share
5-year	0.986	0.973	0.930	0.904	0.298
7-year	0.992	0.987	0.939	0.926	0.362
10-year	0.987	0.987	0.943	0.940	0.470

dataset, XGBoost and Gradient Boosting achieved the highest classification accuracy of approximately 99%. This result reflects performance under the constructed class-balanced sampling scheme and should not be interpreted as an estimate of ECR prevalence or unadjusted population-level performance. Explainable AI techniques, such as SHAP and LIME, provided insights into the significance of different features, reinforcing the importance of non-citation based attributes in ECR identification. The findings contribute to improving collaboration recommendations, research funding allocation, and policymaking for early career researchers. However, this study has certain limitations. A further limitation concerns the interpretation of ECR status and the class distribution used for supervised learning. The present framework

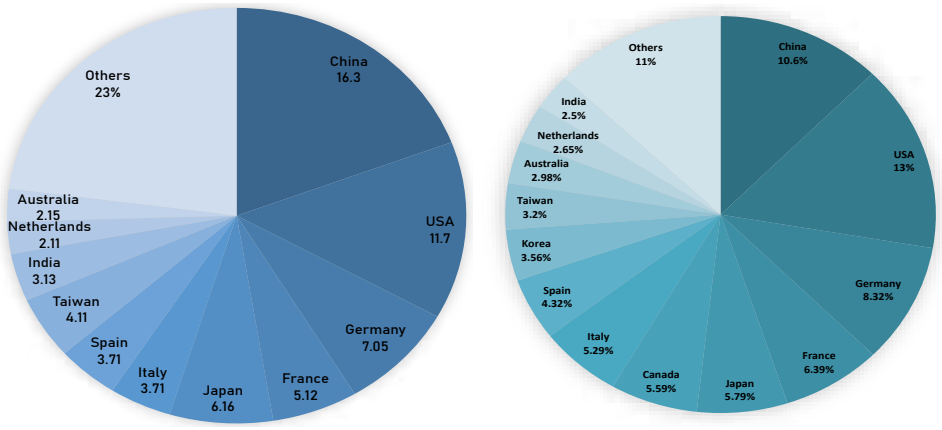
uses bibliographically observable characteristics and an operational career stage criterion; consequently, it does not directly measure research autonomy, independent grant acquisition, principal investigator responsibility, project leadership, student supervision, or other institutional indicators of research maturity. Researchers may also progress through these stages at different rates, meaning that chronological career age may not perfectly correspond to research seniority. Furthermore, because the original AMiner derived population is substantially imbalanced toward ECRs, statistical sampling was used to reduce class imbalance for model development. The resulting classification performance should therefore be interpreted as performance on the constructed model development dataset rather than as a direct estimate of classification performance under the natural population prevalence. Future work should incorporate autonomy- and leadership related indicators, where available, and evaluate the framework on naturally distributed and more diverse bibliographic datasets. It focuses only on the field of Computer Science and Engineering, making its applicability to other disciplines uncertain. However, the analysis is restricted to the AMiner dataset, which may not fully capture global research trends. In the future direction of the research should explore other domains and expand the dataset to include bibliographic sources like Scopus and Web of Science can provide a more comprehensive understanding of ECR characteristics. Furthermore, deep learning techniques, such as graph neural networks and transformers, could enhance classification performance by capturing complex relationships within research networks.We therefore emphasize that the broad developmental distinction between ECRs and established researchers is not itself claimed here as a novel descriptive finding; rather, the contribution of this study is to show how that distinction can be translated into a scalable and empirically validated bibliometric identification framework.

A Appendix: Descriptive Analysis of Dataset

A.1 Geographical Distribution of Authors and Publications

In figure 14(a), we show how scholars are geographically distributed, emphasizing the percentage of authors from various nations. The domination of a small number of nations, including the US, China, Germany, Japan, France, Korea, Taiwan, Canada, Spain, Italy, and India, which together account for a sizable portion of researchers, is a notable finding.However, a number of regions, particularly those with developing economies, are underrepresented and make little contribution to the international scientific community. The observation highlights the structural barriers that researchers in underrepresented areas must overcome, such as limited access to funding for their work, mentorship, and opportunities for collaboration.

Figure 14(b) illustrates the distribution of publications among nations, which provides insights about global research productivity. The key findings are as follows: (a) Nations with a lower number of researchers, as is evident in Figure 14(a), are also likely to have lower publication output; (b) regions with a larger number of researchers are not necessarily associated with proportionally higher publication output; and (c) a small number of countries, namely the United States, China, and a number of European nations, dominate the overall research output.



(a) Author distrib. across Country (b) Author's Publication distrib. across Country

Fig. 14: Figure 14(a) displays pie diagram that shows the proportion of publication writers belonging to each nation; the nations that contribute $\leq 2\%$ of the total share are labeled as "others." Figure 14(b) shows the percentage of publications by writers from each nation in a pie chart. The nations that contribute $\leq 2\%$ of the pie chart's total contribution are indicated as "others" in this figure.

For example, some countries that have a relatively large number of researchers exhibit comparatively lower publication output. This is an interesting observation which emphasizes the importance of institutional support, funding availability and research infrastructure towards scientific productivity.Additionally, certain high researcher regions may reflect structural challenges, such as limited collaboration networks, restricted access to publication resources, or higher rejection rates.

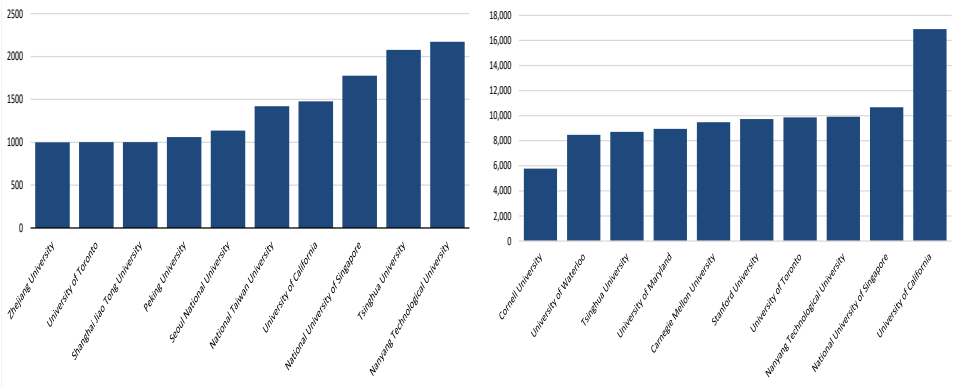
A.2 Institutional Distribution

Figure 15(a) presents the top 10 institutions ranked by the number of researchers, highlighting the concentration of authors within leading academic organizations. Similarly, Figure 15(b) shows the top 10 institutions based on publication output.

The results indicate a strong relationship between the number of researchers and publication productivity across these institutions. However, variations among institutions suggest differences in research support systems, funding availability, and academic infrastructure. These findings emphasize the importance of institutional policies and resources in enhancing research productivity and supporting early career researchers.

Declarations

- **Funding**: No funds, grants, or other support were received to conduct this study.
- **Conflict of interest/Competing interests**: On behalf of all authors, the corresponding author states that there is no conflict of interest.



(a) Author distrib. across Top Institute

(b) Author's Publication distrib. across Top Institute

Fig. 15: Figure 15(a) displays the top 10 institutes in a horizontal bar chart ranked by the number of authors taken from our dataset; the institute name is marked along the y-axis, and the number of authors is marked along the x-axis. Figure 15(b) displays the top 10 institutes in a horizontal bar chart on the basis of publications number by the authors, which was taken from our dataset. The institute name is displayed along the y-axis of the chart, while the number of articles the authors have contributed is displayed along the x-axis.

- **Data availability**: Publicly available.
- **Author contribution**: **Avijit Gayen:** Conceptualization, Methodology, Experiment, Writing. **Subham Singh:** Data cleaning. annotation, Experiment, **Shubhangi Agrawal:** Data cleaning. annotation, Experiment. **Somyajit Chakraborty:** Data collection, Experiment. **Angshuman Jana:** Supervision, Writing review & editing.

References

[1] Horta, H.: Trust and incentives in academic research and the position of universities within innovation systems. Higher Education **84**(6), 1343–1363 (2022)

[2] Schofer, E., Ramirez, F.O., Meyer, J.W.: The effects of science on national economic development, 1970 to 1990. American Sociological Review **65**(6), 866–887 (2000)

[3] Abbasi, A., Altmann, J., Hwang, J.: Evaluating scholars based on their academic collaboration activities: two indices, the rc-index and the cc-index, for quantifying collaboration activities of researchers and scientific communities. Scientometrics **83**(1), 1–13 (2010)

[4] Abramo, G., D'Angelo, C.A., Di Costa, F.: Research collaboration and productivity: is there correlation? Higher education **57**(2), 155–171 (2009)

[5] Lee, S., Bozeman, B.: The impact of research collaboration on scientific productivity. Social studies of science **35**(5), 673–702 (2005)

[6] Lopes, G.R., Moro, M.M., Wives, L.K., De Oliveira, J.P.M.: Collaboration recommendation on academic social networks. In: International Conference on Conceptual Modeling, pp. 190–199 (2010). Springer

[7] Hara, N., Solomon, P., Kim, S.-L., Sonnenwald, D.H.: An emerging view of scientific collaboration: Scientists' perspectives on collaboration and factors that impact collaboration. Journal of the American Society for Information science and Technology **54**(10), 952–965 (2003)

[8] Sampson, V., Clark, D.: The impact of collaboration on the outcomes of scientific argumentation. Science education **93**(3), 448–484 (2009)

[9] Murray, D.L., Morris, D., Lavoie, C., Leavitt, P.R., MacIsaac, H., Masson, M.E., Villard, M.-A.: Bias in research grant evaluation has dire consequences for small universities. PloS one **11**(6), 0155876 (2016)

[10] Browning, L., Thompson, K., Dawson, D.: From early career researcher to research leader: Survival of the fittest? Journal of Higher Education Policy and Management **39**(4), 361–377 (2017)

[11] Wang, W., Xu, B., Liu, J., Cui, Z., Yu, S., Kong, X., Xia, F.: Csteller: forecasting scientific collaboration sustainability based on extreme gradient boosting. World Wide Web **22**(6), 2749–2770 (2019)

[12] Gayen, A., Mukherjea, T., Jana, A.: Characterization of 'early career stage'researchers from their co-authorship network. In: 2023 IEEE Guwahati Subsection Conference (GCON), pp. 1–6 (2023). IEEE

[13] Nicholas, D., Watkinson, A., Boukacem-Zeghmouri, C., Rodríguez-Bravo, B., Xu, J., Abrizah, A., Świgoń, M., Clark, D., Herman, E.: So, are early career researchers the harbingers of change? Learned Publishing **32**(3), 237–247 (2019)

[14] Bridle, H., Vrieling, A., Cardillo, M., Araya, Y., Hinojosa, L.: Preparing for an interdisciplinary future: A perspective from early-career researchers. Futures **53**, 22–32 (2013)

[15] Laudel, G., Bielick, J.: How do field-specific research practices affect mobility decisions of early career researchers? Research Policy **48**(9), 103800 (2019)

[16] Djerasimovic, S., Villani, M.: Constructing academic identity in the european higher education space: Experiences of early career educational researchers. European Educational Research Journal **19**(3), 247–268 (2020)

[17] Kwiek, M.: Global trends in academic careers: A comparative perspective on early

career researchers. Higher Education **84**(3), 607–630 (2022) https://doi.org/10.1007/s10734-022-00954-1

[18] Teichler, U.: Higher education and the world of work: Conceptual frameworks and comparative perspectives. Higher Education **58**(5), 693–708 (2009) https://doi.org/10.1007/s10734-009-9212-4

[19] Merton, R.K.: The Sociology of Science: Theoretical and Empirical Investigations. University of Chicago Press, Chicago (1973)

[20] Cole, J.R., Cole, S.: Social stratification in science. American Journal of Sociology **79**(6), 1110–1141 (1973) https://doi.org/10.1086/225678

[21] Merton, R.K.: The matthew effect in science. Science **159**(3810), 56–63 (1968) https://doi.org/10.1126/science.159.3810.56

[22] McDowell, G.S., Knutsen, J.D., Graham, J.M.: Ghost authorship in the scientific literature: A review. Science and Engineering Ethics **12**(2), 343–355 (2006) https://doi.org/10.1007/s11948-006-0039-6

[23] Hemmings, B.: Doctoral students' perspectives on their academic identity development. Higher Education Research & Development **31**(2), 141–153 (2012) https://doi.org/10.1080/07294360.2011.559197

[24] Levecque, K., Anseel, F., De Beuckelaer, A., Heyden, J., Gisle, L.: Work organization and mental health problems in phd students. Research Policy **46**(4), 868–879 (2017) https://doi.org/10.1016/j.respol.2017.02.008

[25] Nature Editorial: The mental health of researchers demands urgent attention. Nature Human Behaviour **6**, 1037–1039 (2022) https://doi.org/10.1038/s41562-022-01351-5

[26] Momeni, F., Mayr, P., Dietze, S.: Investigating the contribution of author- and publication-specific features to scholars' h-index prediction. EPJ Data Science **12**(1) (2023) https://doi.org/10.1140/epjds/s13688-023-00421-6

[27] Zhu, X., Turney, P., Lemire, D., Vellino, A.: Measuring academic influence: Not all citations are equal. Journal of the Association for Information Science and Technology **66**(2), 408–427 (2015) https://doi.org/10.1002/asi.23179

[28] Frandsen, T.F., Nicolaisen, J.: Defining the early career researcher: A study of publication-based definitions. Learned Publishing **37**(4), 1621 (2024) https://doi.org/10.1002/leap.1621

[29] Maer-Matei, M.M., Mocanu, C., Zamfir, A.-M., Georgescu, T.M.: Skill needs for early career researchers—a text mining approach. Sustainability **11**(10), 2789 (2019)

[30] Barkhuizen, G.: Identity dilemmas of a teacher (educator) researcher: Teacher research versus academic institutional research. Educational Action Research **29**(3), 358–377 (2021)

[31] Castelló, M., McAlpine, L., Sala-Bubaré, A., Inouye, K., Skakni, I.: What perspectives underlie 'researcher identity'? a review of two decades of empirical studies. Higher Education **81**, 567–590 (2021)

[32] Horta, H., Cattaneo, M., Meoli, M.: Phd funding as a determinant of phd and career research performance. Studies in Higher Education **43**(3), 542–570 (2018)

[33] Andrews, E.J., Harper, S., Cashion, T., Palacios-Abrantes, J., Blythe, J., Daly, J., Eger, S., Hoover, C., Talloni-Alvarez, N., Teh, L., *et al.*: Supporting early career researchers: insights from interdisciplinary marine scientists. ICES Journal of Marine Science **77**(2), 476–485 (2020)

[34] Castelló, M., Kobayashi, S., McGinn, M.K., Pechar, H., Vekkaila, J., Wisker, G.: Researcher identity in transition: Signals to identify and manage spheres of activity in a risk-career. Frontline Learning Research **3**(3), 39–54 (2015)

[35] Skakni, I., Calatrava Moreno, M.d.C., Seuba, M.C., McAlpine, L.: Hanging tough: post-phd researchers dealing with career uncertainty. Higher Education Research & Development **38**(7), 1489–1503 (2019)

[36] McDowell, G.S., Knutsen, J.D., Graham, J.M., Oelker, S.K., Lijek, R.S.: Co-reviewing and ghostwriting by early-career researchers in the peer review of manuscripts. Elife **8**, 48425 (2019)

[37] Drosou, N., Del Pinto, M., Al-Shuwaili, M.A., Goodall, S., Marlow, E.: Overcoming fears: a pathway to publishing for early career researchers. Disaster Prevention and Management: An International Journal **29**(3), 340–351 (2020)

[38] Kong, X., Jiang, H., Bekele, T.M., Wang, W., Xu, Z.: Random walk-based beneficial collaborators recommendation exploiting dynamic research interests and academic influence. In: Proceedings of the 26th International Conference on World Wide Web Companion, pp. 1371–1377 (2017)

[39] Team, A.: Arnetminer. https://www.kaggle.com/kmader/aminer-academic-citation-dataset. Accessed 20 February 2021 (2021)

[40] Abrishami, A., Aliakbary, S.: Predicting citation counts based on deep neural network learning techniques. Journal of Informetrics **13**(2), 485–499 (2019) https://doi.org/10.1016/j.joi.2019.02.011

[41] Gayen, A., Chakraborty, S., Mitra, S., Jana, A.: Comeback or dropout: study of discontinued researchers at early career stage. Scientometrics **130**(2), 1201–1236 (2025)

[42] Schriever, V., Grainger, P.: Mentoring an early career researcher: insider perspectives from the mentee and mentor. Reflective Practice **20**(6), 720–731 (2019)

[43] Luque, A., Carrasco, A., Martín, A., Las Heras, A.: The impact of class imbalance in classification performance metrics based on the binary confusion matrix. Pattern Recognition **91**, 216–231 (2019)

[44] Priem, J., Piwowar, H., Orr, R.: OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts (2022). https://arxiv.org/abs/2205.01833

[45] Alperin, J.P., Portenoy, J., Demes, K., Larivière, V., Haustein, S.: An analysis of the suitability of OpenAlex for bibliometric analyses (2024). https://doi.org/10.48550/arXiv.2404.17663 . https://arxiv.org/abs/2404.17663

[46] Culbert, J., Hobert, A., Jahn, N., Haupka, N., Schmidt, M., Donner, P., Mayr, P.: Reference Coverage Analysis of OpenAlex compared to Web of Science and Scopus (2024). https://doi.org/10.48550/arXiv.2401.16359 . https://arxiv.org/abs/2401.16359