Star Scientists' Prediction in the Field of Artificial Intelligence Using Machine Learning Techniques

Koosha Shirouyeh

A Thesis
In the Department of
Mechanical, Industrial, and Aerospace Engineering

Presented in Partial Fulfillment of the Requirements
For the Degree of
Master of Applied Science (Industrial Engineering)

At Concordia University
Montreal, Quebec, Canada

April 2023

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: Koosha Shirouyeh

Entitled: Star Scientists' Prediction in the Field of Artificial Intelligence Using Machine Learning Techniques

and submitted in partial fulfillment of the requirements for the degree of

Master of Applied Science (Industrial Engineering)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final Examining Committee:

_____ Chair

Dr. Kudret Demirli

_____ External Examiner

Dr. Hossein Hashemi Doulabi

_____ Examiner

Dr. Kudret Demirli

_____ Thesis Supervisor

Dr. Andrea Schiffauerova

_____ Thesis Co-Supervisor

Dr. Ashkan Ebadi

Approved by

_____ 2023

Martin D. Pugh, Chair

Department of Mechanical, Industrial and Aerospace Engineering

_____

Dr. Mourad Debbabi, Dean

Faculty of Engineering and Computer Science

Abstract

Star Scientists' Prediction in the Field of Artificial Intelligence Using Machine Learning Techniques

Koosha Shirouyeh

Star scientists are highly influential researchers who have made significant contributions to their field, gained widespread recognition, and often attracted substantial research funding. They are critical for the advancement of science and innovation, and they have a significant influence on the transfer of knowledge and technology to industry. Identifying potential star scientists before their performance becomes outstanding is important for recruitment, collaboration, networking, or research funding decisions. The objectives of this study are to develop a prediction method for star scientists in the artificial intelligence scientific ecosystem and to investigate the features related to their success. Bibliographic data was extracted from Scopus and data mining techniques were employed to gain insights into the authors' discipline, gender, and ethnicity. The h-index was used as a proxy for research performance, and a dynamic profile of authors was established. Rising stars were found to have different patterns compared to their non-rising stars counterparts in almost all the early-career features. Social network analysis showed that certain features such as gender and ethnic diversity play important role in scientific collaboration and that they can significantly impact an author's career development and success. The prediction of rising stars was based on the author's early-career characteristics such as quantity and quality of research output, metrics obtained from social network analysis, and various diversity measures. Several classifiers in machine learning were trained, tested, implemented, and compared in the prediction task. It was shown that the Random Forest outperformed other classifiers and that the most important combination of features in predicting star scientists in the artificial intelligence field is the number of articles, group discipline diversity, and weighted degree centrality. Our findings highlight the importance of considering the authors' characteristics from different categories of features in the early stages of their careers to identify rising stars. This study offers valuable insights for

researchers, practitioners, and funding agencies interested in identifying and supporting talented researchers.

Keywords: Star Scientists, Social Network Analysis, Machine Learning, Data Mining

Acknowledgment

It is with immense pleasure and gratitude that I take this opportunity to express my heartfelt thanks to all those who have helped and supported me during the completion of my thesis. First and foremost, I would like to express my sincere appreciation to my thesis supervisor, Professor Andrea Schiffauerova, for her invaluable guidance, encouragement, and support throughout my research journey. Her expert knowledge, insightful feedback, and unwavering support have been instrumental in shaping my research and guiding me toward the successful completion of my thesis. I am also deeply grateful to my thesis co-supervisor, Dr. Ashkan Ebadi, for his valuable insights, and constructive feedback, and for sharing his expertise with me. His guidance and support have been instrumental in helping me to refine my research and to see it through to completion.

I would also like to extend my heartfelt thanks to my family, especially my mother, father, and brother for their unwavering love, support, and encouragement. They have been a constant source of inspiration and motivation throughout my journey, and I am incredibly grateful for their support. I would also like to extend my sincere thanks to my friends, who have been a constant source of support, understanding, and encouragement. Their presence in my life has been a constant source of joy and motivation, and I am truly grateful for the love and support that they have given me throughout this process.

Finally, I would like to express my appreciation to all those who have contributed to my research, whether through providing feedback, offering help, or support. Without their help, this thesis would not have been possible. Thank you all for your support and for being a part of my journey. I couldn't have done it without you.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Definition
AI	Artificial Intelligence
EDA	Exploratory Data Analysis
EPSRC	Engineering and Physical Sciences Research Council
FN	False Negative
FP	False Positive
GNB	Gaussian Naive Bayes
GUI	Graphical User Interface
HCR	Highly Cited Researchers
ISI	Institute for Scientific Information
LDA	Latent Dirichlet Allocation
LR	Logistic Regression
ML	Machine Learning
NLP	Natural Language Processing
RF	Random Forest
RFE	Recursive Feature Elimination
SJR	SCImago Journal Rank
SL	Supervised Learning
SMOTE	Synthetic Minority Over-sampling Technique
SNA	Social Network Analysis
SVM	Support Vector Machine
TP	True Positive

Chapter1

Introduction

Star scientists are individuals who have made exceptional contributions to their respective fields, characterized by high publication and citation rates, groundbreaking research contributions, and influential collaborations (Hirsch, 2005; Ioannidis et al., 2014). Identifying and understanding the characteristics of star scientists is crucial for advancing research and applications in various fields, as well as recognizing and supporting exceptional researchers (Azoulay et al., 2011). Star scientists are particularly important due to their ability to drive innovation and interdisciplinary collaborations. They often have the expertise and influence to bring together researchers from different fields, leading to new insights and breakthroughs in research (Lee & Bozeman, 2005; Uzzi et al., 2013).

Motivated by the importance of understanding and recognizing star scientists, this thesis focuses on developing a method that can accurately predict the individuals becoming star scientists in their respective fields. Specifically, this research focuses on predicting star scientists in the field of AI, which has seen explosive growth and significant advancements in recent years (Figure 1). The motivation for predicting star scientists in AI is multifaceted. First, the field of AI is highly competitive and fast-paced, with new developments and applications emerging constantly. Predicting star scientists in AI can help researchers, funding agencies, and institutions recognize and support exceptional researchers, as well as identify emerging research directions and collaborations. Second, predicting star scientists in AI can provide valuable insights into the characteristics and patterns of successful researchers. By analyzing the features and patterns of star scientists, such as their collaboration networks along with publication and citation records, this study aims to identify key factors that contribute to research excellence and success in AI.

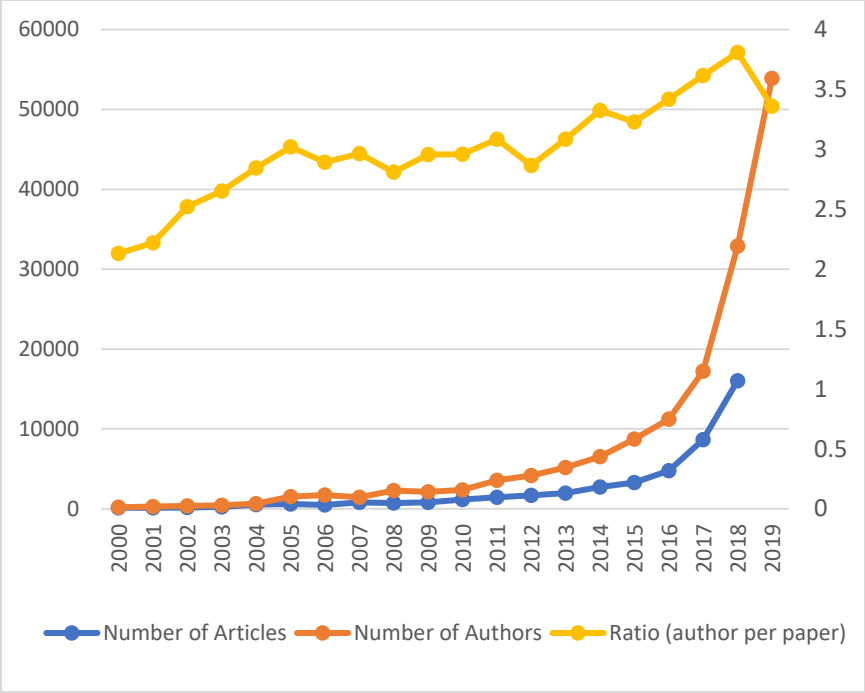


Figure 1. The number of authors and publications over time in the field of AI recorded in Scopus[1] database

To achieve these objectives, this study proposes a prediction method that incorporates features from different categories with a high contribution to the prediction task. Specifically, the research focuses on developing a prediction method that analyzes the collaboration network of researchers in the field of AI. By considering factors such as network measures, diversity, and research output, this thesis aims to predict the individuals becoming star scientists. In addition to predicting star scientists, this research also aims to investigate the differences between rising stars and their peers. By comparing the characteristics and patterns in early career of these groups, we aim to identify key factors that distinguish rising stars and contribute to their success.

[1] Scopus is Elsevier's abstract and citation database launched in 2004. Scopus covers nearly 36,377 titles from approximately 11,678 publishers, of which 34,346 are peer-reviewed journals in top-level subject fields: life sciences, social sciences, physical sciences, and health sciences. http://www.scopus.com

Chapter2

Literature-Review

2.1. Definition of Star Scientists

Measuring scientific performance provides helpful information for comparing academics and highlighting the star scientists – scholars with outstanding performance compared to their peers. Studies on the conceptualization and identification of stars in various academic disciplines or sectors considered several techniques, some of which are described in Table 1. The fundamental belief that stars have extraordinary value creation in organizations serves as the foundation for all conceptualization strategies used by studies of star scientists.

Table 1. Star scientists' definitions in literature

Author(s)	Field/Industry	Star scientists
Lowe & Gonzalez-Brambila (2007)	Six disciplines from biology and chemistry to computer and electrical engineering, and materials	Highly productive scholars that become entrepreneurs.
Hess & Rothaermel (2011)	Pharmaceutical industry	"Researchers who had both published and been cited at a rate of three standard deviations above the mean"
Niosi & Queenton (2010)	Biotechnology firms and academics	Those with more than five patents and more than one major publication per year
Azoulay et al. (2010)	Academic life sciences	Scientists who satisfy at least one of the following criteria for cumulative scientific achievement: (1) Highly funded scientists; (2) highly cited scientists; (3) top patentees; and (4) members of the National Academy of Sciences.
Trippl & Maier (2011)	All scientific disciplines	Authors of highly cited research papers, identified by the number of citations they generated in journals in the ISI databases in the period 1981–2002". (P. 1654)
Schiffauerova & Beaudry (2011)	Biotechnology	Those have more than 20 patents
A. Hess & Rothaermel (2012)	All high-tech scientific academic disciplines	Faculty founders of new tech ventures are star scientists
Oettl (2012)	Immunology	People with high levels of scientific productivity (publications) and helpfulness
Moretti & Wilson (2014)	Biotechnology	Those patent assignees whose patent count over the previous ten years is in the top 5% of patent assignees nationally
Hoser (2013)	Nanotechnology	Those academics with the maximum number of citations
Tartari et al. (2014)	All scientific disciplines	Academics in the top 1% of the distribution of citations in their discipline, and the top 25% of the distribution for grants received from the EPSRC"
Nagane et al. (2018)	All scientific disciplines	Scientists in the list of Highly Cited Researchers (HCR) published by Clarivate Analytics company.
Abramo et al. (2019)	All scientific disciplines	Professors place among the top 10% by fractional scientific strength in each scientific disciplinary sector.
Sá et al. (2020)	All scientific disciplines	Scientists holding research chairs

2.2. Importance of Star Scientists

Star scientists are crucial to the advancement of research and applications in various fields (Wagner & Leydesdorff, 2005). They can drive innovation and interdisciplinary collaborations, often leading to new insights and breakthroughs in research (Wuchty et al.,

2007). Recognizing and supporting exceptional researchers is therefore critical for the progress of science and technology (Azoulay et al., 2019). In this section, we will delve deeper into the importance of star scientists and their impact on research and society.

Star scientists are often characterized by high publication and citation rates, groundbreaking research contributions, and influential collaborations (Lee & Bozeman, 2005). They are often the ones who lead research efforts in their fields, providing guidance and inspiration to others. Moreover, they are often the recipients of prestigious awards and honors, which can have a significant impact on their careers and research trajectory (Bornmann, 2014).

The impact of star scientists goes beyond their individual accomplishments. They have been found to contribute disproportionally across contexts (O'Boyle Jr & Aguinis, 2012) and collaborate in a wider range of scietntists (Abramo et al., 2019). They play a gatekeeper role to boost the knowledge flow within several research groups and affect their neighboring researchers in terms of output and recognition (Azoulay et al., 2010; Oettl, 2012). Stars not only affect academia but also have a significant influence on firms when they transfer advanced knowledge to new technology firms through different channels such as founders or advisors (Zucker et al., 1998). Their work often leads to the development of new technologies and applications, which can have significant economic and societal benefits. In addition, they are often involved in policy-making and public engagement, helping to shape public discourse and opinion on important issues (Leshner, 2003).

2.3. Predicting Star Scientists

A researcher's or professional's career appraisal over time shows that individuals can go through different phases based on their performance during the course of their career (Figure 2). While some people have consistent success throughout their careers, others have fluctuating tendencies. Because productive scientists receive greater recognition, which encourages their future productivity, star scientists are anticipated to have adominant profile during their junior stage, allowing them to reap the benefit of accumulative advantage and become a star (Azoulay et al., 2010). Given the importance of star scientists, the prediction of rising stars in academia has been an active research area in recent years, and various approaches have been proposed to address this task.

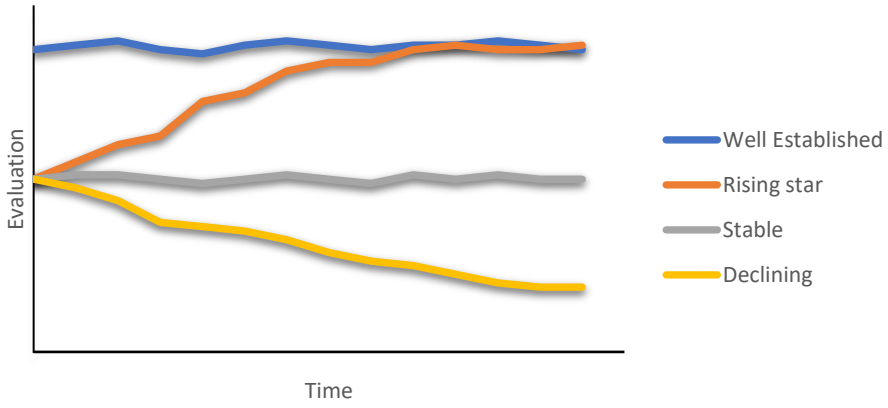


Figure 2. Evaluation of authors over time (Tsatsaronis et al., 2011)

Major techniques pertaining to social network analysis (SNA) and machine learning (ML[2]) prediction employing bibliographic networks[3] were used to identify rising stars. In the beginning, researchers attempted to use the dynamics of author ranking based on compiling a list of each author's important scores to carry out SNA (Daud et al., 2013; Li et al., 2009). There have also been studies that used ML methods to mine potential future star scientists. For instance, Daud et al. (2015) used ML classifiers based on increases in the number of citations coupled with three kinds of characteristics (author, venue[4], and co-authorship) to forecast future rising stars. Their findings demonstrated the significance of the venue characteristic for the prediction job. Regression approaches were also used, taking both temporal and content variables into account (Zhang et al., 2017). They discovered that temporal factors, as opposed to venue features, are the strongest predictors

[2] ML is a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty (Bishop & Nasrabadi, 2006).

[3] Bibliographic networks refer to networks of bibliographic data that capture relationships between scholarly publications, authors, and citations.

[4] Venue characteristics typically refer to the characteristics of the publication venue where a researcher's work is published. The publication venue could be a journal, conference, or any other academic forum where research is presented and published.

of rising stars. Considering different times for computing the performance and features, this technique could be more reasonable than the previous one. Both, however, use citation counts as their assessment measure for identifying rising stars, which may not be a reliable signal to fully assess a researcher's success (Nie et al., 2019).

Nie et al. (2019) proposed a new evaluation method to improve this downside. They used the number of articles, their quality, the number of citations, the domain cited factor, and the co-impact authors to produce a composite score for each scientist. Then, they employed increments in their composite score over two successive five-year periods to create their labels. Their techniques outperform earlier research based on citation increase, according to a comparative ML analysis, and the venue characteristic was revealed to be a key factor in identifying rising stars.

Despite the numerous approaches proposed for identifying rising stars in scientific research, there are still several research gaps in the field. One such gap is the absence of a reliable and all-encompassing method for identifying star scientists. A universal scientific performance evaluation indicator is the cornerstone of an all-encompassing method for identifying rising stars. While Nie et al. (2019) proposed a composite score based on multiple indicators to overcome the limitations of single-number indicators such as citation counts used in prior studies, they acknowledged this approach require a wide range of metadata about citations, coauthors, and venues which may complicate the assessment of scientific performance.

On one hand, there is a need to develop more precise and effective models that can incorporate various features to predict rising stars. Several studies have measured the correlation between current research impact and different attributes. For example, SNA has been utilized, and significant positive measures were found between degree centrality and the h-index (Abbasi & Altmann, 2011). Additionally, diversity measures were investigated and revealed that, in general, diversity has a strong association with research performance (AlShebli et al., 2018). Including such features may improve prediction results.

On the other hand, previous researches (e.g. Daud et al., 2015; Nie et al., 2019) have compared the predictive results between different categories of features and identified the significance of these features based on this comparison. However, a more comprehensive approach would be to consider a combination of features from various

categories rather than looking at each category in isolation. Furthermore, conventional methods have often relied on under-sampling techniques that exclude a portion of the researcher population, which should be avoided in the development of more accurate and inclusive models.

Lastly, most existing approaches for identifying rising stars are focused on identifying individual researchers. However, scientific research is often a collaborative effort, and the success of a researcher may also depend on the success of their collaborations. Therefore, approaches that consider not only individual researchers but also their collaborations and the dynamics of their scientific networks should be explored.

In conclusion, identifying rising stars in scientific research is a complex task that requires the consideration of various factors. Existing approaches have made significant progress in identifying rising stars, but there are still several research gaps that need to be addressed. In this study, we tried to cover some of these gaps. The first objective of this thesis is developing a prediction method in the field of ML to predict star scientists using a combination of features from different categories. The other objective is investigating the relationship of early-career features and comparing rising and non-rising star scientists: Because the current characteristics of academics may impact their future success, we investigate these characteristics to determine if they are linked with each other. Furthermore, we compared if there is a significant difference in these features between rising and non-rising stars.

Chapter3

Methodology

The Methodology of this thesis includes the several steps (Figure 3). To collect data on researchers in the field of AI, various datasets were obtained from the Scopus database. These datasets were cleaned and merged using pre-processing techniques. Then, we extracted significant metadata such as research themes, gender, and ethnicity of the authors, as well as network measures and diversities. Through social network and statistical analysis, we gained a thorough understanding of the authors' behavior and academic relationships. Finally, based on the increase in their scientific impact over two consecutive periods, junior AI scientists were labeled as rising stars or non-rising stars. Then, using over-sampling approaches, several classifiers were trained and tested using features in the first five years of the authors' careers. To get the best prediction outcome, candidate classifiers' comparison results were employed. In addition, the importance of features was examined in order to ascertain which characteristics of scholars in the early stages of their careers are crucial for them to develop into star scientists in the future. This leads to proposing a combination of features from different categories which are important for the prediction task.

3.1. Scientific Production in Artificial Intelligence

3.1.1. Data Collection and Preprocessing

To explore the AI academic ecosystem, considerable data about individuals, their publication, their research performance, and collaborations are required. The basic strategy of the thesis is based on the extensive use of information gathered from Scopus, the world's biggest abstract and citation database of peer-reviewed academic literature. In the field of computer science, Scopus retrieved the highest percentage of articles and indexed a high number of unique articles (Cavacini, 2015). This was the main motivation in selecting Scopus over similar databases since the thesis aims to analyze the AI field.

The collection of the data required several stages. First, Elsevier's Scopus was used to retrieve bibliographic information, which included the title, abstract, keywords,

publication date, author list, and other details. The database was filtered to include research articles, conference papers, book chapters, and books published between 2000 and 2019. Only articles with accessible title and abstract were included. The search term ("artificial intelligence" OR "machine learning" OR "deep learning") is used to find papers on artificial intelligence where at least one of the terms was used in the title, abstract, or keywords (Hajibabaei et al., 2022). Then, articles without an abstract were eliminated which resulted in 45,734 publications written by 162,581 authors. The citation patterns and bibliometric indicators of review papers and survey papers might be different from those of other types of research papers, which could potentially skew the results, specially the performance metrics such as h-index (Radicchi et al., 2008; Radicchi & Castellano, 2012; Stringer et al., 2010). As a response, we excluded from our analysis any review publications that we could locate based on tell-tale phrases in the paper's keywords, such as "literature review", "literature", or "survey". This procedure resulted in the discovery and removal of 965 papers.

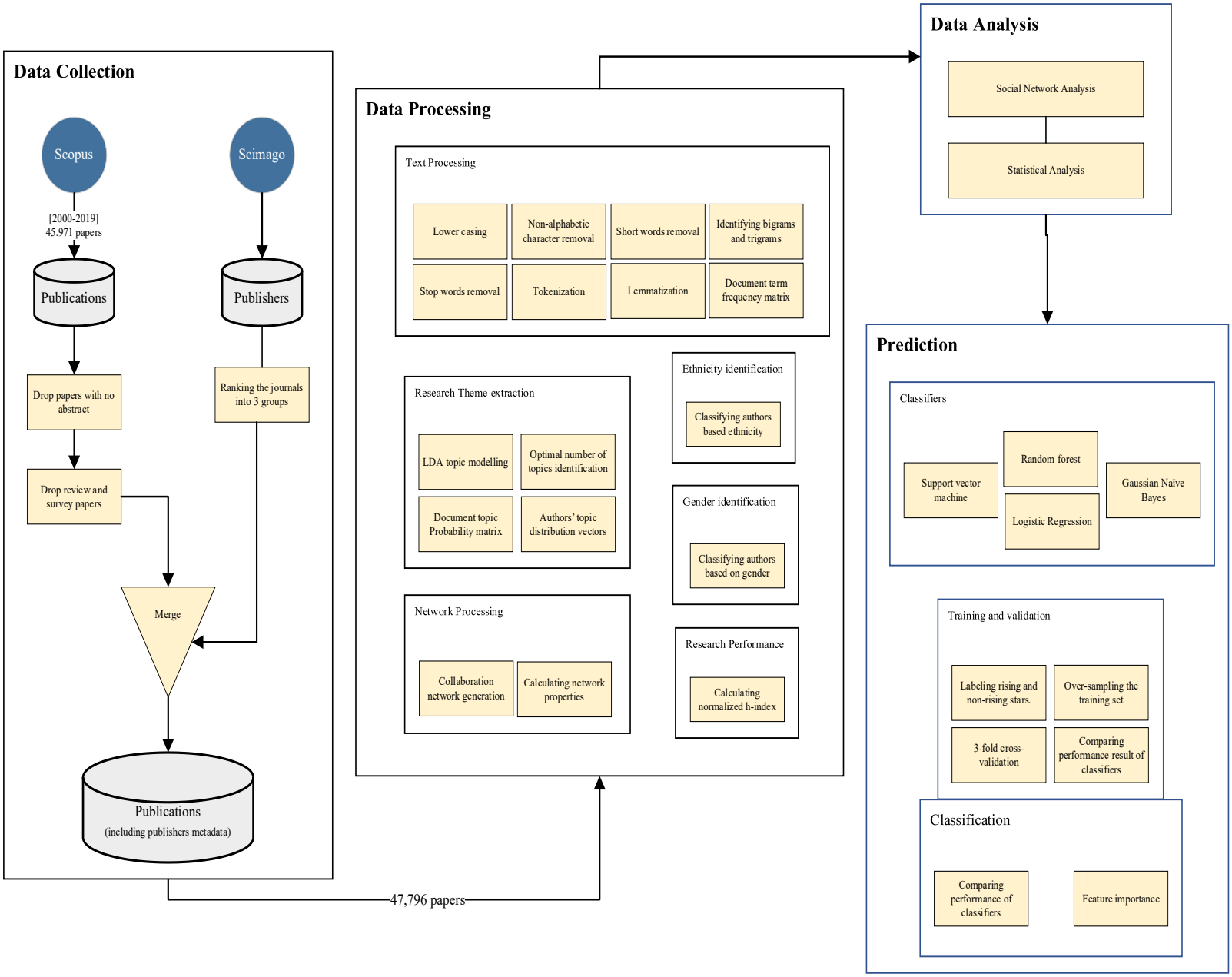
A new query based on these articles was created and retrieved the historical citation count in order to calculate the h-index of the authors throughout different periods of time. In addition, historical publisher metadata was taken from the SCImago[5] Journal website in order to rank the publications according to the publisher's SCImago Journal Rank (SJR[6]) at the time of publication. To this end, publications were ranked into three levels based on the SJR of their publisher at the publishing time: 1) if SJR is more than three standard deviations above the mean of all journals at the publishing time, the rank of the journal is considered as A, 2) if it is more than a standard deviation above the mean but less than three standard deviations above the mean, it is considered as rank B, and 3) if it is less than a standard deviation above the mean, it is considered as rank C.

[5] The SCImago Journal & Country Rank is a publicly available portal that includes the journals and country scientific indicators developed from the information contained in the Scopus database.

https://www.scimagojr.com/

[6] SJR indicator is a measure of the prestige of scholarly journals that accounts for both the number of citations received by a journal and the prestige of the journals where the citations come from.

Figure 3. The conceptual flow



3.1.2. Research Performance Indicators

The scientific performance of scientists the cornerstone of finding star scientists. Scientific performance evaluation is strongly based on the dissemination of scientific research (Lippi & Mattiuzzi, 2017). The use of trustworthy scientific performance metrics in the evaluation of the value of individual contributions to science is acknowledged as being crucial in many academic communities across the globe (Zerem, 2013). In fact,

making these types of analyses is a science itself. Along with evaluating a scientific output, a wide range of other scientific activities can also be used to gauge a scientist's scientific credibility. Examples of these activities include the quantity and quality of extramural funding, leadership in national or international academic societies, service on respected journal editorial boards, participation in government-sponsored national peer review committees, and the number of students who obtained PhD under the individual's supervision (Larivière et al., 2016; Moed, 2010). Although the aforementioned activities are significant and contribute to a certain extent to a scientist's scientific credibility, the pertinent science metrics systems only consider publications and ignore other factors of scientific relevance that are typically taken into account when determining a scientist's promotion and tenure (Greenberg, 2009; Siler et al., 2015). This is due to the fact that these activities, regardless of their significance, are quite heterogeneous since each has unique characteristics and necessitates the use of various assessment criteria (Bornmann & Mutz, 2015). As a result, there are no universal evaluation criteria for these scientifically relevant factors, and their worth is instead mostly determined by the assessment's objective (Wouters, 2014).

At the universal level of analysis, many indicators are routinely established, typically based on both the production of scientists as well as the impact of their documents, such as the impact factor (IF), the total number of documents, the number of citations, the number of citations per document, or the number of highly cited publications. The use of combined indicators that provide information on various elements of scientific output is widely suggested (e.g, van Leeuwen et al., 2003). The h-index, on the other hand, was created in 2005 (Hirsch, 2005), and it combines a measure of the amount and impact of a researcher's scientific output into a single metric. According to Hirsch, "A scientist has index h if h of his or her papers have at least h citations each and the other papers have less than h citations each".

This indicator has stimulated the interest of the scientific community, as seen by a large number of papers on the subject. The fundamental advantage of the h-index is that it combines a measure of quantity and impact into a single indicator. It has been computed in several domains such as physics (Hirsch, 2005), biology (Bornmann & Daniel, 2005), information science (Cronin & Meho, 2006), and business (Saad, 2006). It may be used for

journal evaluation (Braun et al., 2006; Rousseau, 2006), comparative description of scientific themes (Banks, 2006), and awarding of scientific prizes (Glanzel & Persson, 2005). Hirsch demonstrated that this indicator exceeds conventional single-number criteria often used to evaluate a researcher's scientific output. Also, anyone with access to bibliographic databases can easily retrieve the h-index, and it is also straightforward to comprehend.

Despite the advantages, there are still some limitations to this indicator to address. Because of variances in output and citation patterns within areas, there are inter-field disparities in typical h values (Hirsch, 2005), hence the h-index should not be used to compare scientists from various disciplines. The other limitation would be the effect of time. In other words, the h-index is affected by the length of each scientist's career since the pool of publications and citations grows over time (Hirsch, 2005; Kelly & Jennions, 2006). Hirsch (2005) proposed the "m parameter" to compare scientists at different phases of their careers, which is calculated by dividing h by a scientist's scientific age (number of years from the author's first publication). Moreover, highly cited articles are useful for determining the h-index, but once they are chosen to be among the top h papers, the amount of citations they get is immaterial. This is an h-index disadvantage that Egghe (2006) has attempted to alleviate with a new index dubbed the g-index[7].

In this thesis, h-index is considered a proxy of the research performance of the authors and calculated at different stages of the authors' careers. The nature of the study is what led to this choice. First, despite the expectation that stars will produce highly cited research, it may not be feasible to evaluate the scholars using the g-index in order to identify rising stars. The g-index will be heavily impacted by a few highly referenced articles, for instance, if an author has several papers but only one or two of them have high citations while the others have low citation counts. A star scientist's outstanding profile gives the impression that their work is regularly, if not always, excellent. Second, the h-index performs well in this thesis since the focus of study is on authors from the same domain. Finally, as was already noted, the h-index is widely available, allowing people of all knowledge levels and backgrounds to utilize it.

[7] the g-index is the unique largest number such that the top g articles received together at least g^2 citations.

3.2. Data Mining

There are some personal characteristics of authors that should be investigated to address patterns in academic society. For example, it has been established that there is a gender bias in academia, and this has been noted in the community of star scientists as well (Sá et al., 2020). It has also been shown that stars have a higher propensity to collaborate at the international level compared to their non-star peers (Abramo et al., 2019). Furthermore, some diversity measures, such as ethnic diversity, are substantially connected with scholars' scientific achievement (AlShebli et al., 2018). To assess these elements, we must first extract information that is not included in the original data. The following are descriptions of these metadata and data mining approaches.

3.2.1. Authors' Disciplinary Profiles

A topic modeling approach is adopted to find the domain of documents that leads to identifying the disciplinary profile of authors. The title and abstract of a paper can properly present the special keywords and main idea of the research respectively in a few well-chosen words (Ebadi et al., 2020). Latent Dirichlet Allocation (LDA) topic modeling (Blei et al., 2003) on the merged titles and abstracts of the AI-related publications is used to derive their domain. To apply the LDA model, sequences of preprocessing steps were carried out on the corpus including lower-casing, short words removal (i.e., words with less than 3 characters), custom stop-words removal, phrase detection, tokenization, and non-alphabetic characters removal.

To find the best number of topics, we built several LDA baseline models with different numbers of topics and then evaluated them using metrics, such as perplexity and log-likelihood (Griffiths & Steyvers, 2004), in addition to visualizing the inter-topic distance mapping (Figures 4 shows the overlap of models considering 8 and 9 topics) to find the model with the highest performance. Along with quantitative metrics, the quality evaluation was also applied by observing keywords and document-topic distribution of models. Combining both quantitative and qualitative analysis, the best number of topics is found to be 8. Each topic along with its keywords is presented in Table 2. Afterward, Given

the LDA model, each publication has a chance of being related to more than one topic. the authors' research disciplines are identified based on the average topic distribution across their previous publications using the document-topic probability matrix created by the LDA technique. Each author was then represented as an 8-discipline topic distribution vector, with each component representing the average subject distribution of the author's previous publications in the given field.

Figure 4. Inter topic distance for LDA model with 8 and 9 topics (via multidimensional scaling). Circles represent the different topics in the LDA model, where the size of each circle represents the proportion of documents in the corpus that are assigned to that topic. Overlapping circles indicate that some documents in the corpus may be assigned to multiple topics with relatively high probability. The marginal topic distribution percentage shows the percentage of documents that have a high probability of belonging to each topic, based on the LDA model. The LDA model with 8 topics resulted the least overlapping among other LDA models with different number of topics.

Table 2. Topics and keywords

Topic	Keywords
1	network, design, application, time, human, artificial intelligence, information
2	protein, gene, sequence, prediction, expression, biological, DNA
3	knowledge, ontology, text, mining, language, annotation, query
4	image, segmentation, brain, detection, shape, region, magnetic resonance
5	prediction, machine learning, neural network, cluster, kernel, error, performance
6	patient, disease, clinical, cancer, machine learning, diagnosis
7	Classification, feature, support vector machine, accuracy, classifier, recognition, training
8	Technology, management, sensor, service, decision support, environment, software

3.2.2. Gender Determination

The gender of the authors is detected by Hajibabaei et al. (2022), members of our research group. They used ML and Natural Language Processing (NLP[8]) to estimate researchers' gender from a variety of variables such as their first and last names, affiliation, and country of origin, using an automated gender assignment model trained on a large labeled dataset of names. On a massive labeled data set of names, the tool trains a 3-class ML classifier. It benefits from a customized feature engineering component that increases the basic feature set (for example, first and last names and author affiliations) to improve model performance. Many aspects, such as language rules in different nations, were taken into account when designing and developing new features. For example, most of the female surnames in Russian, Czech, and other Slavic languages, finish with the suffix 'ova'. Such rules were incorporated into the inference process. Furthermore, they employed NLP techniques to further expand the feature set by concentrating on various sections of the first and last names. For example, they developed elements that depict the location of researchers based on their association, the last n characters of their last names, and so on. In addition, they determined areas where a particular first or last name appears frequently and added elements to illustrate this information. On this enriched data set, the ML

[8] NLP is a subfield of artificial intelligence and computational linguistics that deals with the interactions between computers and human (natural) languages (Manning & Schutze, 1999).

classifier was trained and produced a label from the set of ["female", "male", "unisex/unknown"].

3.2.3. Author's Ethnicity

The ethnicolr[9] tool is used to determine each scientist's ethnicity. This classifier, in particular, employs a variety of machine-learning approaches to categorize any given name into one of the 13 ethnic groups listed below:

- "Asian, Greater East-Asian, East-Asian"
- "Asian, Greater East-Asian, Japanese"
- "Asian, Indian Sub-Continent"
- "Greater African, Africans"
- "Greater African, Muslim"
- "Greater European, British"
- "Greater European, East-European"
- "Greater European, Jewish"
- "Greater European, West-European, French"
- "Greater European, West-European, Germanic"
- "Greater European, West-European, Hispanic"
- "Greater European, West-European, Italian"
- "Greater European, West-European, Nordic".

3.3. Scientific Collaboration Network Analysis and Structural Metrics

In the early twentieth century, SNA was established with a focus on interactions among social entities, such as group communication, commerce between nations, or business affairs between entities (Boccaletti et al., 2006). It is a diagnostic technique for examining the mechanisms of cooperation and communication among members of various groups (Racherla & Hu, 2010). Applying SNA to a certain group of people enables us to

[9] "ethnicolr" is a package in Python provided by Sood & Laohaprapanon (2018):

https://github.com/appeler/ethnicolr/

recognize network member interactions, the quantity, and structure of subgroups within the networks, as well as their organization and evolution (Anklam, 2003). The identification of both strengths and weaknesses inside and among research organizations, industries, and nations as well as the contribution to scientific advancement and financing policies are some aims for social network analysis that are addressed in the literature (Owen-Smith et al., 2002; Sonnenwald, 2007).

A graph of actors (or vertices) and links is used to depict social networks (ties, relations, or edges). By mapping the graph of authors who have coauthored similar articles, the scientific collaboration network serves as an illustration of a social network (Racherla & Hu, 2010; Staudt, 2011; Yin et al., 2006). A node represents a researcher in this thesis, and a link between two nodes implies that these two scientists have at least one joint publication. When two or more scientists interact in a social context, they create a collaborative environment that enables the exchange of information and the accomplishment of tasks aimed at achieving a mutually shared, overarching goal. This collaborative interaction is an essential aspect of scientific research, as it allows researchers to pool their expertise, resources, and knowledge to tackle complex problems and make important discoveries (Sonnenwald, 2007). This concept states that scientific collaborations frequently take place as a result of formal and informal social connections made by people who are from different disciplinary, organizational, and national backgrounds (Barabâsi et al., 2002; Sonnenwald, 2007).

The significance of scientific collaboration network analysis is in its potential to aid in the understanding of how to efficiently communicate professional and scientific information, as well as in evaluating the performance of individuals, organizations, or the entire social network. For example, they may use a researcher's social network to reflect its collaboration activity within a research community (Abbasi et al., 2010). It is demonstrated that in academic society, researchers' position in collaboration networks could affect their performance (Ebadi & Schiffauerova, 2015, 2016). Hence, we try to explore the academic collaboration network in AI academic ecosystem and extract network-related measures to study the correlation between the position of junior researchers in the social network and their status in the future (star or non-star). Moreover,

we measure diversity to discover individual and group patterns in early-career of rising stars.

3.3.1. The Scientific Collaboration Network Matrix and Properties

Several computer software products are utilized as tools for numerical and visual network analysis. Pajek[10] is one of these powerful tools. It was created specifically to operate, handle, and analyze very large networks with 10^3 to 10^6 nodes. It has been utilized in academic papers for many years because of its versatility and strong Graphical User Interface (GUI), which allows for the simultaneous control of various networks, components, and analytic results (Berryman and Angus, 2010).

Pajek format network files were needed to build in order to prepare the data for analysis. We began by exporting the collaboration relations from our database to a two-column Excel file for each year to analyze the social network through time. The first column provides the authors' identification number (author id) as it appears in the database, whereas the second column comprises the coauthors' ids. In the event of numerous publications coauthored by this pair of scientists, the same row may appear more than once. The number of repetitions is used as a weight for the relationship and shows the regularity with which two researchers collaborate. Furthermore, reciprocal rows might occur in the dataset if the scholar is listed as an author once and as a collaborator another time. Pajek excludes non-identical pairings and persons who name each other reciprocally unless you explicitly advise it otherwise for network visualization. Using the Excel2Pajek[11] program,

[10] Pajek is developed by V. Batagelj and A. Mrvar, Department of Mathematics, Faculty of mathematics and physics, University of Ljubljana, in 1999. It is freely available for noncommercial use and can be downloaded from the following webpage: http://pajek.imfm.si/doku.php?id=download

[11] Excel2Pajek is a windows program developed in Delphi 7 by Jürgen Pfeffer, from FAS.research, Vienna to convert Excel datasets into Pajek format. It can be downloaded from:

http://vlado.fmf.uni-lj.si/pub/networks/pajek/howto/excel2Pajek.htm

the Excel file was then transformed into a one-mode[12] undirected[13] network (.net) format that Pajek can read.

Following the creation of the historical social network matrixes, we analytically examined the network and determined the following measures for each node (representing each author in our network at different points in time).

- **Betweenness Centrality**

  Betweenness centrality may be used to determine if an actor has potential influence over network communication (Abbasi & Altmann, 2011; Chung & Hossain, 2009). It is determined by dividing the total number of shortest pathways by the proportion of shortest paths (between all pairs of nodes) that pass through a certain node (Borgatti, 2005). The most central vertices are indicated with the highest betweenness centrality (which is between 0 and 1). In other words, high betweenness centrality vertices (authors), also known as gatekeepers, are crucial for the information transfer between various nodes that are directly connected to the most central one (Ebadi & Schiffauerova, 2015a). The betweenness centrality of the given node at time t ($bc_t^{(i)}$) is defined as:

$$bc_t^{(i)} = \sum_j \sum_k \frac{\sigma_{jk_t}^{(i)}}{\sigma_{jk_t}} : i \neq j \neq k \qquad \textit{Equation (1)}$$

  Where σ_{jk_t} denotes the total number of the shortest paths between node j and k up to time t , and $\sigma_{jk_t}^{(i)}$ is the number of those paths containing node i in the same period.

- **Degree Centrality**

  Indicators of an actor's communication activity include degree centrality (Abbasi & Altmann, 2011; Chung & Hossain, 2009). The degree of a node ($dc_t^{(i)}$) in a straightforward undirected network indicates how many neighbors node i has up to

[12] One-mode networks are those in which we examine how each actor is connected to every other based on a single connection, such as friendship.

[13] Because both parties are equally invested in a connection, undirected ties (edges) are used to represent those relationships in undirected networks.

time t. Similarly, the degree of each node, which represents a researcher, shows how many co-authors that person has collaborated with in the past.

- **Weighted Degree Centrality**

  The weight of the linkage w_{ij} between nodes i and j represents the strength of their collaboration tie, which reflects how many times they have collaborated. We determined each author's weighted degree by dividing the sum of their link weights (total number of co-authorships) by the total number of distinct co-authors. Scholars who have a strong relationship (often co-authorship with the same partner) are seen to be loyal (Abbasi & Altmann, 2011).

- **Clustering Coefficient**

  A vertex's clustering coefficient in a network graph indicates how near its neighbors are to being a clique[14] up to time t (complete graph). In other words, it demonstrates how closely each scientist is tied to his or her colleagues and the likelihood that they will form a closed research network. The clustering coefficient is simply the number of edges between neighbors divided by the maximum possible for the type of network. It is worth noting that the clustering coefficient has been decreasing over periods, with only about a 20% chance of two scientists collaborating if both have collaborated with a third scientist (Perc, 2010). For a node i with $k_t^{(i)}$ neighbors at time t, the local clustering coefficient t is defined as:

$$cc_t^{(i)} = \frac{2e_t^{(i)}}{k_t^{(i)}(k_t^{(i)}-1)} \quad \textit{Equation (2)}$$

where $e_t^{(i)}$ is the number of edges connecting the $k_t^{(i)}$ neighbors of node i to each other at time t.

3.3.2. Diversity

Modern societies place great importance on diversity (Ager & Brückner, 2013; Puritty et al., 2017; Wagner & Jonkers, 2017). Diversity has inspired numerous governmental and employment practices and has the potential to have far-reaching and long-lasting

[14] Based on the graph theory a clique in an undirected graph is a subset of its vertices such that every two vertices in the subset are connected by an edge.

consequences for society (Arcidiacono et al., 2015; Brown & Langer, 2015; Wagner & Jonkers, 2017). Academia is one area where diversity and its effects are widely explored (Hong & Page, 2004; Woolley et al., 2010). Co-authorships, which typically involve scientists from diverse places, fields, and backgrounds, reveal the structure of academic collaboration (Deville et al., 2014; Jia et al., 2017). The tendency towards the collaboration network analysis has already prompted endeavors to comprehend the fundamental characteristics that contribute to academic performance (Fortunato et al., 2018). Many of these characteristics, such as discipline (Hajibabaei et al., 2022), gender (AlShebli et al., 2018), academic age (Jones & Weinberg, 2011), ethnicity (Freeman & Huang, 2015), and affiliation (Gershenson, 2014; Jones et al., 2008), have been researched, and their impact on research performance has been described.

When exploring diversity in research collaborations, we investigate five classes of diversity:

- **Ethnic Diversity**

  This type of diversity considers each co-author's ethnic background. As explained in the preceding section, we utilize the name ethnicity classifier to determine each scientist's ethnicity.

- **Gender Diversity**

  This class of diversity takes into consideration the gender of researchers in the collaboration networks.

- **Age Diversity**: The term age in this context refers to a scientist's academic age, which is calculated for each author from the first publication. Classification of authors' age is based on the following bins:

  Academic age group 1: 0-5 years of experience.

  Academic age group 2: 5-10 years of experience.

  Academic age group 3: 10-15 years of experience.

  Academic age group 4: 15-20 years of experience.

- **Discipline Diversity**

  Both independently and collaboratively, this category of diversity may be quantified. The diversity of co-authors' fields of competence, which correspond to the topics with the highest likelihood in each disciplinary author's profile, have been

considered at the group level. At the individual level, this diversity is identified by the diversity of fields represented in his/her publications.

- **Affiliation Diversity**

  This class of diversity takes into consideration the country of the affiliations of the co-authors of a paper and measures how diverse this collaboration is at the international level.

Any given diversity measure reveals how unlike its components are from one another. A metric of diversity must be used to analyze the connection between this feature and the success of the linked group. Shannon entropy (Shannon, 1948) is a widely used diversity metric (Aydinoglu et al., 2016; Feng & Kirkley, 2020; Gray, 2011) that evaluates the diversity in predicting the type of an element picked at random from the set under study. Shannon entropy of set X is calculated as follow:

$$H(x) = \begin{cases} -\frac{1}{\log(|X|)} \sum_{x \in X} p(x) \log p(x) & \forall\ |X| > 1 \\ 0 & \forall\ |X| = 1 \end{cases} \quad \textit{Equation (3)}$$

Where $p(x)$ is the probability of element x in set X. It is notable that in the case of categorical elements, this probability is equal to the fraction of the element in the whole set.

Chapter4

Result

4.1. Network Visualization and Mathematical Analysis

A visual depiction of social networks allows for a comprehensive knowledge of huge and complicated societies, such as academic research groups (Racherla & Hu, 2010). To explore the scientific collaboration network, we constructed a network for each year of the study. Analyzing the network of the authors can provide valuable information about the structure of the network and its key players. For example, it can help identify the most influential authors, the clusters of authors who work closely together, and the key pathways or bridges that connect different clusters. To gain a comprehensive understanding of scientific collaboration network among AI researchers, we explored their network in 2014 as an instance.

At first, by visualizing the network of most central authors, who are often the top 1% in terms of betweenness centrality, we gained insights into structure of network, including how densely connected different groups or clusters of nodes are, and how these clusters may be linked together through the central nodes. This examination can also identify which individuals or groups are most influential or have the greatest potential to spread information or influence within the network. Overall, it can reveal the dynamics, structure, and key players within a collaboration network (Figure 5).

The analysis of this network revealed the total number of 99,479 connections (edges) between authors, including 5,919 multiple lines representing repeated collaborations. The density of the network was found to be 0.0002, indicating that only a small percentage (0.019%) of all possible edges are present. This low density is typical in large networks, as the number of connections that can be maintained by everyone is limited in comparison to the rapidly increasing number of possible connections as the network size increases (de Nooy et al., 2005). The average degree centrality of the network was 5.98, indicating that on average, each vertex is involved in almost 6 connections. A higher degree leads to a denser network. The average degree is a better measure of overall consistency within the network than density, as it is not affected by network size and can be compared across

networks of different sizes (de Nooy et al., 2005). The average degree with summed lines was 3.18, indicating the average number of vertices connected to a specific vertex (or its neighbors). Betweenness centralization, which is the variation in betweenness centrality scores divided by the maximum possible variation, was found to be 0.0005. Additionally, the network clustering coefficient was 0.85, indicating that nodes tend to cluster together and create tightly knit groups with a relatively high density of connections. In general, high transitivity is considered to be highly clustered or cliquish[15].

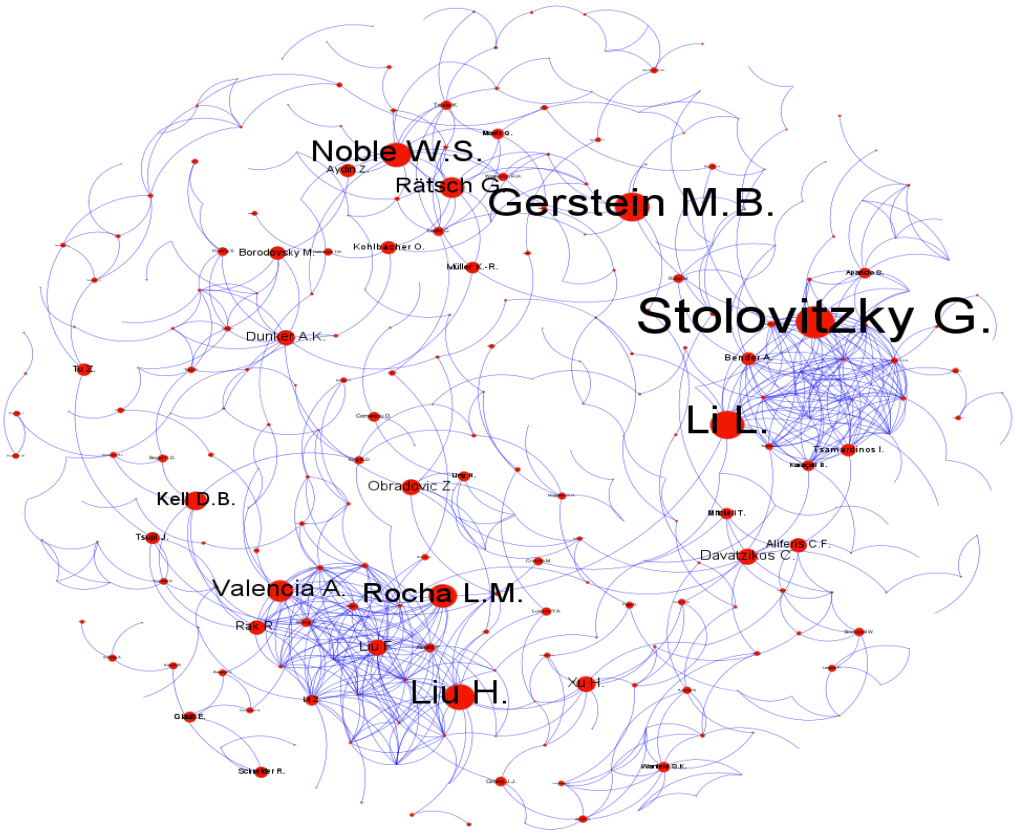


Figure 5. Snapshot of most central authors' network (who stand for top 1% in terms of betweenness centrality) in 2014.

[15] A network that is highly clustered or cliquish means that there are a high number of closed triangles of connections among the nodes in the network.

Table 3. The analysis of the network in 2014 (macro level analysis)

Number of vertices (n): 31,288	
Total number of lines (Edges)	99,479
Number of multiple lines	5,919
Density (no loops allowed)	0.0002
Average degree	5.98
With summed lines average degree	3.18
Network betweenness centralization	0.0005
Network clustering coefficient (transitivity)	0.85

4.2. Identify and Characterize the Rising Star Scientists

To study the junior researchers who become star scientists, this thesis includes only authors who had their first publication between 2006 and 2010 during the initial decade of their careers with at least one collaboration (9,391 authors). The 10-year time frame was chosen as it is widely used in studies of rising stars (e.g., Daud et al., 2015; Nie et al., 2019), allowing for a better understanding of their upward trend during the early-career stage. By choosing authors who began their careers within similar time frame, the study allows for a reasonable comparison of relative success while accounting for potential confounding factors such as historical context and technological advancements that may have impacted the career paths of authors who started at different times.

In this section, we aim to identify researchers who are rising stars, in order to analyze their research performance and early-career behavior. Rising stars are considered to be authors who have an h-index growth rate that is significantly higher than the average, typically three standard deviations above the mean. This growth rate is calculated by comparing the h-index of an author between the first and second five-year period of their career.

$$H_{GR} = \frac{(h_2 - h_1)}{(t_2 - t_1)} \quad \textit{Equation (4)}$$

Where h_1 and h_2 are the h-index in the first five years and first ten years respectively, and t_1 and t_2 is the time of the first and second period respectively.

Table 4. The top ten authors who exhibited the highest growth rate in their h-index during the second five of their careers.

Name	Year of first publication	h-index in the first 5 years	h-index in the first 10 years	h-index growth rate
You Z.-H.	2010	3	14	2.2
Xu H.	2006	2	10	1.6
Tang B.	2009	2	10	1.6
Xu J.	2009	2	9	1.4
Müller K.-R.	2007	3	10	1.4
Ekins S.	2009	1	8	1.4
Ceriotti M.	2010	1	8	1.4
Heider D.	2009	4	10	1.2
Ballester P.J.	2010	4	10	1.2
Sánchez C.I.	2010	1	7	1.2

We identified 171 rising stars, representing roughly 2% of the total population, and analyzed their early-career characteristics such as research productivity, diversity indicators, and social network measures during the first 5 years of their career. The following table lists these features.

Table 5. Features

Feature
Total number of Articles
Number of articles based on publisher ranking
Citations
h-index
Individual discipline diversity
Group discipline diversity
Ethnic diversity
Gender diversity
Affiliation diversity
Age diversity
Degree centrality
Weighted degree centrality
Clustering Coefficient
Betweenness centrality

4.3. Statistical Analysis

Initially, we investigated the relationship between the early-career attributes and growth rate of research performance. To do this, we calculated the correlation coefficient between these variables. As depicted in Figure 6, there is a strong correlation between the growth rate of the h-index and the number of articles, particularly level B articles. Additionally, a high correlation was found between the h-index, weighted degree centrality, and group discipline diversity within the first five years of a researcher's career. Furthermore, a correlation was established between certain diversity, performance, and social network measures. For example, gender diversity was found to be correlated with degree centrality, weighted degree centrality, and clustering coefficient, while the highest correlation was observed between ethnic diversity and clustering coefficient among these two groups of features. This could suggest that individuals from diverse ethnic groups tend to form close relationships and create dense connections, resulting in a high clustering coefficient in the early career of authors. Similarly, it could indicate that authors who collaborate with individuals of different genders, rather than only individuals of the same gender tend to have a larger network of connections. This highlights the importance of gender diversity in scientific collaboration and the potential impact it can have on the

growth and success of an author's career. Furthermore, both the number of articles and citations demonstrated a correlation with group discipline diversity, which could indicate that a diverse research group in terms of the research background of individuals is likely to produce more published papers. Finally, as expected, the number of level A articles was found to have a strong correlation with citation count, implying that publishing in high-level venues brings more recognition which can lead to a higher number of citations.

Furthermore, we aimed to investigate whether rising stars possessed different early-career characteristics compared to the general population of scientists. To do this, we used a statistical method called pairwise comparison hypothesis testing. This method allowed us to make decisions based on data from the study and determine if the results were statistically significant. A two-sample t-test was implemented to test the null hypothesis, which stated that there was no significant difference in the means of the two groups being compared. We also assumed that the variances of the two groups were known and that there was independence between the samples. The goal of the hypothesis testing was to determine if there was enough evidence to reject the null hypothesis and conclude that there was a significant difference between the two groups being compared. The result showed that a significant difference exists between these two groups in all the features except ethnic diversity (Table 6).

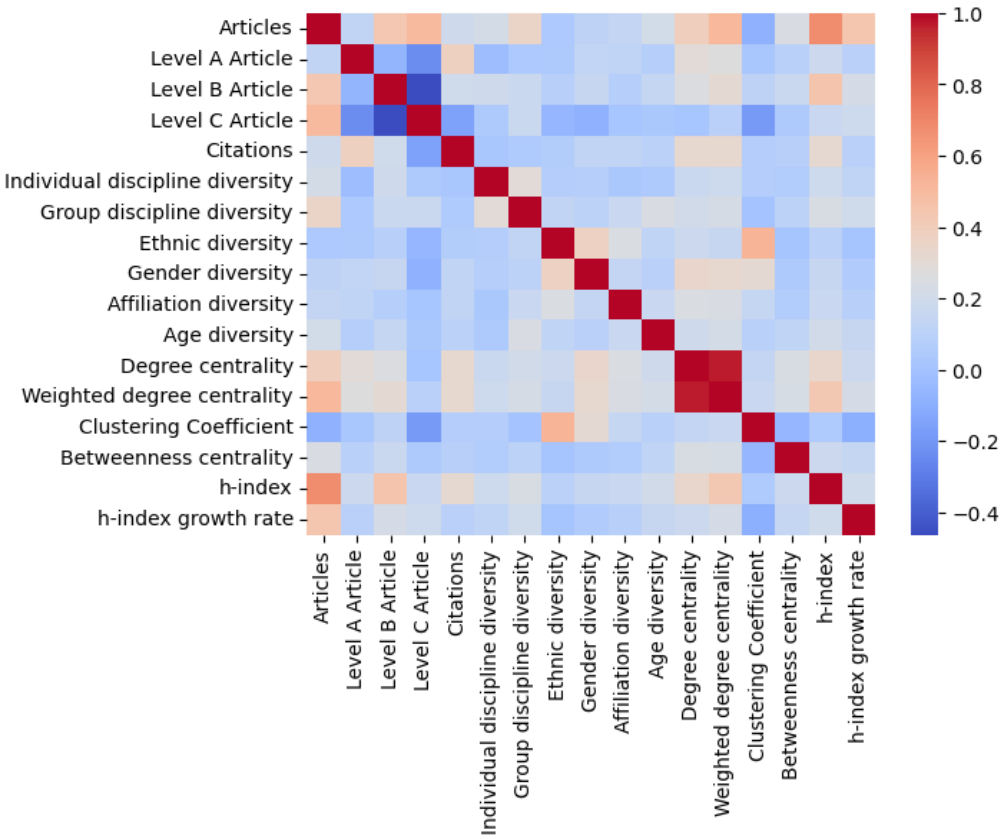


Figure 6. Correlation heatmap

Table 6. Two-sample t-test

Feature	Reject Null Hypothesis
Articles	TRUE
Articles with level A publisher	TRUE
Articles with level B publisher	TRUE
Articles with level C publisher	TRUE
Citations	TRUE
Individual discipline diversity	TRUE
Group discipline diversity	TRUE
Ethnic diversity	FALSE
Gender diversity	TRUE
Affiliation diversity	TRUE
Age diversity	TRUE
Degree centrality	TRUE
Weighted degree centrality	TRUE
Clustering Coefficient	TRUE
Betweenness centrality	TRUE
h-index	TRUE

We have conducted an Exploratory Data Analysis (EDA) of our datasets to understand their key characteristics. EDA allows us to gain insights from the data beyond traditional modeling and hypothesis testing. The approach was first introduced by Tukey (1977) to encourage statisticians to actively explore the data and potentially develop new hypotheses for further research. The tables presented in this thesis provide basic statistics for the early-career features of rising star scientists and non-rising star scientists.

A comparison of the EDA between rising and non-rising stars has uncovered some interesting findings about the factors that contribute to career success in academia and research. Firstly, we can observe that rising stars have a higher average weighted degree centrality than their average degree centrality, while the average weighted degree and degree centrality of non-rising stars are almost the same. Furthermore, rising stars have a higher average in both weighted degree centrality and degree centrality compared to non-rising stars. These findings suggest that rising stars tend to collaborate with the same partners repeatedly, resulting in a higher weighted degree centrality score. This pattern may reflect the establishment of strong and productive collaborations, which could be a key factor contributing to their success (Wuchty et al., 2007). On the other hand, non-rising stars may have a more diverse set of collaborators, resulting in a lower weighted degree centrality score. Secondly, rising stars tend to publish more and receive citations from a wider range of publications in their early career compared to non-rising stars. This suggests that they are more successful in achieving recognition for their work. They also published more higher-level articles on average.

Together, these findings highlight the importance of both collaboration and publication quality in establishing a successful career in academia and research. While it is important to produce a substantial amount of work, it is equally important to ensure that the work is of high quality, receives recognition from a diverse range of sources, and is accomplished through strong and productive collaborations.

Table 7. EDA for rising stars

Feature	N	Mode	Mean	STD	Sum	Min	Max
Articles	171	[1,3]	2.82	1.74	482	1	10
Level A Article	171	0	0.29	0.6	49	0	3
Level B Article	171	0	1.35	1.34	231	0	6
Level C Article	171	[4,10,11,15,32]	1.18	1.44	202	0	9
Citations	171	10	42.89	41.27	7334	0	237
Individual discipline diversity	171	0.2	0.19	0.03	32.32	0.09	0.25
Group discipline diversity	171	0	0.15	0.14	25.39	0	0.36
Ethnic diversity	171	0.35	0.25	0.1	42.62	0	0.37
Gender diversity	171	0	0.23	0.12	38.81	0	0.37
Affiliation diversity	171	0	0.15	0.13	25.1	0	0.37
Age diversity	171	0	0.12	0.12	20.54	0	0.35
Degree centrality	171	[3,4]	10.8	9.56	1846	1	63
Weighted degree centrality	171	4	13.12	12.02	2243	1	76
Clustering Coefficient	171	0.07	0.05	0.03	9.32	0	0.15
Betweenness centrality	171	0	0.0002	0.0006	0.03	0	0.005
h-index	171	1	1.79	0.95	306	0	5
h-index growth rate	171	0.6	0.74	0.24	126	0.6	2.2

Table 8. EDA for non-rising star

Feature	N	Mode	Mean	STD	Sum	Min	Max
Articles	9453	1	1.25	0.67	11,769	1	13
Level A Article	9453	0	0.1	0.32	932	0	4
Level B Article	9453	0	0.45	0.67	4,221	0	12
Level C Article	9453	1	0.7	0.74	6,616	0	12
Citations	9453	0	19.63	35.1	185,604	0	757
Individual discipline diversity	9453	0.18	0.16	0.03	1,561.14	0.04	0.25
Group discipline diversity	9453	0	0.04	0.1	410.23	0	0.37
Ethnic diversity	9453	0	0.23	0.14	2,220.02	0	0.37
Gender diversity	9453	0	0.17	0.15	1,614.34	0	0.37
Affiliation diversity	9453	0	0.09	0.13	847.89	0	0.37
Age diversity	9453	0	0.05	0.1	430.33	0	0.37
Degree centrality	9453	0	5.2	4.89	49,194	0	57
Weighted degree centrality	9453	3	5.58	5.59	52,736	0	68
Clustering Coefficient	9453	0.07	0.06	0.03	645.99	0	0.26
Betweenness centrality	9453	0	0.0002	0.0002	0.15	0	0.001
h-index	9453	1	0.58	0.58	9,585	0	8
h-index growth rate	9453	0	0.04	0.09	372.6	0	0.4

4.4. Predicting Star Scientists

Supervised Learning (SL[16]) can handle the prediction problem quite well (Witten et al., 2005), including the prediction of stars (Daud et al., 2015; Nie et al., 2019). When the output is categorical, the problem is called classification. In general, instances in a dataset are classified according to predefined classifications. Both organized and unstructured datasets can benefit from classification (Kadhim, 2019; Sen et al., 2020). Classification terminology includes classification model, classification algorithm, and feature. A classification algorithm, also known as a classifier, learns from the training dataset and assigns each new data point to one of several classes. A classification model, on the other hand, employs a mapping function derived from the training dataset to predict the class label for the test data. The following sections demonstrate the classification approach used in this thesis along with the results.

4.4.1. Classifiers

Classification is a widely utilized method in ML for solving prediction problems with categorical outputs. In the context of rising star prediction, binary classification is often employed to categorize individuals as either a rising star or not. In this study, we investigate four popular classification algorithms that can be used for this purpose: Logistic Regression (LR), Support Vector Machine (SVM), Gaussian Naive Bayes (GNB), and Random Forest (RF).

- **LR** is a simple and well-established algorithm that models the relationship between a dependent variable and one or more independent variables using a logistic function (Hosmer Jr et al., 2013). It is particularly useful for binary classification problems and is known for its interpretability and ease of implementation.

[16] Supervised learning is a method of machine learning where a model is trained on a labeled dataset, which means the input data is paired with corresponding desired output labels. The model is then able to make predictions on new, unseen input data based on the patterns it has learned from the training data. Examples of supervised learning tasks include classification and regression.

- **SVM** is a robust classification algorithm that can handle non-linearly separable data by using a boundary, known as a hyperplane, to separate the data into different classes (Cortes & Vapnik, 1995). SVM is well-suited for high-dimensional data and is known for its robustness to overfitting.
- **GNB** is a fast and efficient algorithm that is based on Bayes' theorem and the assumption of independence between features (H. Zhang, 2004). It is particularly useful for text classification problems and is known for its speed and ease of implementation.
- **RF** is an ensemble learning method that combines the outputs of multiple decision trees to make a final prediction (Breiman, 2001). It is useful for improving the accuracy and stability of predictions in complex, non-linear problems and is versatile in handling both continuous and categorical variables.

4.4.2. Training and Test Sets

The dataset used in this study includes authors who published their first paper between 2006 and 2009 (7,311 authors) as the training set and authors who published their first paper in 2010 (2,313 authors) as the test set. This division was chosen in order to have a clear temporal separation between the training and test sets and to ensure that the predictions are made for a relatively recent period of time.

To address the imbalance in the distribution of rising stars and non-rising stars in the training set, an over-sampling method, the Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002) in the Imbalanced-learn[17] package, was used to balance the classification. This method helps to increase the number of samples in the minority class while still maintaining the characteristics of the original data distribution.

[17] Imbalanced-learn (imported as imblearn) is an open source, MIT-licensed library in Python and provides tools when dealing with classification with imbalanced classes.

4.4.3. Classification

In this section, we will focus on the classification process used in the study. The study used four different classifiers: LR, SVM, GNB, and RF in the Scikit-learn library in Python (Pedregosa et al., 2011). To ensure that the results were robust and reliable, we used an expanding window cross-validation approach for hyperparameter tuning. Expanding window cross-validation is a technique used to evaluate the performance of an ML model in different studies (e.g., Varma & Simon 2006). The technique is called "expanding window" because the size of the training set grows as the validation set moves forward in time. In other words, the validation set starts with a small size and expands over time, giving the model more and more data to learn from.

In our study, we used a 3-fold expanding window cross-validation, meaning that the validation set consisted of 3 equal-sized (1-year) windows of data (Figure 7). The validation process started with the first window as the validation set and the remaining data as the training set. Then, the validation set was expanded to include the next window, and so on, until the entire dataset was used for validation. This approach allowed us to evaluate the performance of the model over time, ensuring that the results were robust and reliable.

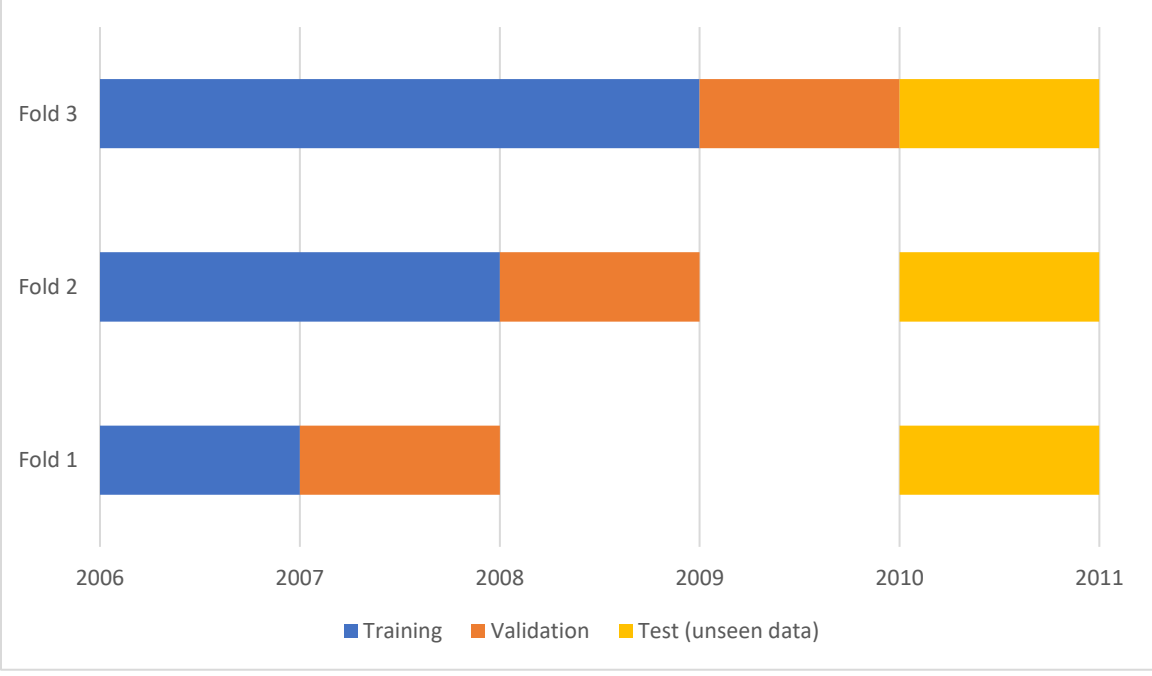


Figure 7. Expanding window cross-validation

The expanding window cross-validation technique is particularly useful when dealing with time series data, as it allows us to consider the temporal dependencies that may exist in the data. By evaluating the performance of the model over time, we can obtain a more accurate assessment of the model's ability to generalize to new, unseen data. It is worth noting that the expanding window cross-validation technique is more complex than other validation techniques, such as k-fold cross-validation, but it can provide more reliable results for time series data.

Additionally, to select the most relevant features for each classifier, we used a Recursive Feature Elimination (RFE) (Guyon et al., 2002). RFE is a feature selection technique used in ML and data mining to select the most important features in a dataset. The goal of RFE is to reduce the dimensionality of the data by removing the least important features, while retaining the most important ones. This is achieved by recursively removing the feature with the lowest weight until a specified number of features is reached. RFE is particularly useful when dealing with high-dimensional datasets, as it can help to improve the performance of classifiers by reducing the number of irrelevant or redundant features. By removing these features, RFE can also help to mitigate the risk of overfitting, where a classifier becomes too complex and performs poorly on new, unseen data. In our study, we used RFE to select the most important features for each of the four classifiers. By doing so, we aimed to improve the performance of the classifiers and reduce the risk of overfitting.

After performing the classification, we compared the results of each classifier by calculating the F1 and ROC AUC scores. The F1 score is a measure of the balance between precision[18] and recall[19], which are both important in classification problems. The F1 score provides a comprehensive evaluation of the performance of a classifier, considering both the number of false positive and false negative predictions (González, 2010).

[18] Precision calculates the percentage of correct positive predictions out of all the predicted positives.

[19] Recall calculates the percentage of actual positives that were correctly identified by the model or algorithm.

$$F1\ score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad \textit{Equation (5)}$$

$$precision = \frac{TP}{TP+FP} \quad \textit{Equation (6)}$$

$$Recall = \frac{TP}{TP+FN} \quad \textit{Equation (7)}$$

Where a true positive (TP) is a result that accurately confirms the presence of a certain condition. On the other hand, a false positive (FP) is a result that falsely indicates the presence of a condition, while a false negative (FN) wrongly suggests the absence of a condition that exists.

The F1 score does not take into account the ability of the model to distinguish between the positive and negative classes, which is important for evaluating the overall quality of the model's predictions. The ROC-AUC[20] score, on the other hand, measures the model's ability to distinguish between the positive and negative classes across all possible thresholds. It takes into account both the TP rate and FP rate, and provides a single value that reflects the overall quality of the model's predictions. By combining these two metrics, we get a better sense of the model's performance, especially in this thesis where the dataset is highly imbalanced.

Based on these scores, the RF classifier demonstrated superior performance with an F1 score of 0.6 and an AUC-ROC score of 0.75 (Figure 8). This indicates that the selected features exhibit strong predictive ability. Nevertheless, the differences in these scores between RF and SVM are not substantial. Therefore, it is crucial to consider the advantages of each classifier concerning their procedures, data, or other relevant factors. Notably, one advantage of RF over SVM is that RF is less sensitive to overfitting, especially when dealing with high-dimensional data. RF works by creating many decision trees on random subsets of the training data and then averaging their results. This process helps to reduce the risk of overfitting and can lead to better generalization performance on unseen data. The features selected for the classification based on RFE include the number of articles, citation count, individual discipline diversity, ethnic diversity, gender diversity, weighted degree centrality, clustering coefficient, and betweenness centrality.

[20] AUC-ROC is The area under ROC (Receiver Operating Characteristic) curve that plots the TP rate against the FP rate for different threshold values.

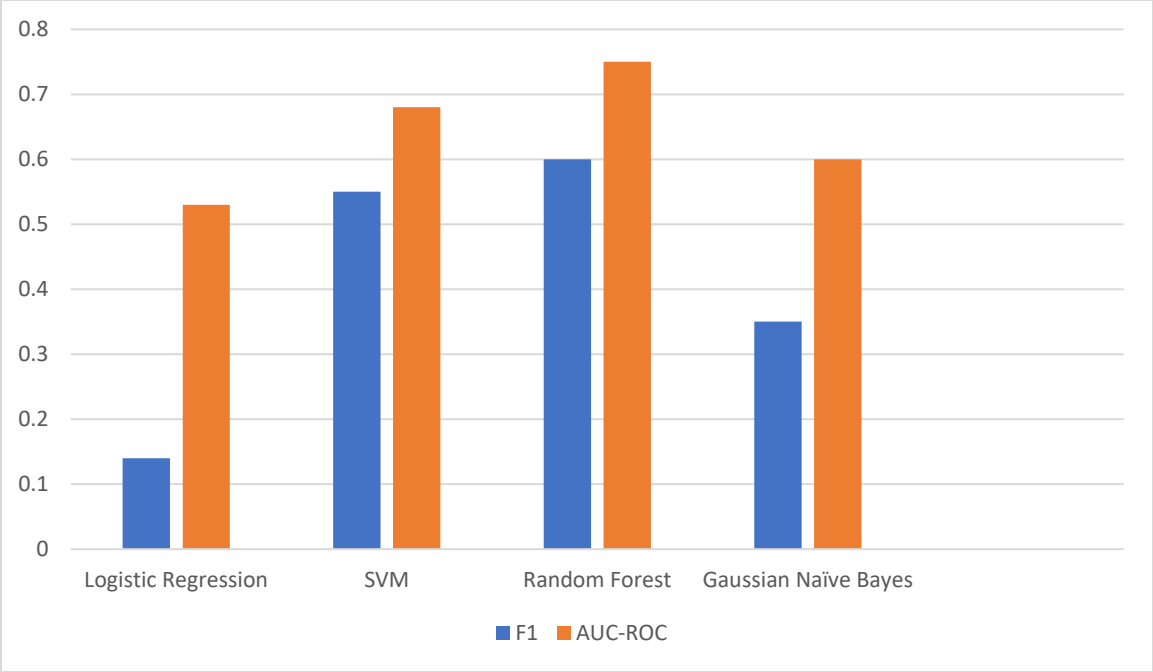


Figure 8. F1 and AUC-ROC scores of classifiers.

Our prediction involved using a combination of features in several classifiers to determine the status of authors. In the test set, we achieved satisfactory accuracy scores for identifying the status of authors using the RF classifier. This is an important accomplishment, as it allows us to understand the publishing landscape and identify rising stars more effectively. Specifically, our classifier accurately identified 36 out of 55 rising stars in the test set. This is a significant finding, as rising stars represent a group of authors who are more likely to achieve broader recognition in their field. By identifying these individuals early on, we can better support their work and foster the development of promising talent. Overall, our findings demonstrate the power of utilizing advanced ML techniques to analyze complex datasets. By leveraging a combination of features in RF classifiers, we are able to extract valuable insights and make informed decisions that can help drive scientific progress forward.

Chapter5

Conclusion and Future Work

The study focuses on the prediction of star scientists in the field of AI using ML techniques and data extracted from the Scopus database. One of the main research objectives is to determine if ML algorithms can identify early-career scientists who will become the star scientists of the AI field in the future. To answer this research question, multiple datasets including publication, citation count, and publisher data were merged and processed with NLP techniques to extract additional meaningful metadata such as author gender, discipline, and ethnicity. social network analysis, diversity measures, and research output information were used to further understand the characteristics of authors in the AI field. Only authors who published their first paper between 2006 and 2010 were considered for the study, and their features were calculated in the first five years of their careers and labeled based on the outliers of the h-index growth rate between the first and second five years of their careers.

Social network analysis and visualization were used to gain a deeper understanding of the co-authorship network among AI researchers. Moreover, the statistical analysis led to valuable insight into the AI researcher, especially rising stars. For example, a strong correlation was found between ethnicity diversity and the clustering coefficient of researchers at the beginning of their career which suggests that authors from different ethnic backgrounds tend to form tight connections and establish dense relationships, leading to a notable clustering coefficient. Moreover, a similar relation was found between gender diversity and degree centrality which suggest that authors who collaborate with a diverse group of individuals in terms of gender tend to have more connections in their network compared to those who do not. This underscores the significance of gender diversity in the realm of scientific collaboration and how it can significantly impact an author's career development and success. On other hand, the strong correlation between group discipline diversity of authors with the number of published papers indicates that a diverse research group in terms of the research background of individuals is likely to produce more published papers. In addition to the correlation test, a hypothesis test

revealed that rising stars are significantly different in all the features except ethnic diversity compared to the general population.

This work also brings a methodological contribution. Several classifiers were implemented for predictions and their performance was compared. The expanding window cross-validation and RFE combined with these classifiers, and it was shown that the RF classifier outperformed the other classifiers and that the most important features in the prediction task were the number of articles, citation count, individual discipline diversity, ethnic diversity, gender diversity, weighted degree centrality, clustering coefficient, and betweenness centrality. The combination of features from different groups of characteristics expands upon previous studies, which only evaluated the significance of each group of features.

In conclusion, this thesis contributes to the growing body of literature on the use of ML algorithms to predict the success of early-career scientists and highlights the potential of these techniques for advancing our understanding of the scientific ecosystem. The results of this study suggest that the combination of various ML techniques and NLP provides dynamic data on authors and the co-authorship network and can help predict the star scientists in the field. The findings of this research can provide valuable insights for researchers, practitioners, and funding agencies in the field of AI.

There were several limitations to this study that need to be considered when interpreting the results. Firstly, the dataset used in this study was limited in terms of both period and metadata. The period of the data may have influenced the results and a longer period may have provided more insights into the research performance of the researchers. In addition, the lack of additional metadata, such as the publisher and source of the citations, may have limited the accuracy of the research performance metric used.

Another limitation of this study was the research performance metric itself. While the metric used provided some understanding of the research performance, it was limited by the lack of complementary data about the citations. Information about cited-by articles could have helped to better understand the impact and reach of the researcher's work. For instance, self-citations, where an author cites their own work, can impact research metrics such as the h-index and other bibliometrics that measure the impact of a researcher's work. However, they are generally not considered to have the same weight as citations from other

researchers, as self-citations can inflate the apparent impact of an author's work. Considering complementary information and adjusting the research performance metric accordingly could have improved the accuracy of the research performance metric and the overall results of the study.

Finally, this study was limited to only one sub-field of academia, which is AI. By considering only one sub-field, the results may not be generalizable to other fields. It would be interesting to explore rising stars in other areas of academia to gain a more generalized insight into the factors that contribute to research success.

In order to address the limitations of this study, there are several directions for future research. Firstly, future research could consider collecting data over a longer period and including additional metadata from different databases to provide a more comprehensive understanding of the research performance. This would help to improve the accuracy of the research performance metric and the overall results of the study. Another direction for future research could be to incorporate other metrics into the research performance metric. This would provide a more accurate representation of the impact and reach of the researcher's work, which could improve the results of the study. Additionally, future research can further extend this study by considering other factors that may impact the h-index growth rate, such as the impact of funding.

References

Abbasi, A., & Altmann, J. (2011). On the correlation between research performance and social network analysis measures applied to research collaboration networks. *2011 44th Hawaii International Conference on System Sciences*, 1–10.

Abbasi, A., Altmann, J., & Hwang, J. (2010). 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.

Abramo, G., D'Angelo, C. A., & di Costa, F. (2019). A gender analysis of top scientists' collaboration behavior: evidence from Italy. *Scientometrics*, *120*(2), 405–418. https://doi.org/10.1007/s11192-019-03136-6

Ager, P., & Brückner, M. (2013). Cultural diversity and economic growth: Evidence from the US during the age of mass migration. *European Economic Review*, *64*, 76–97.

AlShebli, B. K., Rahwan, T., & Woon, W. L. (2018). The preeminence of ethnic diversity in scientific collaboration. *Nature Communications*, *9*(1), 1–10.

Anklam, P. (2003). Tapping social networks to leverage knowledge and innovation. *INFO TODAY*, 81–93.

Arcidiacono, P., Lovenheim, M., & Zhu, M. (2015). Affirmative action in undergraduate education. *Annu. Rev. Econ.*, *7*(1), 487–518.

Aydinoglu, A. U., Allard, S., & Mitchell, C. (2016). Measuring diversity in disciplinary collaboration in research teams: An ecological perspective. *Research Evaluation*, *25*(1), 18–36.

Azoulay, P., Fons-Rosen, C., & Zivin, J. S. G. (2019). Does science advance one funeral at a time? *American Economic Review*, *109*(8), 2889–2920.

Azoulay, P., Graff Zivin, J. S., & Manso, G. (2011). Incentives and creativity: evidence from the academic life sciences. *The RAND Journal of Economics*, *42*(3), 527–554.

Azoulay, P., Zivin, J. S. G., & Wang, J. (2010). Superstar extinction. *Quarterly Journal of Economics*, *125*(2), 549–589. https://doi.org/10.1162/qjec.2010.125.2.549

Banks, M. G. (2006). An extension of the Hirsch index: Indexing scientific topics and compounds. *Scientometrics*, *69*(1), 161–168.

Barabâsi, A.-L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and Its Applications*, *311*(3–4), 590–614.

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4, Issue 4). Springer.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*(4–5), 993–1022.

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., & Hwang, D.-U. (2006). Complex networks: Structure and dynamics. *Physics Reports*, *424*(4–5), 175–308.

Borgatti, S. P. (2005). Centrality and network flow. *Social Networks*, *27*(1), 55–71.

Bornmann, L. (2014). Do altmetrics point to the broader impact of research? An overview of benefits and disadvantages of altmetrics. *Journal of Informetrics*, *8*(4), 895–903.

Bornmann, L., & Daniel, H.-D. (2005). Does the h-index for ranking of scientists really work? *Scientometrics*, *65*(3), 391–392.

Bornmann, L., & Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, *66*(11), 2215–2222.

Braun, T., Glänzel, W., & Schubert, A. (2006). A Hirsch-type index for journals. *Scientometrics*, *69*(1), 169–173.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Brown, G. K., & Langer, A. (2015). Does affirmative action work: lessons from around the world. *Foreign Aff.*, *94*, 49.

Cavacini, A. (2015). What is the best database for computer science journal articles? *Scientometrics*, *102*(3), 2059–2071. https://doi.org/10.1007/s11192-014-1506-1

Chawla, N. v, Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Chung, K. S. K., & Hossain, L. (2009). Measuring performance of knowledge-intensive workgroups through social networks. *Project Management Journal*, *40*(2), 34–58.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*, 273–297.

Cronin, B., & Meho, L. (2006). Using the h-index to rank influential information scientistss. *Journal of the American Society for Information Science and Technology*, *57*(9), 1275–1278.

Daud, A., Abbasi, R., & Muhammad, F. (2013). Finding rising stars in social networks. In *Lecture Notes in Computer Science (including subseries Lecture Notes in*

Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 7825 LNCS (Issue PART 1). https://doi.org/10.1007/978-3-642-37487-6_4

Daud, A., Ahmad, M., Malik, M. S. I., & Che, D. (2015). Using machine learning techniques for rising star prediction in co-author network. *Scientometrics*, *102*(2), 1687–1711. https://doi.org/10.1007/s11192-014-1455-8

de Nooy, W., Mrvar, A., & Batagelj, V. (2005). *Structural analysis in the social sciences*. Cambridge University Press Cambridge, UK:

Deville, P., Wang, D., Sinatra, R., Song, C., Blondel, V. D., & Barabási, A.-L. (2014). Career on the move: Geography, stratification and scientific impact. *Scientific Reports*, *4*(1), 1–7.

Ebadi, A., & Schiffauerova, A. (2015a). How to become an important player in scientific collaboration networks? *Journal of Informetrics*, *9*(4), 809–825.

Ebadi, A., & Schiffauerova, A. (2015b). How to receive more funding for your research? Get connected to the right people! *PloS One*, *10*(7), e0133061.

Ebadi, A., & Schiffauerova, A. (2016). How to boost scientific production? A statistical analysis of research funding and other influencing factors. *Scientometrics*, *106*(3), 1093–1116.

Ebadi, A., Tremblay, S., Goutte, C., & Schiffauerova, A. (2020). Application of machine learning techniques to assess the trends and alignment of the funded research output. *Journal of Informetrics*, *14*(2). https://doi.org/10.1016/j.joi.2020.101018

Egghe, L. (2006). Theory and practise of the g-index. *Scientometrics*, *69*(1), 131–152. https://doi.org/10.1007/s11192-006-0144-7

Feng, S., & Kirkley, A. (2020). Mixing patterns in interdisciplinary co-authorship networks at multiple scales. *Scientific Reports*, *10*(1), 1–11.

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., Petersen, A. M., Radicchi, F., Sinatra, R., & Uzzi, B. (2018). Science of science. *Science*, *359*(6379), eaao0185.

Freeman, R. B., & Huang, W. (2015). Collaborating with people like me: Ethnic coauthorship within the United States. *Journal of Labor Economics*, *33*(S1), S289–S318.

Gershenson, C. (2014). Collaborations: The fourth age of research. *Complexity*, *19*(1).

Glanzel, W., & Persson, O. (2005). H-index for Prize medalist. *ISSI Newsletter*, *1*(4), 15–18.

González, F. A. (2010). *An Introduction to Machine Learning*.

Gray, R. M. (2011). *Entropy and information theory*. Springer Science & Business Media.

Greenberg, S. A. (2009). How citation distortions create unfounded authority: analysis of a citation network. *Bmj*, *339*.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(SUPPL. 1), 5228–5235. https://doi.org/10.1073/pnas.0307752101

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, *46*, 389–422.

Hajibabaei, A., Schiffauerova, A., & Ebadi, A. (2022). Gender-specific patterns in the artificial intelligence scientific ecosystem. *Journal of Informetrics*, *16*(2). https://doi.org/10.1016/j.joi.2022.101275

Hess, A. M., & Rothaermel, F. T. (2011). When are assets complementary? Star scientists, strategic alliances, and innovation in the pharmaceutical industry. *Strategic Management Journal*, *32*(8), 895–909. https://doi.org/10.1002/smj.916

Hess, A., & Rothaermel, F. T. (2012). Intellectual human capital and the emergence of biotechnology: Trends and patterns, 1974-2006. *IEEE Transactions on Engineering Management*, *59*(1), 65–76. https://doi.org/10.1109/TEM.2010.2082550

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(46), 16569–16572. https://doi.org/10.1073/pnas.0507655102

Hong, L., & Page, S. E. (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, *101*(46), 16385–16389.

Hoser, N. (2013). Public funding in the academic field of nanotechnology: A multi-agent based model. *Computational and Mathematical Organization Theory*, *19*(2), 253–281. https://doi.org/10.1007/s10588-013-9158-x

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

Ioannidis, J. P. A., Boyack, K. W., & Klavans, R. (2014). Estimates of the continuously publishing core in the scientific workforce. *PloS One*, *9*(7), e101698.

Jia, T., Wang, D., & Szymanski, B. K. (2017). Quantifying patterns of research-interest evolution. *Nature Human Behaviour*, *1*(4), 1–7.

Jones, B. F., & Weinberg, B. A. (2011). Age dynamics in scientific creativity. *Proceedings of the National Academy of Sciences*, *108*(47), 18910–18914.

Jones, B. F., Wuchty, S., & Uzzi, B. (2008). Multi-university research teams: Shifting impact, geography, and stratification in science. *Science*, *322*(5905), 1259–1262.

Kadhim, A. I. (2019). Survey on supervised machine learning techniques for automatic text classification. *Artificial Intelligence Review*, *52*(1), 273–292.

Kelly, C. D., & Jennions, M. D. (2006). The h index and career assessment by numbers. *Trends in Ecology & Evolution*, *21*(4), 167–170.

Larivière, V., Kiermer, V., MacCallum, C. J., McNutt, M., Patterson, M., Pulverer, B., Swaminathan, S., Taylor, S., & Curry, S. (2016). A simple proposal for the publication of journal citation distributions. *BioRxiv*, 062109.

Lee, S., & Bozeman, B. (2005). The impact of research collaboration on scientific productivity. *Social Studies of Science*, *35*(5), 673–702.

Leshner, A. I. (2003). Public engagement with science. In *Science* (Vol. 299, Issue 5609, p. 977). American Association for the Advancement of Science.

Lippi, G., & Mattiuzzi, C. (2017). Scientist impact factor (SIF): a new metric for improving scientists' evaluation? *Annals of Translational Medicine*, *5*(15).

Li, X.-L., Foo, C. S., Tew, K. L., & Ng, S.-K. (2009). Searching for rising stars in bibliography networks. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 5463). https://doi.org/10.1007/978-3-642-00887-0_25

Lowe, R. A., & Gonzalez-Brambila, C. (2007). Faculty entrepreneurs and research productivity. *Journal of Technology Transfer*, *32*(3), 173–194. https://doi.org/10.1007/s10961-006-9014-y

Manning, C., & Schutze, H. (1999). *Foundations of statistical natural language processing*. MIT press.

Moed, H. F. (2010). Measuring contextual citation impact of scientific journals. *Journal of Informetrics*, *4*(3), 265–277.

Moretti, E., & Wilson, D. J. (2014). State incentives for innovation, star scientists and jobs: Evidence from biotech. *Journal of Urban Economics*, *79*, 20–38. https://doi.org/10.1016/j.jue.2013.07.002

Nagane, H. S., Fukudome, Y., & Maki, K. (2018). An Analysis of Star Scientists in Japan. *2018 IEEE International Conference on Engineering, Technology and Innovation, ICE/ITMC 2018 - Proceedings*. https://doi.org/10.1109/ICE.2018.8436388

Nie, Y., Zhu, Y., Lin, Q., Zhang, S., Shi, P., & Niu, Z. (2019). Academic rising star prediction via scholar's evaluation model and machine learning techniques. *Scientometrics*, *120*(2), 461–476. https://doi.org/10.1007/s11192-019-03131-x

O'Boyle Jr, E., & Aguinis, H. (2012). The best and the rest: Revisiting the norm of normality of individual performance. *Personnel Psychology*, *65*(1), 79–119.

Oettl, A. (2012). Reconceptualizing stars: Scientist helpfulness and peer performance. *Management Science*, *58*(6), 1122–1140. https://doi.org/10.1287/mnsc.1110.1470

Otte, E., & Rousseau, R. (2002). Social network analysis: a powerful strategy, also for the information sciences. *Journal of Information Science*, *28*(6), 441–453.

Owen-Smith, J., Riccaboni, M., Pammolli, F., & Powell, W. W. (2002). A comparison of US and European university-industry relations in the life sciences. *Management Science*, *48*(1), 24–43.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*, 2825–2830.

Perc, M. (2010). Growth and structure of Slovenia's scientific collaboration network. *Journal of Informetrics*, *4*(4), 475–482.

Puritty, C., Strickland, L. R., Alia, E., Blonder, B., Klein, E., Kohl, M. T., McGee, E., Quintana, M., Ridley, R. E., & Tellman, B. (2017). Without inclusion, diversity initiatives may not be enough. *Science*, *357*(6356), 1101–1102.

Racherla, P., & Hu, C. (2010). A social network perspective of tourism research collaborations. *Annals of Tourism Research*, *37*(4), 1012–1034.

Radicchi, F., & Castellano, C. (2012). Testing the fairness of citation indicators for comparison across scientific domains: The case of fractional citation counts. *Journal of Informetrics*, *6*(1), 121–130.

Radicchi, F., Fortunato, S., & Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences*, *105*(45), 17268–17272.

Rousseau, R. (2006). A case study: Evolution of JASIS'Hirsch index. 19 Ağustos 2008 tarihinde http://eprints. rclis. org/archive/00005430/01. *Evolution_of_h_JASIS_rev. Pdf Adresinden Erişildi*.

Sá, C., Cowley, S., Martinez, M., Kachynska, N., & Sabzalieva, E. (2020). Gender gaps in research productivity and recognition among elite scientists in the U.S., Canada, and South Africa. *PLoS ONE*, *15*(10 October). https://doi.org/10.1371/journal.pone.0240903

Schiffauerova, A., & Beaudry, C. (2011). Star scientists and their positions in the Canadian biotechnology network. *Economics of Innovation and New Technology*, *20*(4), 343–366. https://doi.org/10.1080/10438591003696886

Sen, P. C., Hajra, M., & Ghosh, M. (2020). Supervised classification algorithms in machine learning: A survey and review. In *Emerging technology in modelling and graphics* (pp. 99–111). Springer.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423.

Siler, K., Lee, K., & Bero, L. (2015). Measuring the effectiveness of scientific gatekeeping. *Proceedings of the National Academy of Sciences*, *112*(2), 360–365.

Sonnenwald, D. H. (2007). Scientific Collaboration: a Synthesis of Challenges and Strategies, en: Cronin, B. *Annual Review of Information Science and Technology*, *41*.

Sood, G., & Laohaprapanon, S. (2018). Predicting race and ethnicity from the sequence of characters in a name. *ArXiv Preprint ArXiv:1805.02109*.

Staudt, C. L. (2011). Analysis of scientific collaboration networks: Social factors, evolution, and topical clustering. *University of the State of Baden-Wuerttemberg*.

Stringer, M. J., Sales-Pardo, M., & Amaral, L. A. N. (2010). Statistical validation of a global model for the distribution of the ultimate number of citations accrued by

papers published in a scientific journal. *Journal of the American Society for Information Science and Technology*, *61*(7), 1377–1385.

Tartari, V., Perkmann, M., & Salter, A. (2014). In good company: The influence of peers on industry engagement by academic scientists. *Research Policy*, *43*(7), 1189–1203. https://doi.org/10.1016/j.respol.2014.02.003

Trippl, M., & Maier, G. (2011). Star scientists as drivers of the development of regions. In *Advances in Spatial Science* (Vol. 66). https://doi.org/10.1007/978-3-642-14965-8_6

Tsatsaronis, G., Varlamis, I., Torge, S., Reimann, M., Nørvåg, K., Schroeder, M., & Zschunke, M. (2011). How to become a group leader? or modeling author types based on graph mining. *International Conference on Theory and Practice of Digital Libraries*, 15–26.

Tukey, J. W. (1977). *Exploratory data analysis* (Vol. 2). Reading, MA.

Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical combinations and scientific impact. *Science*, *342*(6157), 468–472.

van Leeuwen, T., Visser, M., Moed, H., Nederhof, T., & van Raan, A. (2003). The Holy Grail of science policy: Exploring and combining bibliometric tools in search of scientific excellence. *Scientometrics*, *57*(2), 257–280.

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, *7*(1), 1–8.

Wagner, C. S., & Jonkers, K. (2017). Open countries have strong science. *Nature*, *550*(7674), 32–33.

Wagner, C. S., & Leydesdorff, L. (2005). Network structure, self-organization, and the growth of international collaboration in science. *Research Policy*, *34*(10), 1608–1618.

Wetherell, C., Plakans, A., & Wellman, B. (1994). Social networks, kinship, and community in Eastern Europe. *The Journal of Interdisciplinary History*, *24*(4), 639–663.

Witten, I. H., Frank, E., Hall, M. A., Pal, C. J., & DATA, M. (2005). Practical machine learning tools and techniques. *Data Mining*, *2*(4).

Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *Science*, *330*(6004), 686–688.

Wouters, P. (2014). The citation: From culture to infrastructure. *Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact*, 47–66.

Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science*, *316*(5827), 1036–1039.

Yin, L., Kretschmer, H., Hanneman, R. A., & Liu, Z. (2006). Connection and stratification in research collaboration: An analysis of the COLLNET network. *Information Processing & Management*, *42*(6), 1599–1613.

Zhang, C., Liu, C., Yu, L., Zhang, Z.-K., & Zhou, T. (2017). Identifying the academic rising stars via pairwise citation increment ranking. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 10366 LNCS*. https://doi.org/10.1007/978-3-319-63579-8_36

Zhang, H. (2004). The optimality of naive Bayes. *Aa*, *1*(2), 3.

Zucker, L. G., Darby, M. R., & Brewer, M. B. (1998). Intellectual Human Capital and the Birth of U.S. Biotechnology Enterprises. *American Economic Review*, *88*(1), 290–306.