Identification of ‘Early Career Stage’ Researchers from large scale Bibliographic dataset

Avijit Gayen[1*], Tamoghna Mukherjea[2], Soumik Mukherjee[2], and Angshuman Jana[1]

[1]Indian Institute of Information Technology Guwahati, India
[1]{avijit.gayen, angshuman}@iiitg.ac.in,
[2*]Techno India University, Kolkata, India
[2]tamoghnamukherjea4@gmail.com

***Abstract*—In the realm of scientific research, collaboration plays a pivotal role. Identifying and supporting Early Career-stage Researchers (ECRs) is crucial for fostering their career development and ensuring the growth of scientific knowledge in their respective fields. In this paper, we present a machine learning-based approach to identify ECRs from large-scale bibliographic datasets. Our proposed framework leverages various features extracted from the dataset, including publication records, citation patterns, academic affiliations, co-/authorship networks, etc. To train and evaluate our model, we utilize a comprehensive and diverse AMiner Author dataset. Further, we have divided the 5000 sample data points into a 7:3 ratio for training and testing purposes. However, with the application of machine learning techniques, we achieve high accuracy in identifying ECRs, effectively distinguishing them from established researchers (‘h-index’ distributions are used to identify the established researchers).**

***Index Terms*—Collaboration network, Co-authorship network, Early career-stage, Collaboration recommendation.**

I. Introduction

Scientific research and development takes forward the advancement of any nation which implicitly drives the economic growth of the nation [1]. In the current trend of research, not only in the science and technology domain but also in other domains, collaborative research work is highly popular. It can effectively solve the complex problem by joint endeavour of research collaborators. It enhances the research quality [2], generates a greater amount of knowledge which further led to innovation and productivity [3]. The steady growth in the need for academic and scientific collaborations has been observed in the last few decades [4]. The growth of research and development has led to the creation of large sizes of exponentially increasing bibliographic metadata. This bibliographic data in the form of electronic transcription has given a lot of opportunity to the researcher to explore and understand nature of scientific growth and its dynamics which are significantly useful in research policy making.

In recent time, early career-stage researchers take a prime role in collaborative research work [5]. ECR are those researchers who are generally not older than 35. They either have received their doctorate and are currently in a post doctorate research position. ECR also includes those researchers who have been in research position but are currently doing a doctorate [6]. It has been observed that though ECR have comparatively less experience and a limited amount of knowledge, their utmost level of enthusiasm and curiosity toward the new things takes the research to a comparatively higher level. It is their capability to cope with the challenges of research and their diligence, perseverance help to resolve the complex research problem. Though the guidance of experienced researchers towards those ECRs can not be neglected, ECRs constant effort to solve the complex problem is always appreciable [7]. Thus ECRs plays a major role in change of scholarly which in turn promotes research and development of the nation.

Though the need of ECR in collaborative research work is inevitable, finding a effective collaboration to cope up with a potential research problem is quite challenging [8]. This task become more complex due to ever-increasing bibliographic information as well. Although it is a generic problem to any researcher irrespective to their career age, ECR in such a situation may find it more difficult issue to choose their ideal academic collaborator due to the lack of proper domain knowledge as well as domain experts. This situation leads to inappropriate collaborations [9], [10] or missing out on a valuable collaboration, resulting in an overall lack of good collaborations at the early stage of their research career. Beside this major challenge, ECR also face some very serious problems which affect their career growth. It has been observed that due to lack of attention towards ECR, ECR face strong competition with well-established researchers or group of researchers, to acheive funding for their projects [11], they are dominated by recognized researchers in terms of recognition of their works on identified problems [12].

In literature, we find the lack of proper definition of ECR, which might lead to improper identification ECR from large bibliographic dataset. This situation promotes the less effective collaborations recommendations for those ECR, and it may result in some improper collaborations. People already observed that those improper collaborations are relatively less sustainable [13]. On the other side it also leads to the improper allocation of funding resulting in ill-impact in the growth of the development and research. To find the ECR with help of existing methodology in literature is quite challenging for two major reasons: a) the lack of proper definition which can segregate the ECR from the other researchers, b) it is very challenging to identify the ECR from large scale of bibliographic data, e.g., some of the researcher's *h*-index is comparatively low though they are not in their 'early career stage'. This might be due to the domain in which they work which might have fewer citations or they work in close small groups or work individually which in turn receives fewer citations. Hence, we take the cues from the observation taken by Gayen et. al [] in a detailed study of publication and collaboration patterns of researchers at their early career stage. We observed it as classification problem where we classify the researchers into ECR and non-ECR from bibliographic dataset with the help of machine learning model.

For the development of the our ML model to identify ECR from bibliographic metadata, we have used the AMiner dataset. Initially, we cleaned the missing parameters data points from the dataset. With the help of earlier collaboration and publication analysis on ECR in their initial career stage and the distinguished differences with the well-established researchers of ECR observed by Gayen et al., we have prepared the potential feature set for our classifier. It includes citation as well as non-citation based features. To create the labeled dataset, we have used three expert decision method to break the ties for proper annoation of the dataset. As the bibliographic dataset is highly populated with low profile researchers which leads to class imbalance problem in the training data. To overcome this situtaion, we have used standard statistical method of sampling to prepare a dataset of 5000 data points. Further, we have divided the 5000 sample data points into 7:3 ratio for training and testing purpose. We employ nine different supervised machine learning classifier i.e Logistic Regression, Naive Bayes, SVM, Decison tree, Gradient Boosting and XGBoost etc. to train the model. The Gradient Boosting and XGBoost in this dataset outperform the other model of learning.

II. Related Works

In this section, we identify the contributory works related to the ECR. Initially, we highlight the contributions made toward scientific collaborative research work with recent trends in collaboration developments. Further, we outline the works related with ECR and their role in research. We also highlight the problems and challenges of the ECRs in research. We further underline the challenges of identification of ECR. Finally, we point out the scope of the work mentioning the research gap in the literature.

a) **Collaborative Research Trend:** Collaborative research work in scientific community had been a usual activity since long ago. Although, it has been observed that there is a steady growth in the need for academic and scientific collaborations in the past few decades [4]. In a work [2], abbasi et al. had observed local and online collaborations increased their research output [2]. In an another work [3], Melin et al. showed that collaborations generated more knowledge which in turn led to innovation and academic productivity. The study outlines that 38% of the academic respondents mentioned "Increased Knowledge" as the major benefit they received from collaborations. In [14], authors observed that most scientific output was a result of collaborative work. It has been observed in earlier research [15] that there is an ever-increasing need for international scientific cooperation. In a recent work [16], authors observed the fast growing demand for new researchers. In an another work [17], Kong et al. also observed the growth of research collaborations due to the increase in scale of the internet.

b) **Early career stage Researcher:** The recent trend of collaborative research work reveals that the early career stage researcher take a crucial role in the research community to overcome the complex scientific challenges. In literature, we find some earlier works [6], [18], [19], where authors provide a qualitative definition of ECR.In those literature, authors has assumed that people at their Ph.D. or at post Ph.D. position are being considered as ECR. In some earlier work [18], authors observed that interdisciplinary research can enhance the development in science and technology. They also observed that early career stage researchers initiative improves the scope of interdisciplinary work. In another work [20], authors observed the several types of evolving academic identity among a group of European educational ECRs who moves across their country. It has also been observed that they relatively have strong orientation to academic values, accompanied by a spirit of agency and entrepreneurship. Thus this type of activities amongst the ECR foster the knowlege and eduaction.

c) **Challenges and Problems of ECR:** The existence of ECR is any domain of research is obvious and the need for the participation of ECR in solving critical problems is an undoubted truth. Though, ECR not only face the various challenges and issues in cross-domain research activity but also in their own area of interest. In an earlier work [21], [22], authors analyzed Early stage research positions are often include teaching activities to a great extent. Thus, ECR faces the issues regarding the time management, lack of focus in research, etc. In another work [23], authors highlight that the role of identity developed by the ECR plays a important role in their career advancement. Funding in research works plays a major role for the advancement of scientific activities. In an important work [24], Horta et al. observed that funding and recognition of ECR shapes the performance of their activities. The barriers of the ECR folded multiple times for the interdisciplinary domains e.g., ocean and coastal sustainability. In a recent work [25], authors identified major barriers of ECR in ocean and coastal sustainability and the overcome methods of those.

In some earlier works [26], authors remarked that due to the current higher education policy, ECRs face a 'risk-career' in which they may find seldom a predictable, stable academic careers. The study of an important work [27], reveals the impact of career uncertainty on post-PhD researchers' experiences. The findings of the work conclude that post-PhD researchers in the course of their work experiences, they face two kinds of career uncertainty, i.e., intellectual uncertainty and occupational uncertainty. The uncertainty in research career drives the potential good researchers towards the other kind of career option. Mcdowell et al. in their study of review methods of the standard peered-review Journals observed that there is a 'ghost' review writing process which has been done by the ECR without any recognition [28]. This scenario not only point out the compromisation of the review method which in turn bring down the quality of research but also the exploitation of ECR. In an another work [29], Drosou et al. also highlight that ECR always face challenge to publish their research report in standard journal.

The increasing trend in collaborative research work has urged the searching of potential collaboration in researcher community. Though the searching in this continuously populated bibliographic dataset in quite challenging. This situation become more difficult for the ECR due to the lack of information regarding the domain experts and complex method collaboration development for the career advancement. Though the introduction of Collaboration Recommendation system improved the situation, Most of works till now have focused on recommending most possible collaborations without focusing on recommending beneficial collaborators (BCs) and sustainable nature of collaboration. In [30], Kong et al. observed that BCs can not only collaborate with researchers but also can provide academic guidance. BCs could help them pave the way to make achievements rapidly. Another Studies have shown that a sustainable collaboration has a significant positive impact on the productivity and influence for a researcher [13]. In the plethora of the works related to collaboration recommendation, there exists very few works focus on ECRs.

The existing literature identifies that there exists very strong urge to focus towards'early-stage' researchers for the sake of development of research. Though the literature agree with the existing issues of the ECR, surprisingly that none of the work tried to give some significant solution of the problems regarding the issues of the ECR as a main focus of their work. Finding ECRs based on the earlier definition from large bibliographic dataset is quite challenging. Thus there is a strong need of formal methodology to identify ECR to focus the issues regarding them. In this paper, we propose machine learning based methodology on the electronically updated bibliographic dataset to figure out the researchers at their 'early-stage'.

III. Methodology

In this section, we initially provide a detailed description of the dataset used in this work. We further describe the data prepossessing method used to clean the dataset and the method adopted to properly annotate the dataset for supervised learning. We also discuss in detail the several citation-based and non-citation-based features used in the model for classification of the authors into ECR and non-ECR type. later, we describe various machine learning classifier used in our work to learn the bibliographic data. We also provide a schematic diagram 1 to show the work flow to execute the method.

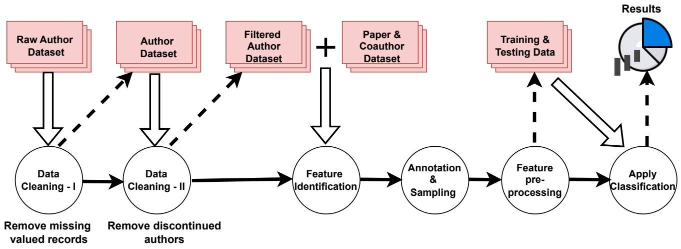

Fig. 1: Figure shows the schematic diagram of the flowchart followed in the proposed methodology.

A. Dataset

We are using the AMiner dataset [31] for our analysis. It is a free and open-source dataset with a Creative Commons license (CC BY-NC-SA 4.0) maintained by K. Scott Mader, and was last updated on the 12th of December, 2018. The dataset contains the publication's bibliographic information in the duration of 1936—2014. In our analysis, we have used the paper publication bibliographic data and detailed data regarding authors and their coauthor relationship. This information is stored in three separate files, organized into the following attributes shown in table I.

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 I: Dataset attributes and their definitions

Arnet Miner dataset	Raw	Filtered
Number of valid papers	20,92,356	20,54,704
Number of papers with no author details	37505	-
Number of papers with no publication year	147	-
Number of authors	17,12,433	17,12,431
Number of Authors without proper details	2	-
Avg. number of papers per author	-	2.887
Avg. number of authors per paper	-	3.262
Total number of distinct coauthorship relation	-	42,58,946

TABLE II: General information of raw and filtered Arnet Miner dataset(Publication tenure 1936—2014).

B. Data cleaning & Annotation

In our dataset, we find many papers do not contain any author details and in the author dataset few authors do not have complete information (some of them h-index, affiliation missing). We have removed those data from our dataset. In our dataset, we find around 90% researchers of the complete dataset after preprocessing who do not have any publications in the consecutive three years of their research career. The idea behind the identification of the gap of three consecutive years of publication by a researcher is to embark on the discontinuation of his / her research work. The discontinuation might be either due to the completion of graduation from the institute where he/she might have undergone research work or a career switch in the due course of time. We segregate them as *'discontinued'* researchers as the majority of them either do not have a further publication or have very few publications later on. We have not considered that *'discontinued'* researchers in our analysis as these researchers have the least impact in the co-authorship network. We provide the general information regarding the dataset after the preprocessing in table II. To do the unbiased annotation, we have adopted three expert judgement technique to label the data points. At the time of annotation, we find that the majority($\approx 90\%$) of researcher in the bibliographic dataset are ECR. Thus to keep the balanced two classes of the annotated dataset to be used in our classifier, we have applied statistical sampling technique to improve the class imbalance issue. In our annotated dataset, it contains 5000 data points where 57.3% are ECR and rest are non-ECR. We have classified the 5000 data points into two part— a)training data b)testing data, where 70% data have been used in classifier for training purpose and rest have been used for testing of performance.

C. Features used in Learning model

In this section, we describe the various features used for the classification and the preprocessing method applied on the several features before applying the machine learning classifier. We have used citation-based as well as non-citation-based features in our model. In table III, we list the features used and given a brief description of the each feature. Though the citation-based features available in the dataset do not require any preprocessing before feeding the data into classifier, the non-citation-based features except publication count require the preprocessing.

1) Preprocessing of features: The number of collaborators feature was estimated by identifying the distinct coauthors of each author from the his/her collaboration edges stored in "Coauthor" dataset. The Number of year gap in research publication feature has been derived by calculating the total number of years in which the author does not have any publications from the paper information dataset. On the other hand publication tenure of the researcher was measured by the difference between the publication year of his first publication and latest publication available in dataset. We mark the lowest publication year of the paper of the concerned researcher as first publication.

2) Vectorization: The 'affiliation' and 'area of interest' features were vectorized using the CountVectorizer class from the $sklearn.feature\_extraction.text$ module. An instance of CountVectorizer was created for each feature. The $fit\_transform$ method of the vectorizers was applied to convert the text data into a numerical representation. The $fit\_transform$ method learns the vocabulary and performs the vectorization. It tokenizes the text, counts the occurrences of each token, and transforms the text into a matrix representation. In summary, the 'affiliation' and 'area of interest' features are vectorized using CountVectorizer, which tokenizes the text and converts it into numerical representations by counting the occurrences of each token.

D. Machine Learning Algorithms

In our proposed method, we have used nine different supervised classifier i.e., Logistic Regression, Support Vector Machine, Random Forest, Decision Tree, K-Nearest Neighbour, Gradient Boosting, AdaBoost,XGBoost and Naive Bayes. In this section, we briefly describe the machine learning classifiers used in our proposed model and also justify the implementation of that classifier in this context.

a) ***Logistic Regression*****:** Logistic regression is specifically designed for binary classification problems, where the goal is to classify instances into one of two classes, in this case, ECR or non-ECR. Logistic regression can work well with limited amounts of data. If the author dataset is relatively small, logistic regression can still yield reliable results without requiring an extensive amount of training data. By examining the coefficients associated with each feature, we can determine which factors have the most significant impact on the probability of being an ECR.

b) ***Support Vector Machine*****:** SVM performs well even when the number of features (attributes of researcher) is greater than the number of samples. In our dataset, each author is described by multiple attributes such as name, affiliations, publication count, citations, h-index, etc. SVM can handle these high-dimensional feature spaces effectively. SVM can provide insights into feature importance, allowing you to understand which attributes contribute the most to the ECR classification.

c) ***Random Forest*****:** In the context of identifying ECR (Early Career Researchers), the Random Forest algorithm can be a good choice because it can handle a large number of input features and non-linear relationships between them. Random Forest also has the ability to reduce overfitting by randomly selecting a subset of features for each tree, making it less sensitive to noise and outliers in the data.

d) ***Decision Tree*****:** Decision Tree classifier provides a clear and interpretable set of rules for classification. The tree structure represents a series of if-else conditions based on different features, allowing us to understand the decision-making process of the classifier. Decision Trees are capable of capturing complex nonlinear relationships between features and the target variable. In author datasets, the relationship between the features e.g., publication count, citation count, h-index, etc. and target value may not follow a simple linear pattern. Decision Trees can handle such non-linearity

Features(Citation-based)	Description
citation count	In author dataset, the field "cn" represents the total number of citations received by each individual author. It measures the number of times that an author's work has been cited by other researchers in their own publications. It is the most popular and simple metric to measure the impact and influence of a researcher's work within their field.
h-index	The *h*-index is a most popular citation-based metric used to estimate the impact and productivity of a researcher's publications. In our author dataset it is attributed by ‘hi’. It takes into account both the number of publications and the number of citations received by those publications. The *h*-index is the highest number *h* for which the researcher has at least *h* publications that have received *h* or more citations.
p-index	The p-index is stands for the measure of popularity metric of the researcher in his/ her research community. The pi (p-index with equal A-index) represents the p-index value when the A-index (authorship index) is equal for all publications. The A-index refers to the average number of authors per paper. In this case, the p-index considers the author's productivity and impact assuming an equal contribution in all publications.
up-index	The "upi" (p-index with unequal A-index) represents the popularity (p-index) value when the A-index is unequal, i.e., the author has varying degrees of contribution across their publications. This variation accounts for situations where an author may have publications with different levels of contribution or authorship positions.
Features(Non-Citation-based)	Description
Publication count	The "Publication count" feature in our author dataset refers to the number of papers published by each author. It represents the total count of research papers that have been authored by a particular researcher.
Affiliation	In the author dataset, the "Affiliations" feature refers to the institutions or organizations with which the author is associated. The researcher could have multiple affiliations which are separated by semicolons in the dataset.
Area of Interest	This feature in our author dataset refers to the field or areas of research that each author is interested in or has expertise in. It provides information about the specific topics or subjects that authors focus on or have published papers related to.
Number of collaborators	It represents the total number of authors that a particular author had collaborated with across the different publications done in his/ her research career.
Number of year gap in research publication	This feature in our author dataset refers to the total number of years in his/her total research tenure in which that specific author do not have any publication.
Publication tenure	The "Publication Tenure" feature refers to the total research tenure of an author. It represents the length of time an author has been actively engaged in their academic or research career. It provides information about the author's experience and seniority in their field.

TABLE III: Features used in Machine Learning Classifiers.

effectively by creating splits based on various thresholds and combinations of features.

*e) **K-Nearest Neighbour**:* KNN can capture feature interactions that contribute to ECR classification. KNN takes into account the collective influence of multiple features. This enables the algorithm to capture synergistic effects and dependencies between different features, potentially leading to improved classification accuracy. Since KNN classifies instances based on their nearest neighbours, the relative distribution of the classes in the neighbourhood can effectively influence the classification decision. As we sufficiently improve the class imbalance problem at the time of dataset preparation, we apply KNN in our work.

*f) **Gradient Boosting**:* In ECR identification, there might be an imbalance between the number of ECR and non-ECR instances in the dataset. It can be handled well by Gradient Boosting. This classifier is known for its ability to capture complex nonlinear relationships between features and the target variable.

*g) **AdaBoost**:* AdaBoost provides a measure of feature importance, allowing you to understand which attributes in our author dataset contribute most to identifying ECRs. This analysis helps to gain the insights into the relevant features and their impact on the ECR classification process. Iterative adjustment of weights on misclassified instances, it can learn and focus on those instances which have more probable to misclassification, improving the classifier's overall performance.

*h) **XGBoost**:* XGBoost offers high predictive performance and handles imbalanced data well. It can capture complex patterns, has feature importance analysis, and handles large datasets efficiently. XGBoost's regularization techniques prevent overfitting, and it can handle missing values and outliers. Its scalability and efficiency make it a reliable choice for ECR classification.

*i) **Naive Bayes**:* Naive Bayes is computationally fast and requires fewer training data compared to more complex models. This can be advantageous when working with large datasets, e.g., Aminer author dataset. It provides interpretability by estimating the probability of a sample belonging to a particular class based on its feature values. This can offer insights into the factors contributing to the classification of ECRs, helping understand the decision-making process.

IV. RESULTS & DISCUSSIONS

A. Performance Metrics

The different metric has been used to measure the performance of this work as follows:

*a) **Confusion Matrix**:* The performance of the machine learning classifier for the identification of ECR researchers from the bibliographic dataset can be evaluated using confusion matix as given below [32]:

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

Where,

1) **TN=** Total number of Non-ECR correctly classified as Non-ECR.
2) **TP**=Total number of ECR correctly classified as ECR.
3) **FP**=Total number of ECR misclassified as Non-ECR.
4) **FN**=Total number of Non-ECR misclassified as ECR.

b) Accuracy: Accuracy is a commonly used evaluation metric for classification algorithms. It measures the percentage of correct predictions made by the algorithm on a given dataset. Mathematically, accuracy is calculated by dividing the number of correct predictions by the total number of predictions:

$$Accuracy = \frac{TN + TP}{TP + FN + FP + TN} \tag{1}$$

c) Precision: Precision is a performance metric that evaluates the performance of a classification algorithm, specifically for positive class predictions. Mathematically, precision is calculated as:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

d) Recall: Recall is a performance metric that measures the ability of the model to correctly identify positive instances or the proportion of actual positive instances that are correctly identified by the algorithm. The formula to calculate recall is as follows:

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

e) F1-score: The F1-score is a metric commonly used in classification tasks to measure the performance of an algorithm. It combines precision and recall into a single value to provide a balanced assessment of the algorithm's effectiveness. The F1-score is the harmonic mean of precision and recall, calculated using the following formula:

$$F_1 = \frac{2 \times (Precision \times Recall)}{Precision + Recall} \tag{4}$$

f) ROC: The Receiver Operating Characteristic (ROC) curve is a graphical representation of the performance of a binary classification algorithm. It illustrates the trade-off between the true positive rate (TPR) and the false positive rate (FPR) for different classification thresholds.

$$TPR = \frac{TP}{TP + FN} \tag{5}$$

$$FPR = \frac{TN}{TN + FP} \tag{6}$$

g) AUC: In machine learning, the AUC (Area Under the Curve) is a metric used to evaluate the performance of binary classification models, typically in the context of a Receiver Operating Characteristic (ROC) curve.

B. Performance of Machine Learning Algorithms

In this section, we highlight the performance of the machine learning classifier applied in our work and also compare the performce metric of the classifier. In table IV, we show the comparison the ten classifier in terms of five performance metrices. The figure 2 shows the comparions of the ROC of the nine different classifier used in our work. Though the result shows in table IV shows that Gradient Boosting and XGBoost outperform the other classifier, ROC implies decision tree and random forest gives the comparatively better result.

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 IV: Performance of Machine Learning Model applied for classification.

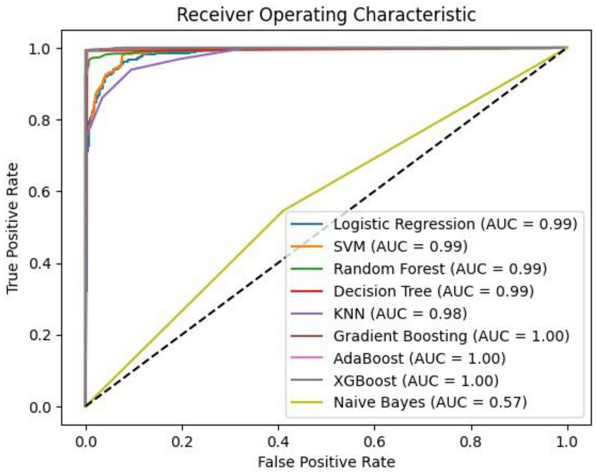

Fig. 2: Figure shows comparison among ROC of several classifiers used in our work.

V. Conclusion and Future Work

Collaboration among researchers always exerts a crucial influence on scientific research endeavors. However, it has been observed that the formation of new collaborations predominantly occurs among established researchers, as evident from the h-index distributions of the authors, which serve as a defining criterion for established researchers. This paper focuses on the identification of ECRs from a large-scale bibliographic dataset using machine learning techniques. The study begins by recognizing the importance of distinguishing ECRs due to their unique needs and potential for growth during their initial ten years, juxtaposed with those of established researchers during their early career stages. We performed an extensive analysis of the dataset to identify several key features and characteristics that are indicative of researchers in their early career stages. Our experimental results are conducted on the AMiner Author dataset using machine learning algorithms when trained on these distinguishing features, can accurately classify ECRs with a high level of precision. This work provides a solid foundation for ECR identification using machine learning, there are still avenues for future exploration. Possible directions include refining the feature selection process, evaluating the performance of different machine learning algorithms, and extending the research to incorporate additional datasets from different domains or regions.

References

[1] "MS Windows 10 kernel description," https://www.imf.org/en/Blogs/Articles/2021/10/06/blog-ch3-weo-why-basic-science-matters-for-economic-growth, accessed: 2023-05-23.

[2] A. Abbasi, J. Altmann, and J. Hwang, "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*, vol. 83, no. 1, pp. 1–13, 2010.

[3] G. Melin, "Pragmatism and self-organization: Research collaboration on the individual level," *Research policy*, vol. 29, no. 1, pp. 31–40, 2000.

[4] C. S. Wagner, H. W. Park, and L. Leydesdorff, "The continuing growth of global cooperation networks in research: A conundrum for national governments," *PloS one*, vol. 10, no. 7, p. e0131816, 2015.

[5] D. Nicholas, A. Watkinson, C. Boukacem-Zeghmouri, B. Rodríguez-Bravo, J. Xu, A. Abrizah, M. Świgoń, and E. Herman, "Early career researchers: Scholarly behaviour and the prospect of change," *Learned Publishing*, vol. 30, no. 2, pp. 157–166, 2017.

[6] D. Nicholas, A. Watkinson, C. Boukacem-Zeghmouri, B. Rodríguez-Bravo, J. Xu, A. Abrizah, M. Świgoń, D. Clark, and E. Herman, "So, are early career researchers the harbingers of change?" *Learned Publishing*, vol. 32, no. 3, pp. 237–247, 2019.

[7] M. C. Evans and C. Cvitanovic, "An introduction to achieving policy impact for early career researchers," *Palgrave Communications*, vol. 4, no. 1, pp. 1–12, 2018.

[8] G. R. Lopes, M. M. Moro, L. K. Wives, and J. P. M. De Oliveira, "Collaboration recommendation on academic social networks," in *International conference on conceptual modeling*. Springer, 2010, pp. 190–199.

[9] N. Hara, P. Solomon, S.-L. Kim, and D. H. Sonnenwald, "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*, vol. 54, no. 10, pp. 952–965, 2003.

[10] V. Sampson and D. Clark, "The impact of collaboration on the outcomes of scientific argumentation," *Science education*, vol. 93, no. 3, pp. 448–484, 2009.

[11] D. L. Murray, D. Morris, C. Lavoie, P. R. Leavitt, H. MacIsaac, M. E. Masson, and M.-A. Villard, "Bias in research grant evaluation has dire consequences for small universities," *PloS one*, vol. 11, no. 6, p. e0155876, 2016.

[12] L. Browning, K. Thompson, and D. Dawson, "From early career researcher to research leader: Survival of the fittest?" *Journal of Higher Education Policy and Management*, vol. 39, no. 4, pp. 361–377, 2017.

[13] W. Wang, B. Xu, J. Liu, Z. Cui, S. Yu, X. Kong, and F. Xia, "Csteller: forecasting scientific collaboration sustainability based on extreme gradient boosting," *World Wide Web*, vol. 22, no. 6, pp. 2749–2770, 2019.

[14] C. L. Borgman and J. Furner, "Scholarly communication and bibliometrics," *Annual review of information science and technology*, vol. 36, no. 1, pp. 2–72, 2002.

[15] M. Leclerc and J. Gagné, "International scientific cooperation: The continentalization of science," *Scientometrics*, vol. 31, no. 3, pp. 261–292, 1994.

[16] F. G. Montoya, A. Alcayde, R. Baños, and F. Manzano-Agugliaro, "A fast method for identifying worldwide scientific collaborations using the scopus database," *Telematics and Informatics*, vol. 35, no. 1, pp. 168–185, 2018.

[17] X. Kong, H. Jiang, Z. Yang, Z. Xu, F. Xia, and A. Tolba, "Exploiting publication contents and collaboration networks for collaborator recommendation," *PloS one*, vol. 11, no. 2, p. e0148492, 2016.

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

[19] G. Laudel and J. Bielick, "How do field-specific research practices affect mobility decisions of early career researchers?" *Research Policy*, vol. 48, no. 9, p. 103800, 2019.

[20] S. Djerasimovic and M. Villani, "Constructing academic identity in the european higher education space: Experiences of early career educational researchers," *European Educational Research Journal*, vol. 19, no. 3, pp. 247–268, 2020.

[21] M. M. Maer-Matei, C. Mocanu, A.-M. Zamfir, and T. M. Georgescu, "Skill needs for early career researchers—a text mining approach," *Sustainability*, vol. 11, no. 10, p. 2789, 2019.

[22] G. Barkhuizen, "Identity dilemmas of a teacher (educator) researcher: Teacher research versus academic institutional research," *Educational Action Research*, vol. 29, no. 3, pp. 358–377, 2021.

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

[24] H. Horta, M. Cattaneo, and M. Meoli, "Phd funding as a determinant of phd and career research performance," *Studies in Higher Education*, vol. 43, no. 3, pp. 542–570, 2018.

[25] E. J. Andrews, S. Harper, T. Cashion, J. Palacios-Abrantes, J. Blythe, J. Daly, S. Eger, C. Hoover, N. Talloni-Alvarez, L. Teh *et al.*, "Supporting early career researchers: insights from interdisciplinary marine scientists," *ICES Journal of Marine Science*, vol. 77, no. 2, pp. 476–485, 2020.

[26] M. Castelló, S. Kobayashi, M. K. McGinn, H. Pechar, J. Vekkaila, and G. Wisker, "Researcher identity in transition: Signals to identify and manage spheres of activity in a risk-career." *Frontline Learning Research*, vol. 3, no. 3, pp. 39–54, 2015.

[27] I. Skakni, M. d. C. Calatrava Moreno, M. C. Seuba, and L. McAlpine, "Hanging tough: post-phd researchers dealing with career uncertainty," *Higher Education Research & Development*, vol. 38, no. 7, pp. 1489–1503, 2019.

[28] G. S. McDowell, J. D. Knutsen, J. M. Graham, S. K. Oelker, and R. S. Lijek, "Co-reviewing and ghostwriting by early-career researchers in the peer review of manuscripts," *Elife*, vol. 8, p. e48425, 2019.

[29] N. Drosou, M. Del Pinto, M. A. Al-Shuwaili, S. Goodall, and E. Marlow, "Overcoming fears: a pathway to publishing for early career researchers," *Disaster Prevention and Management: An International Journal*, vol. 29, no. 3, pp. 340–351, 2020.

[30] X. Kong, H. Jiang, T. M. Bekele, W. Wang, and Z. Xu, "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*, 2017, pp. 1371–1377.

[31] "Arnetminer," https://www.kaggle.com/kmader/aminer-academic-citation-dataset, accessed 20th February, 2021.

[32] A. Luque, A. Carrasco, A. Martín, and A. de Las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," *Pattern Recognition*, vol. 91, pp. 216–231, 2019.