Discovering the Evolution and Co-authorship Patterns of Artificial Intelligence in Cancer Research using Machine Learning and Link Prediction

Shahab Mosallaie

A Thesis

In the Department of

Concordia Institute for Information Systems Engineering

Presented in Partial Fulfillment of the Requirements

For the Degree of

Master of Applied Science in Quality Systems Engineering

at Concordia University

Montréal, Quebec, Canada

December 2021

© Shahab Mosallaie, 2021

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: Shahab Mosallaie

Entitled: Discovering the Evolution and Co-authorship Patterns in Cancer Research Using Machine Learning and Link Prediction.

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

Master of Applied Science (**Quality Systems 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. J. Y. Yu

_____ Examiner
Dr. D. Terekhov

_____ Examiner
Dr. J. Y. Yu

_____ Supervisors
Dr. A. Schiffauerova

_____ Supervisors
Dr. A. Ebadi

Approved by _____
Dr. A. Ben Hamza, Director
Concordia Institute for Information Systems Engineering

_____
Dr. M. Debbabi, Dean
Faculty of Engineering and Computer Science

Date: _____

Abstract

Discovering the Evolution and Co-authorship Patterns of Artificial Intelligence in Cancer Research using Machine Learning and Link Prediction

Shahab Mosallaie

Applications of artificial intelligence play an increasingly important role in diagnosing and treating cancer. The complexity of this dynamic research field requires scientists coming from various backgrounds to continuously collaborate and create multi-disciplinary teams. However, finding the potential collaborator(s) for high-quality research effectively and efficiently is considered a difficult task for many stakeholders. In this thesis, we address these problems through developing co-authorship predictive model not only to predict potential co-authorships but also to interpret and explain which factors play essential roles in selecting potential collaborations. Finding these factors may help policymakers and research organizations investigate drivers as well as constraints in order to facilitate fruitful research collaboration and the formation of strong research teams.

The thesis has two research objectives. The first one is to characterize and map the recent research landscape in the field of artificial intelligence applications for cancer diagnosis and treatment. The second one is to explore the driving factors for different co-authorship patterns of researchers working in this field. We first used NLP techniques to characterize the evolution of artificial intelligence in the cancer research area. We observed great interest of researchers who have been gradually moving from conventional to advanced learning techniques. We then employed complex network analysis with co-authorship as a proxy for collaboration and constructed several co-

authorship networks of researchers working in this field. We extracted different structure-based and attribute-based metrics related to the individual authors and their collaboration patterns. Finally, we used machine learning and interpretability techniques to predict collaboration links of co-authorship patterns and to interpret the machine learning models. We were able to successfully predict future co-authorship links for various collaboration patterns, namely new co-authorships, persistent co-authorships and discontinued co-authorships. In general, our results show that common neighbors-based and discipline similarity factors have a positive impact on the appearance of co-authorship links. We conclude that using machine learning models and interpretability techniques is a useful and effective way to predict potential co-authorships and derive driving factors for collaboration.

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my supervisors, Dr. Andrea Schiffauerova and Dr. Ashkan Ebadi, for their guidance and feedback throughout this project. This work would not have been possible without their endless support and commitment during this journey.

Finally, my deep heartfelt gratitude goes to my parents, my sister, Sepideh, my niece, Nila, who have always supported me and have inspired me to move forward at most difficult times. They are far and yet very near.

Table of contents

List of figures

List of tables

List of acronyms

Acronym	Definition
AI	Artificial Intelligence
AP	Average Precision
AUC	Area Under the Receiver Operating Characteristic Curve
DTM	Dynamic Topic Modeling
LDA	Latent Dirichlet Allocation
ML	Machine Learning
NLP	Natural Language Processing
SHAP	SHapley Additive exPlanations
XGBoost	Extreme gradient boosting

Chapter 1 Introduction

1.1 Background and motivation

Cancer was identified as the second reason for death worldwide in 2018, with more than 9.5 million reported deaths. Despite being a deadly disease, many patients can be cured if diagnosed in the early stages (World Health Organization 2018). Early diagnosis and treatment of cancer can impede the disease's development and also improve outcomes (Wardle et al. 2015). This highlights the importance of cancer diagnosis strategies and processes to save many patients' lives from this deadly disease (Cho et al. 2014).

Recent developments in information technology have led to increased medical digital data volume (Pramanik et al. 2017). Due to the enormous healthcare data volume, the conventional analytic methods seem to be not very efficient anymore (Mehta et al. 2019). Hence, researchers have been employing more advanced approaches like artificial intelligence (AI), machine learning (ML), and deep learning to process, analyze and interpret massive data (Abbod et al. 2007; Dilsizian and Siegel 2014; Jiang et al. 2017). AI-based techniques are attracting increasing interest by researchers in the medical domain including cancer research (Jiang et al. 2017). According to a comprehensive bibliometric study conducted by Xuan Tran et al. (2019), the number of publications on AI applications in the medical domain has tripled from 2016 to 2019. They also identified that the highest interest has been in cancer research. These developments have led to significant scientific advancements in cancer diagnosis and treatment, such as imaging diagnosis, pathological diagnosis, oncology surgery, oncology radiotherapy, drug development and cancer

screening (Shen and Fu 2018). Computer science algorithms embedded in decision support systems can help physicians diagnose cancer faster and even more accurately, improving cancer treatment and patients' outcomes (Abbod et al. 2007). Characterization of the research landscape of AI-based techniques for cancer diagnosis and treatment is considered to be essential. Since it motivates research and development by facilitating our understanding of current techniques and the new trends in this field, it also helps us understand the evolution of this research area (Tran et al. 2019).

Researchers characterized cancer care as a sophisticated, but poorly organized and fragmented process (Bidassie et al. 2017; Brazil et al. 2004; Husain et al. 2013). Effective and high-quality cancer care requires health care professionals to function as a multidisciplinary team (Bidassie et al. 2017; Morgan et al. 2012; Renshaw 2007; Sayed et al. 2013). Therefore, the collaboration of researchers and professionals with different disciplines, skills, and backgrounds is required for different cancer care processes, such as cancer diagnosis and treatment (Knoop et al. 2017; Mosallaie et al. 2021).

Studies conducted by researchers show that collaboration leads to improved health outcomes, which are beneficial to patients suffering from diseases (Bosch and Mansell 2015; Greene and Cross 2017; Morgan et al. 2020). Moreover, collaboration facilitates the transfer of knowledge, information sharing and improved decision-making (Morley and Cashell 2017). Some studies attribute scientific discovery to the collaboration between individual researchers (Bougrain and Haudeville 2002; Sonnenwald 2007a). Researchers also indicate collaboration's positive impact on research productivity (Katz and Martin 1997).

The accessibility of bibliometric data has led to researchers and policy makers considering co-authorship as a proxy for research collaboration (Ponomariov and Boardman 2016). Besides that, the average number of co-authors per publication has increased over the past century (Fanelli and Larivière 2016; Henriksen 2016; Parish et al. 2018). These are some of the critical reasons that have motivated researchers to view co-authorship as the outcome of scientific research collaboration (Ponomariov and Boardman 2016; Subramanyam 1983).

Even though collaboration has many benefits, finding the potential collaborator(s) is considered a challenging task for researchers (Pavlov and Ichise 2007; Yu et al. 2014). For example, one of the greatest concerns of individual authors is how to find a suitable collaborator. This highlights the importance of co-authorship link predictors to predict future collaborations. Over the last decades, with the rise of Netflix[1], Amazon[2] and YouTube[3], recommender systems have played an important role to suggest movies, products, videos to users. The recommender systems facilitate the complexity of decision-making for their users. Finding new collaborations are often quite sophisticated and co-authorship link predictors can be incorporated in recommender systems to help authors to find potential collaborators (Aslan and Kaya 2019).

Even though co-authorship link predictors are considered as valuable tools, studying driving factors for the predicted co-authorship links and research collaboration in general is essential since it helps us understand how different factors drive collaboration behavior. These findings would help policymakers and research organizations look into drivers as well as constraints to improve

[1] Netflix is a subscription streaming service that offers a library of movies and television series. https://www.netflix.com/
[2] Amazon is an e-ecommerce company that sells a wide variety of products. https://www.amazon.com/
[3] YouTube is an online video sharing platform. https://www.youtube.com/

the research collaboration (Cheng et al. 2013). There has been significant work in understanding the process of research collaboration and the factors that drive it. Thakur, Wang, and Cozzens (2011) identified different driving factors for research collaboration at the macro and micro levels.

In the first part, we leveraged Dynamic Topic Modeling (DTM) and natural language processing (NLP) techniques to characterize and map the recent AI applications landscape for cancer diagnosis and treatment and analyze the evolution of this field. We analyzed more than 12,000 scientific publications (extracted from Elsevier's Scopus and PubMed) of researchers with the main focus on AI application for cancer diagnosis and treatment. To the best of our knowledge, this is the first research that uses multiple data sources and considers the temporal aspect of data in analyzing AI in cancer publications.

In the second part, we constructed several co-authorship networks of researchers with the main focus of AI in cancer research. We used link prediction and ML techniques to predict the existence and absence of possible links related to different co-authorship patterns. We finally identified driving factors for different co-authorship patterns by interpreting the constructed ML models. To the best of our knowledge, this is the first study that considers different co-authorship patterns to predict the existence and absence of possible co-authorship links. Moreover, we used an interpretability technique to identify driving factors for each co-authorship pattern.

We organize the thesis as follows. Each part of the thesis discussed above represents one research objective. We define the research objectives in the next section. Chapter 2 presents the literature review of the subjects corresponding to our first and second research objectives. Chapter 3

discusses the data used throughout this research — Chapter 4 and Chapter 5 present the methodology, results and discussion of the first and second research objectives, respectively. Finally, in Chapter 6, the conclusions, the limitations, and the future research are provided.

1.2 Research objectives

The research objectives for this research are as follows:

Objective 1*: Characterize and map the recent research landscape of AI applications for cancer diagnosis and treatment*

- Analyze the AI in cancer publications trend in 21st century
- Analyze the trend of most frequent computer science algorithms for cancer research
- Extract and uncover latent research topics of AI in cancer and analyze their evolution and trend over time

Objective 2*: Explore driving factors for different co-authorship patterns of researchers working in the field of AI in cancer research*

- Predict new, persistent, and discontinued co-authorship links as follows:
    - New co-authorship pattern: co-authorship links between authors who have not previously collaborated on a joint publication
    - Persistent co-authorship pattern: The continuity of co-authorship links between authors who have previously co-authored a joint paper(s)

    - Discontinued co-authorship pattern: The discontinuity of co-authorship links between authors who have previously co-authored a joint paper(s)

- Identify and interpret driving factors for new, persistent, and discontinued co-authorship patterns

Chapter 2 Literature review

This chapter provides the related literature relevant to the objectives of this research. In the first section of this chapter, we explore the literature that analyzed the research landscape of AI applications for cancer diagnosis and treatment and we compare previous works. In the second section, we investigate the literature that explored driving factors for co-authorship.

2.1 Characterization of recent research landscape of AI applications for cancer diagnosis and treatment

The number of publications and their diversity has been increasing at a growing rate (Tran et al. 2019b). A comprehensive understanding of any research area's past and recent trends can be beneficial to researchers and stakeholders working in that field. It facilitates our understanding of what topics researchers were working on and how these topics have evolved. A comprehensive understanding of any research area also helps us concentrate on essential problems (Johri et al. 2011). Empirical efforts, including interviews and surveys to understand a research area faced many challenges including bias and unreliability (Johri et al. 2011). Due to the inefficiency of empirical and traditional methods to analyze this massive amount of publications data, researchers leverage data mining to uncover information hidden in the scientific literature (Nie and Sun 2017).

Text mining is a field of research and a subset area of data mining that analyzes and processes textual data (Choudhary et al. 2009). This research area which emerged in the late '80s analyzes

large amounts of textual data in order to discover hidden information and research trends (Hearst 1999; Kostoff et al. 1999; Kostoff et al. 2000). Therefore, researchers can leverage text mining and bibliometric data to uncover research trends in different research fields (Viator and Pestorius 2001).

There has been a vast literature showing the use of text mining approaches by researchers to uncover the research trends in many fields. Viator and Pestorius (2001) used a textual mining software named Technology Opportunities Analysis of Scientific Information System (TECH OASIS) to analyze textual data of scientific publications related to acoustic research from 1970 to 1999. TECH OASIS is used for various text mining tasks, including but not limited to counting the occurrence of particular terms in scientific publications, providing a list of most frequent authors and organizing research topics. Their objective was to uncover acoustic research trends. After analyzing the scientific publications with the software, they identified a shifting characteristic of acoustic research from 1970 to 1999 in four areas, including 1) the US versus non-US affiliations, 2) research areas by year 3) research areas by world region, 4) breadth of coverage of the Journal of the Acoustical Society of America (JASA) in three acoustic areas in 1999. Perez-Iratxeta et al. (2007) used text mining to explore the research trends in bioinformatics. They extracted publications related to bioinformatics from the Medline database using a custom query that involved various bioinformatics keywords from 1996 to 2005 and only considered the abstract of publications for textual analysis. They analyzed and compared the frequency of bioinformatics terms in the literature. They observed that microarray analysis was one of the widely used topics by the bioinformatics community from 1996 to 2005.

With the advancement of the NLP techniques including topic modeling, researchers started using different topic modeling approaches for text mining. Topic modeling is a popular area of research used recently by researchers to uncover hidden topics in textual documents (Vayansky and Kumar 2020). Topic modeling approaches were used to explore research topics and trends in different research areas such as transportation (Sun and Yin 2017), management research (Hannigan et al. 2019), communication research (Maier et al. 2018), marketing (Reisenbichler and Reutterer 2019), hydropower (Jiang et al. 2016), and smart factory (Yang et al. 2018). For instance, Johri et al. (2011) used topic modeling to discover emerging and growing topics in engineering education for the period of 2000-2008. They used Latent Dirichlet Allocation (LDA) (Blei et al. 2003) as a topic modeling approach to extract topics and their corresponding top 20 keywords in engineering education. They also extracted key phrases and their corresponding frequency values to understand their frequency trends over time. According to their results, some topics such as the global and interaction aspect of engineering education experienced a significant spike, whereas other topics remained constant over time. Ayele and Juell-Skielse (2020) explored self-driving cars' evolving topics and trends. They extracted 5425 publications with the focus of self-driving cars research from Scopus within 2000-2019. Unlike researchers who used LDA as a topic modeling approach, they used DTM . The reason for selecting DTM was that LDA does not consider the temporal aspect of topics (how topics have evolved). In contrast, DTM is a probabilistic time-series model to analyze the evolution of topics over time (Blei and Lafferty 2006). The result of their study showed the evolution of twenty topics related to self-driving cars, including but not limited to software system architecture and design, brake system and safety and navigation in self-driving.

Researchers have conducted systematic studies that show the supremacy of AI-based techniques in cancer research for different types of cancer (Bashiri et al. 2017; Jalalian et al. 2013; Lisboa and

Taktak 2006; Sadoughi et al. 2018; Spelt et al. 2012). Tran et al. (2019) were the first researchers that leveraged text mining to investigate all applications of AI in cancer care and analyzed these approaches' trends over time. They extracted publications focusing on AI in cancer care from the Web of Science (WOS) from 1991 to 2018. Their study focused on four main parts as follows:

1) Co-occurrence analysis of publications keywords with the most frequent groups of terms.
2) Content analysis, listing of the top 50 emerging research domains for AI in cancer care.
3) Content analysis using LDA topic modeling approach (Blei et al. 2003), uncovering and labeling research topics extracted from the abstract section of publications.
4) Dissimilarity analysis of research disciplines of AI in cancer care

Their result showed the growing trend of publications over recent years and the expansion of multidisciplinary approaches resulting from machine learning, artificial neural network and AI in clinical practices. Moreover, the extracted research topics indicate that the development of AI in cancer care is concentrated on enhancing the prediction in cancer screening, AI-based therapeutics and personalized medicine. Their results show that some areas of AI in cancer care, including cancer outcomes and survivorship topic has recently received less attention compared to other research areas due to the growing cancer survivor population that results from progresses in early detection and treatment of cancer (Miller et al. 2019). Their result provided a supplementary study to their previous research on AI in medicine (Tran et al. 2019a). Compared to past studies that only investigated AI applications for specific types of cancer (Bashiri et al. 2017; Jalalian et al. 2013; Sadoughi et al. 2018), their study provided a relatively comprehensive view in describing the research trends of AI in cancer care.

2.1.1 Identified open research problems in characterization of research landscape of AI applications for cancer diagnosis and treatment

Although many researchers investigated AI applications for different types of cancer (Bashiri et al. 2017; Jalalian et al. 2013; Sadoughi et al. 2018; Tran et al. 2019), they did not consider the temporal aspect in analyzing AI in cancer research areas. In other words, they did not show how these research topics have evolved. The investigation of topic evolution is essential to accelerate research and development in cancer. It allows researchers to understand what methodologies, tools and techniques were employed by researchers at different periods. Moreover, researchers did not consider multiple data sources to characterize AI in cancer research that might result in less comprehensive understanding of this field.

2.2 Exploration of driving factors for different co-authorship patterns

Scientific collaboration is considered an integral part of academic research (Fonseca et al. 2016). It positively impacts research productivity and is often attributed to scientific discovery and innovation (Bougrain and Haudeville 2002; Katz and Martin 1997; Sonnenwald 2007b). Researchers are considered dependent players who collaborate to address different scientific problems that often involve multidisciplinary approaches (Sonnenwald 2007b). Collaboration of researchers results in increased sharing of ideas, resources and information, which often leads to new knowledge and innovation while reducing cost and increasing efficiency (Fonseca et al. 2016; Smith and Sotala 2011).

Researchers consider co-authorship as a proxy for collaboration (Katz and Martin 1997), the two main reasons being the easy accessibility of bibliometric data (Cronin et al. 2003) and the

important role of co-authored papers in scientific development (Acedo et al. 2006). Many studies state that intellectual collaboration is the result of co-authorship in which the collaboration between authors leads to a scientific output with higher quality than could be achieved by an individual author (Bandodkar and Grover 2016). Co-authorship has experienced a growing trend in all fields (Cummings and Kiesler 2005; Fanelli and Larivière 2016; Henriksen 2016; Ioannidis 2008; Parish et al. 2018). For instance, Henriksen (2016) investigated the rise in co-authorship in social sciences over 34 years. They considered 4.5 million peer-review articles published from 1980 through 2013 and indexed in 56 subject categories of the Web of Science's Social Citation Index. They saw an increase in the average number of authors and co-authorships for most subject categories.

The research community interested in the driving factors for co-authorship focused on two main standpoints: driving factors based on social characteristics and driving factors extracted from co-authorship networks. These two standpoints are not separable, meaning that the latter standpoint might be viewed as the result of the former standpoint. In other words, the assumptions for quantitively calculating different driving factors have roots in their social, intellectual, and economic definition.

2.2.1 Driving factors based on intellectual, economic and social interactions

In general, research collaboration (whether it results in co-authorship or not) at a macro level has three major driving factors: intellectual, economic and social (Thakur et al. 2011). Intellectual requirements are often represented by specialization and multi-disciplinary research collaboration (Katz and Martin 1997). Researchers emphasize the importance of economic driving factors characterized as different forms including huge funding resulting in more collaboration and

sharing resources (Hwang 2008). Researchers consider social interactions as an essential driving factor that results in knowledge and career advancement. For instance, Thakur et al. (2011) provided an example for social driving factors such as junior researchers collaborating with the senior researchers to be able to take advantage of resources (e.g., equipment, data, resources) and also the prestige of their work that results from them collaborating with high ranking researchers. Also, some researchers proposed that the social status of researchers is one of the driving factors of collaboration. For instance, highly productive authors might prefer to work with the authors of the same productivity levels (Glänzel and Schubert 2006). Thakur et al. (2011) stated that at a micro-level, collaboration might result from different factors that depend on the context of collaboration. As shown in Table 1, Beaver and Rosen (1978) provided different motivations for collaboration at a micro-level.

Table 1. Micro-level motivations for research collaboration (Beaver and Rosen 1978)

1	Access to expertise
2	Access to equipment, resources
3	Improve access to funds
4	To obtain prestige or visibility; for professional advancement
5	Efficiency: multiple hands and minds
6	To make progress more rapidly
7	To tackle complicated problems (more critical, more comprehensive, more complex and global)
8	To enhance productivity
9	To get to know people, to create network

10	To learn new skills and techniques, usually to break into a new field, subfield, or problem
11	To satisfy the curiosity, inetellectual interest
12	To share the excitement of an area with other people
13	To find flaws more efficiently, reduce errors and mistakes
14	To keep one more focused on research because researchers are counting on each other
15	To reduce isolation and to increase the researcher's energy and excitement
16	To educate a student, graduate student or oneself
17	To advance knowledge and learning
18	To increase the pleasure

Ponomariov and Boardman (2016) conducted a comprehensive review on co-authorship and driving factors affecting co-authorship. They categorized the driving social factors into two main groups, resource-based and non-resource-based predictors. The resource-based predictors address the resource or capital contribution of collaborators as a good reason for collaboration. In contrast, non-resource-based relations are purely relational. In other words, non-resource-based relations characterize relationships among collaborators and distinguish research collaborations from each other. They stated that even though the collaboration predictors are categorized into different groups, they are related and not separable.

Ponomariov and Boardman (2016) categorized the resource-based predictors into formal and informal predictors and formulated different hypotheses for these predictors. Finally, they validated these hypotheses against a survey data of researchers. For instance, for a formal resource-based relation, they hypothesized that a mentoring relationship (e.g., thesis advisor and student) increases the probability of co-authorship. However, their validation result provided weak support for that by showing that the student-advisor relationships might be important at the beginning of

the scientific career age of authors. However, over time the co-authorship relationship fades away. Similarly, their hypothesis that stated "Authors who are close to each other in terms of the length of relationship and trust outside of their profession are more likely to collaborate" showed to be only true for authors with low and medium productivity.

Other researchers also investigated the impact of the authors' characteristics on co-authorship. For instance, Gallivan and Ahuja (2015) investigated the impact of homophily[4] of proximity, gender and geography on co-authorship among Information Systems (IS) researchers. They analyzed publications of high-ranking IS journals. Based on the result, they observed that IS researchers worldwide tend to collaborate with co-authors of the same sex and the same Ph.D. program. Thakur, Wang, and Cozzens (2011) narrowed down the identification of driving factors for co-authorship to international co-authorship in the biofuels field. They interviewed a range of biofuel researchers. Their interview results showed that different factors drive international co-authorship in the biofuel field, including intellectual, economic and social motivations. The intellectual motivation often results from the need for an interdisciplinary project to leverage different expertise/specializations to accomplish its goal that may not be present in each of the individual countries. They suggested that the economic motivation behind the research project is to share different resources such as data, expansive facilities, a group of personnel and massive funding. For instance, small and developing countries seek international collaboration because of their limited resources and their need to access more resources. Finally, they viewed the social

[4] Homophily is the tendency for people to seek out or be attracted to those who are similar to themselves ("homophily : Oxford English Dictionary").

motivations behind a research project as two standpoints. In the first standpoint, they suggested that researchers seek to gain more visibility by collaborating with influencing researchers in their field. The second standpoint is the difference between collaborators' social status, such as the tendency of highly productive people to collaborate with their highly productive peers. In general, these researchers conducted interviews with authors/researchers and used survey data to identify the driving factors for co-authorship. For instance, Hwang (2008) conducted interviews with Korean and British scientists to understand the motivations that drive international collaboration. Thakur et al. (2011) interviewed a group of researchers who had published papers in the field of bio-fuels to find driving factors and motivations of co-authorship.

2.2.2 Driving factors identification using co-authorship networks

Researchers used different methods to determine driving factors for co-authorship. As discussed in the previous section, some studies conducted interviews with researchers to identify these factors. With the growth of research on complex networks including social networks, where nodes/vertices represent entities and edges/links show interaction between them, researchers started to use new techniques to understand the social networks. One of the research areas of social networks has been understanding the dynamicity of social networks (Breiger 2004). Social networks are highly dynamic objects. In other words, they evolve over time through the addition and/or removal of new nodes and edges. Co-authorship network as a special case of social network, in which nodes represent authors and edges represent co-authorship links, is not an exception. Like social networks, they are also dynamic objects. New authors and co-authorship links are added to co-authorship networks over time. Also, authors might stop collaborating with each other.

Nowell and Kleinberg (2003) first introduced the link prediction problem in social networks to understand how social networks evolve. They formulated the problem as "Given a snapshot of a social network, can we infer which new interactions among its members are likely to occur in near feature?" They considered a co-authorship graph/network of researchers and suggested that scientists who are "similar" to each other in the network are more likely to collaborate in future and similarity of authors could be a good indication of their future collaboration. Therefore, they proposed different "similarity-based metrics" to measure the similarity of authors in the co-authorship network to be able to predict the possibility of co-authorship links in future. These similarity-based metrics were adapted from techniques in graph theory and social network analysis.

Researchers introduced different similarity-based metrics to define similarity between authors in the co-authorship networks with the assumption that authors who are similar in a co-authorship network are more likely to collaborate in future (Liben-Nowell and Kleinberg 2003; Martínez et al. 2016). Researchers used these metrics to predict the possibility of collaboration for author pairs in co-authorship networks. Predicting the co-authorship links would help us to gain valuable information by understanding the driving factors of collaboration and how different co-authorship networks evolve. Moreover, the predicted co-authorship links would serve as reasonable suggestions for potential collaboration to build strong research teams (Pavlov and Ichise 2007; Yu et al. 2014). Liben-Nowell and Kleinberg (2003) used similarity-based metrics to predict the new co-authorship links in future. They did expriment on five co-authorship networks extracted from

different sections of physics e-Print arXiv[5] from 1994 to 1999. They considered the first three years as the training interval to extract different similarity-based metrics and used the latter three years for the test interval to predict the future co-authorship links between researchers "who had not collaborated before". They extracted different similarity-based metrics including but not limited to common neighbors, shortest path, preferential attachment, Adamic-Adar index and Jaccard coefficient. For instance, for common neighbors (CN) as a similarity-based metric, they assumed that author pairs with a higher number of common neighbors are more likely to collaborate in future. Shortest path (SP) was another similarity-based metric they used with the assumption that authors who are far from each other in the network are less likely to collaborate than authors who are close to each other in the co-authorship network. Other similarity-based scores were defined as different variations of common neighbors or by considering more information from the co-authorship network. In the methodology section, we will describe some similarity-based metrics that we used throughout this research,

Finally, Liben-Nowell and Kleinberg (2003) identified that most similarity-based metrics outperformed different baselines in co-authorship link prediction. Even though their work was considered successful and was the first study in the link prediction field for social networks, they considered each similarity-based metric individually and did not consider the effect of considering similarity-based features combined as a feature vector to predict the future co-authorship links. Moreover, they defined similarity-based metrics based on only the topology of the network and not the attribute of researchers. On the other hand, Pavlov and Ichise (2007) proposed an improved

[5] "arXiv is a free distribution service and an open-access archive for 1,976,918 scholarly articles in the fields of physics, mathematics, computer science, quantitative biology, quantitative finance, statistics, electrical engineering and systems science, and economics. Materials on this site are not peer-reviewed by arXiv" https://arxiv.org/

method for link prediction by considering multiple similarity-based scores as a feature vector and used supervised ML algorithms to predict the new co-authorship links. They conducted several experiments on the co-authorship network of Japanese researchers from 1993 to 2006. They extracted different similarity-based metrics from co-authorship networks including but not limited to common neighbors, Jaccard coefficient, Adamic Adar index, preferential attachment and combined them as a feature vector. They formulated the problem as a binary classification task and used different ML classifiers including Support Vector Machines (SVMs), decision trees, and AdaBoost to predict future co-authorship links. They improved the link prediction results by leveraging the ML algorithms that used the feature vector of similarity-based metrics as input. Unlike Pavlov and Ichise (2007) that used type similarity-based metrics that only consider the structure of co-authorship networks, Hasan et al. (2006) suggested that adding attribute-based similarity metrics to the feature vector of the ML algorithms significantly improves the performance of the co-authorship link prediction. For instance, they considered the size of the intersection set of the publications' keywords of a pair of authors as their similarity score and identified that authors with higher values of the keyword intersection size are similar to each other and thus are more likely to collaborate in future. Similar to the previous researchers, they constructed several ML models including but not limited to decision tree, support vector machines and k-nearest neighbors for co-authorship link prediction. Yu et al. (2014) extracted different similarity-based metrics such as common neighbors, Adamic Adar index, preferential attachment from the co-authorship networks of researchers working in the field of coronary artery disease. These features served as an input for two ML models including logistic regression and support vector machines. They were able to predict the co-authorship links with a good performance and identified that structure-based features might be good indicators or driving factors for future

collaboration and could facilitate the development of strong research teams. Similar to Pavlov and Ichise (2007), they only used structure-based metrics for co-authorship link prediction. Chuan et al. (2018) considered different similarity-based link prediction metrics and introduced a new hybrid attribute-based feature called LDAcosin to predict co-authorship links. The LDAcosin metric considers the content of authors' publications and uses a LDA topic modeling algorithm to extract the topics of publications and calculate a discipline similarity score for a pair of authors. They suggested that authors are more likely to collaborate with authors doing research within similar disciplines. They considered three co-authorship networks in three domains of physics. Later on, they constructed a support vector machine (SVM) classifier and used unweighted and weighted similarity-based and the LDAcosin metrics. They were able to successfully predict the future co-authorship links. They identified the importance of considering the content of publications (to calculate the similarity score of authors and papers) to improve the performance of co-authorship link prediction.

2.2.3 Open research problems in co-authorship link prediction

As discussed in the previous section, many researchers successfully adopted link prediction approaches and ML techniques to predict the co-authorship links. For instance, Yu et al. (2014) were able to predict approximately 70% of new co-authorship links using ML techniques and similarity-based metrics as inputs. Similarly, Chuan et al. (2018) were able to predict more than 70% of new co-authorship links using unweighted and weighted similarity-based and LDAcosin metrics. However, these researchers only predicted the possibility of collaboration between researchers who have not previously collaborated. Excluding the investigation of the continuity and discontinuity of future collaboration of authors who have previously collaborated on a joint

paper might not give us a comprehensive view to fully understand the driving factors for different co-authorship patterns. Moreover, the co-authorship predictive models can be better leveraged by co-authorship recommender systems if they give us a full picture of all possible co-authorship links leading to the development of strong research teams.

Moreover, researchers did not interpret how the ML classifiers that they used predicted the co-authorship links. In other words, they did not identify the metrics/factors that had the highest contribution for co-authorship link prediction. These metrics might be considered as reasonable indictors or driving factors for future co-authorship links (Yu et al. 2014). Furthermore, researchers emphasized the importance of interpretability in predictive models. They suggested that lack of interpretability poses several drawbacks, such as reducing the practical applicability and trust to use these models by researchers, practitioners, and policymakers (ElShawi et al. 2020; Katuwal and Chen 2016). The interpretability is vital, especially when these ML models are incorporated in decision-making by policymakers to set strategies, or they are used by researchers to locate potential co-authors and research organizations to build strong research teams.

Even though Yu et al. (2014) predicted the co-authorship links in the medical domain (Q. Yu et al. 2014), no researchers investigated the co-authorship network of researchers with the main focus of AI applications for cancer diagnosis and treatment. This research domain is essential due to the complexity of this disease and the need for collaboration of researchers from different disciplines, including but not limited to computer scientists and medical experts.

Chapter 3. Data

This chapter will describe the data used to satisfy the research objectives of the thesis. We introduce the data collection step and the data sources that we used for this research. Elsevier's Scopus and PubMed are two databases exploited in this thesis. Scopus is Elsevier's abstract and citation database launched in 2004 that includes a large number of articles across many disciplines ("About | Elsevier Scopus Blog" (2021). PubMed was launched in 1997 by National Library of Medicine (NLM) and has been identified as one of the most reliable source for researchers and clinicians specifically in the field of medicine (Falagas et al. 2008).

For the first research objective of this research which is characterizing and mapping the research landscape of AI applications for Cancer diagnosis and treatment, we considered both Scopus and PubMed data sources. In contrast, for the second research objective, we considered only Scopus because of the result obtained from the comparison of these two data sources. They have a significant overlap, with ~86% of PubMed publications being covered in Scopus. We will describe the comparison of the results in the section 4.2.1. We used the following search query to extract publications from Elsevier's Scopus and PubMed with the [2000, 2018] time interval:

("artificial intelligence" OR "machine learning" OR "deep learning" OR "neural network" OR "neural net") AND ("cancer")

We included journal articles, conference papers, book chapters, and books and collected 10,071 and 2,206 publications from Scopus and PubMed, respectively.

Chapter 4. Research landscape of AI applications for cancer diagnosis and treatment

In this section, we will describe in detail data processing steps as well as methodologies used to characterize the landscape of AI applications for cancer diagnosis and treatment within 2000-2018. Our main objective is to quantify the development of research in this field, and confirm quantitatively if the subject field has been evolving over time. We used NLP and DTM to extract hidden patterns and latent research themes and analyze their evolution. The rest of this section proceeds with "Data processing and methodology" section which describes the data and techniques in greater detail, and the findings of the research are then presented in section 4.2.1 to 4.2.3 and discussed in section 4.2.4.

4.1 Data processing and methodology

As discussed in the data chapter, using the search query, we extracted 10,071 and 2,206 publications from Scopus and PubMed, respectively. We considered three data scenarios for the analysis: 1) Scopus publications only, 2) PubMed publications only, and 3) The intersection of Scopus and PubMed publications.

We merged titles and abstracts of the collected publications, preprocessed the textual data, and used the DTM algorithm (Blei and Lafferty 2006) to extract latent research themes in each of the three data scenarios. We decided to concatenate titles and abstracts as the abstract provides a condensed representation of articles with detailed information compared to publications' titles, but titles can also contain some complementary information such as specific keywords or key phrases

about the research (Ebadi et al. 2020). We performed several preprocessing steps, including but not limited to lowercase conversion, stop-words and punctuation removal, tokenization, and lemmatization. We considered uni- and bi-grams to extract key phrases and investigated their trends over the examined period. We then used uni-grams to calculate the document-term frequency (TF) matrix that the DTM algorithm consumed to build the model. We performed intensive hyperparameter tuning on the DTM model. The DTM algorithm required the number of topics to be defined in advance. We first built several LDA (Blei et al. 2003) baseline models for each dataset to find a range for the number of topics by investigating the extracted topics' coherence. We found the best range for the number of topics to be five to eight topics. Next, we built four separate DTM models for each dataset, with five to eight topics. We assessed the quality of their topics quantitatively with the *CV* coherence score (Röder et al. 2015) and qualitatively by verifying with three domain experts. The DTM models with six topics were found to be the best for Scopus and PubMed. For the Scopus-PubMed intersection data scenario, the optimal number of topics was found to be five. In addition to the number of topics, we further tuned other hyperparameters such as chunk size.

The DTM algorithm does not assign a representative label to the extracted topics. To reduce the subjectivity impact, we used a panel of three domain experts to analyze the expanded set of keywords generated for each topic and assign a short informative label to them. Since we were interested in finding the primary "AI in cancer" research themes in this study, the experts were tasked to discover more generic and inclusive topic labels, preferably with a simple and interpretable name rather than a phrase.

The extracted topics are almost the same in the three data scenarios. The only difference between Scopus only and PubMed only scenarios is that the "clinical decision support system" is observed in the former while "colorectal cancer" is instead observed in the latter. We developed and programmed the entire analytics pipeline in the Python programming language. As shown in Figure 1, this pipeline contains three main components, i.e., data collection, preprocessing, and data analytics. In the data collection step, we extracted the publications with the main focus of "AI in Cancer" for [2000-2018]. Then, abstracts and titles of the publications were merged and preprocessed. Next, we extracted key phrases and analyze their trends over time. We developed a DTM model to extract latent research topics. Finally, the trends and evolution of the extracted research topics were investigated.

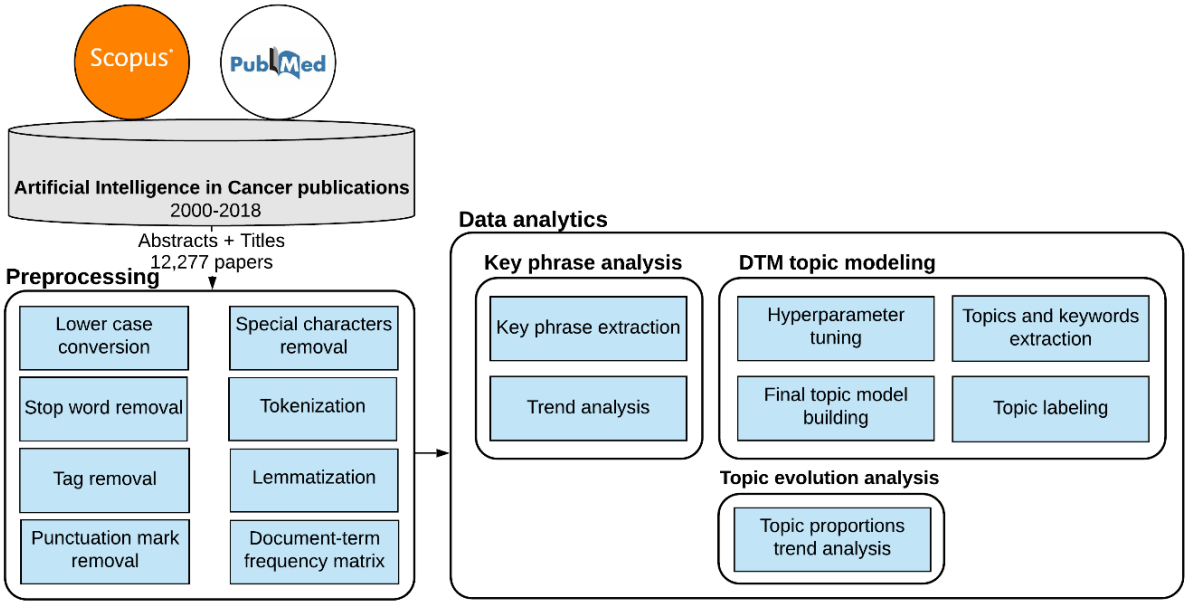


Figure 1. Conceptual flow of the analytics pipeline for research landscape of AI in cancer.

4.2. Results and discussions

4.2.1 Publications trend

The total number of publications in Scopus and PubMed datasets is 12,277 (Scopus dataset with 10,071 and PubMed dataset with 2,206 publications). The two datasets overlap significantly, with ~86% of PubMed publications being covered in Scopus. The overlap rate ranges from ~81% in 2000 to ~76% (i.e., the minimum) in 2018, with a ~95% peak in 2006. There are 331 (3.2%) and 8,196 (78.8%) unique publications in PubMed and Scopus datasets, respectively.

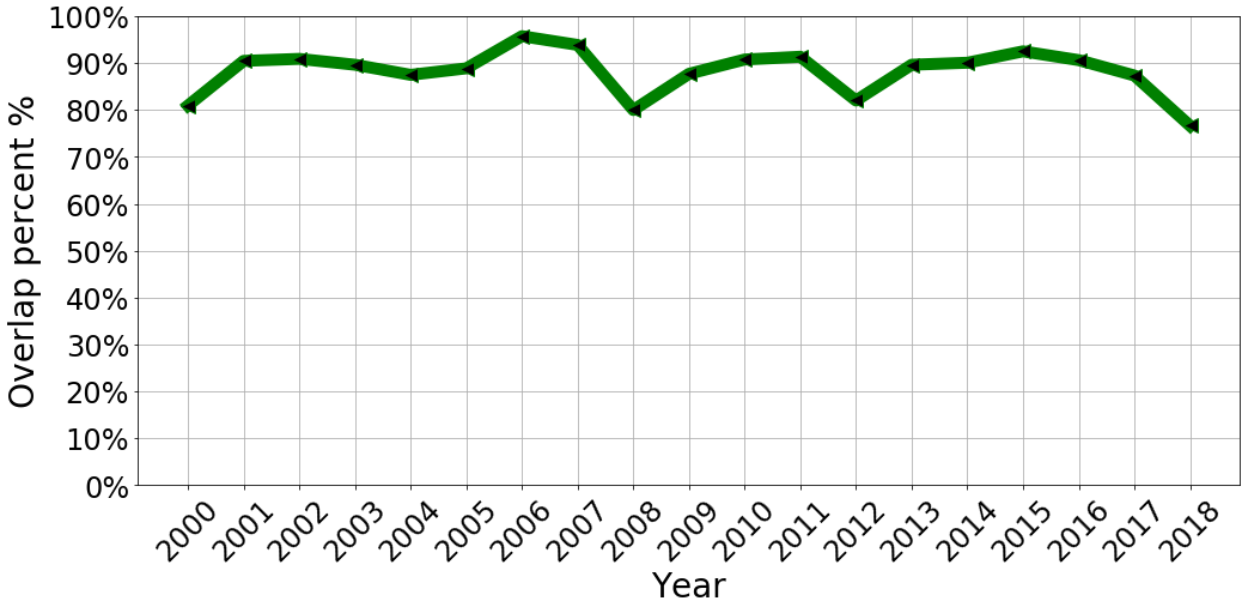


Figure 2. Scopus and PubMed overlap trend.

As shown in Figure 3, the number of Scopus and PubMed publications has increased continuously from 2000 to 2018. We saw a significant increase in the last two years. The number of Scopus publications had a ~40% and ~65% growth in 2017 and 2018. In PubMed, a growth rate of 67%

is observed in 2018. Although this could be due to the better coverage of publications in the final years, it could also partially imply researchers' closer attention to the subject topic, i.e., AI in cancer.

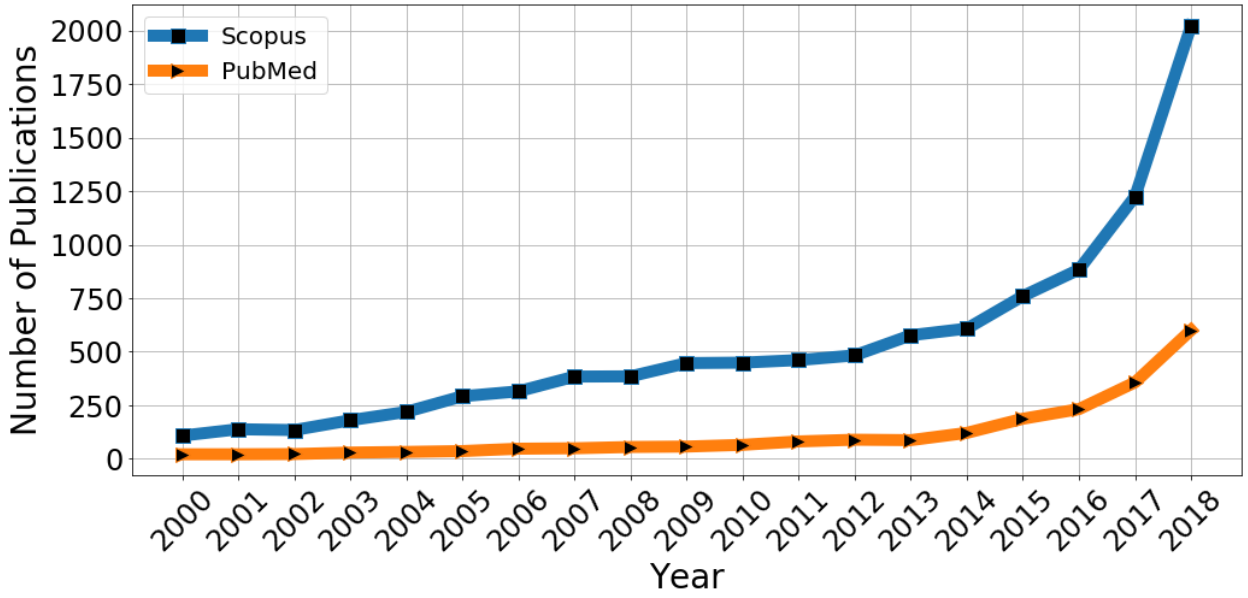


Figure 3. The number of "AI in cancer" publications extracted from Scopus and PubMed within 2000-2018.

4.2.2 Key phrase analysis

We extracted the most frequent computer science algorithms/techniques mentioned in the collected publications, i.e., machine learning, random forests, logistic regression, principal component analysis (PCA), deep learning, and convolutional neural network (CNN).

As shown in Figure 4, we analyzed proportion of these key phrases over time in Scopus and PubMed. In both Scopus and PubMed datasets, “machine learning” has the highest ratio with some

fluctuations from 2000 to 2002, especially in PubMed. “machine learning” almost experienced a constant trend from 2003 to 2015. However, it follows a decreasing trend after 2015 in both datasets. In the Scopus dataset, the “random forests” algorithm first appeared among the most frequent key phrases in 2005, and its proportion slightly increases after 2009, whereas in PubMed, it is first seen in 2009, following an almost constant trend until the final period. Additionally, in the Scopus dataset, “deep learning” and “convolutional neural network” are observed for the first time in 2012 and 2015, respectively, and their proportion increases after that. Similarly, in the PubMed dataset, we can see the appearance of “deep learning” and “convolutional neural network” key phrases in 2011, followed by a sharp increase after 2015. In both datasets, the trends for “logistic regression” have decreased over time. These observations may partially confirm the researchers active in the healthcare and cancer diagnosis domain have gradually shifted from conventional statistical analysis to more advanced computer science algorithms due to several factors such as convolutional techniques limitations in handling massive data and/or new types of data, e.g., medical images (Yu et al. 2018).

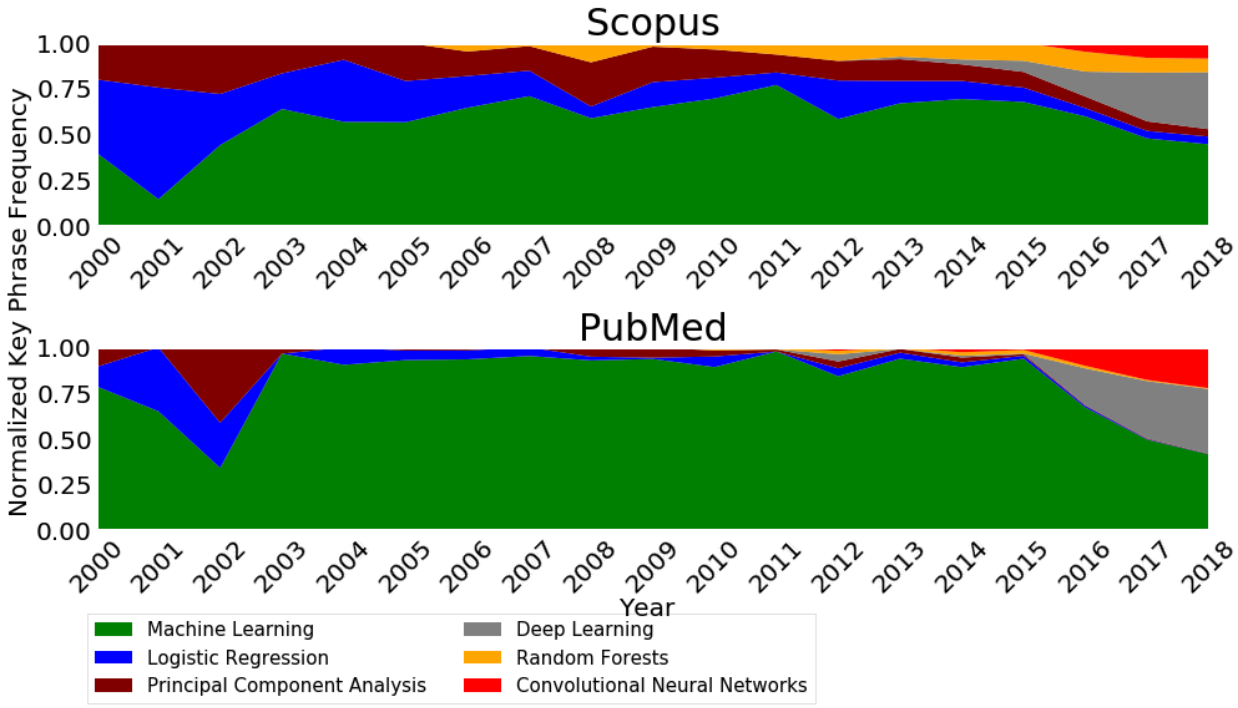


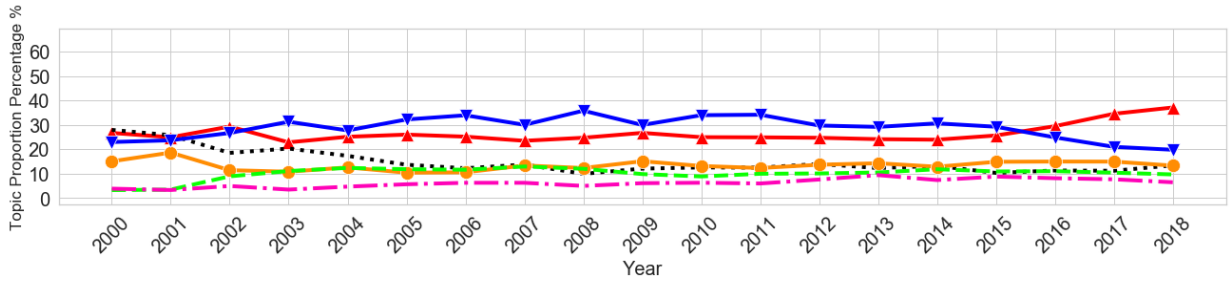
Figure 4. The temporal trend of the proportions of the top key phrases observed in Scopus and PubMed.

4.2.3 Topic evolution analysis

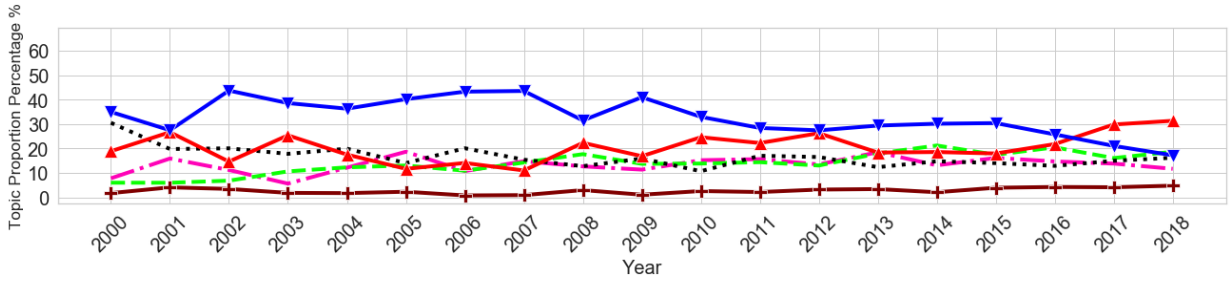
Topics prevalence in Scopus and PubMed datasets, as well as their intersection, are depicted in Figure 5. As explained in the "Data processing and methods" section, six topics were extracted from the Scopus dataset. As seen in Figure 5-a, the "machine learning" topic has been the most prevalent topic in almost all the examined period, being replaced by the "medical image analytics" topic in the last three years. This would imply researchers' increasing interest in using ML algorithms to solve complex problems such as cancer diagnosis. This finding is in line with Kourou et al. (2015).

Interestingly, after 2015, the “medical image analytics,” is represented by deep neural networks key phrases (e.g., CNNs), takes the place of the “machine learning” topic as the most prevalent topic. This confirms our findings in the previous section about applying deep learning approaches to analyze new data types such as medical images (Figure 4). Medical image analytics has been used in practice for many years, especially in computer-aided diagnosis (CAD) systems, to improve radiologists' and clinicians' performance (Anwar et al. 2018). Many publications have investigated the application of CNN algorithms for diagnosing different types of cancers such as breast cancer, lung cancer or prostate cancer (Alakwaa et al. 2017; Le et al. 2017; Zou et al. 2019). The “cancer survival” topic proportion in high at the beginning of the examined period, having an almost constant prevalence after 2008, despite some fluctuations. This might also indicate the application of AI in medical subfields. Our data also suggests that “clinical decision support system” and “cancer genomics” research themes have attracted researcher's attention, following an almost steady trend after 2002. This results from the rapid growth of healthcare data in different forms, including genomic datasets and the need for intelligent decision support systems to analyze and process the massive amount of data quickly and efficiently (Dias and Torkamani 2019; Sutton et al. 2020). The “drug design” topic’s proportion has slightly increased from 2000 to 2012, and it remained almost constant afterward. The drug development process has been influenced a lot by AI and advanced techniques. For instance, recent AI advancements have improved the drug development process, such as drug repurposing, predicting the mode-of-action of compounds, and selecting a population for clinical trials. AI contributed a higher efficiency, and lower research and development (R&D) costs to the drug development process (Mak and Pichika 2019).

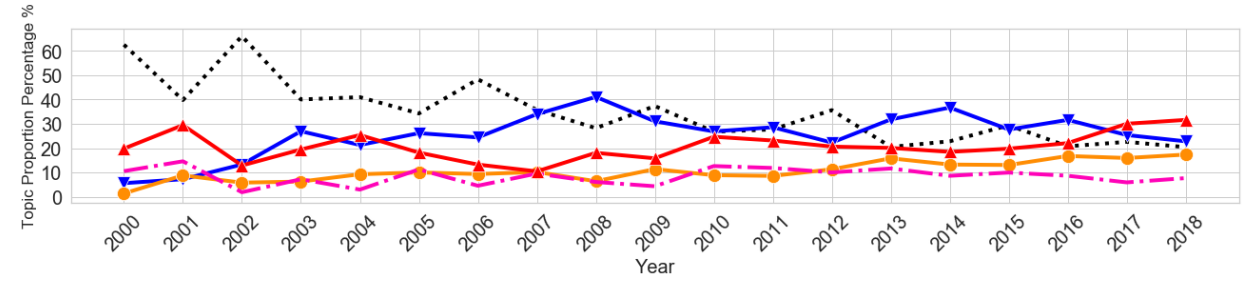
a)



b)



c)



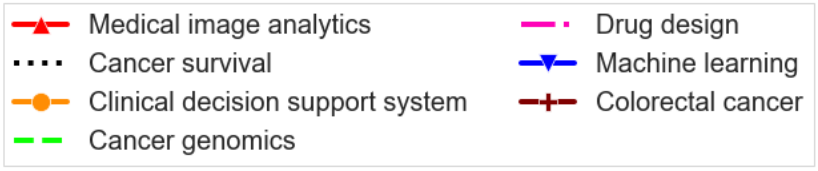


Figure 5. Topic prevalence from 2000 to 2018, a) Scopus, b) PubMed, c) the intersection of Scopus and PubMed.

Figure 5-b shows the prevalence of the topics for the PubMed dataset. Six topics were extracted from the PubMed dataset, including “medical image analytics”, “cancer survival”, “drug design”, “cancer genomics”, “machine learning” and “colorectal cancer”. Comparing Figures 5-a and 5-b, the “machine learning” and “medical image analytics” topics have been the most prevalent research themes in both Scopus and PubMed datasets, having similar patterns. The proportion of the “colorectal cancer” topic (a unique topic extracted from the PubMed dataset) remained almost constant over the examined period, being the least prevalent topic in the entire time interval. Although other topics are observed to have a similar trend to the ones in Scopus, fewer fluctuations are observed in the Scopus topics prevalence that might be due to the highest number of publications in the Scopus dataset. Figure 4-c shows the prevalence of the five extracted topics from the Scopus-PubMed intersection dataset. As seen, “cancer survival” has been the most prevalent topic in the first seven years of the examined period, losing its place to the “machine learning” and “medical image analytics” topics from time to time. Being among the top three most prevalent topics in the entire time interval, “medical image analytics” becomes the most prevalent topic in the final year.

4.2.4 Discussion

As an aggressive disease with a low median survival rate, cancer has a lengthy and costly treatment process. Early diagnosis can enhance a patient’s survival chances. Over the last years, scientists have been using statistics and computational methods to diagnose/predict the disease. With the emergence of AI and learning techniques, many scientists are applying machine/deep learning to clinical cancer research, where the performance of the cancer prediction models has been promising (Huang et al. 2020). For example, deep convolutional neural networks were shown to

improve the diagnostic accuracy of solid tumors in thyroid cancer (X. Li et al. 2019) or classification of malignant and benign masses in digital breast tomosynthesis (Zhu et al. 2019). Apart from the performance, such technologies provide unique opportunities to analyze new data types such as medical imaging. For example, AI has been applied to medical imaging such as magnetic resonance imagery (MRI) and computerized tomography (CT) scans, facilitating the analysis of new data types.

By doing a cross publication search engine study within the period of 2000-2018 and using DTM and NLP techniques, this thesis provides a concrete perspective on how scientists employ AI for cancer detection and treatment, highlighting the current research trend. Our findings suggest a shift from conventional analytic techniques towards learning techniques. Besides, medical image analytics was a prevalent theme that has been increasingly attracting researchers' attention over time. Deep learning techniques, specifically, are assisting data scientists in analyzing and interpreting imaging data more precisely (Ferroni et al. 2019; N. Jiang and Xu 2019), ensuring fewer false positives than radiologists (Huang et al. 2020).

The complexity of medical chemistry research calls for the application of emerging technologies in the design of new drugs (Workman et al. 2019). ML and AI are being used in the new drug discovery process, aiming to make it faster, cheaper and more effective (Fleming 2018). We identified "drug design" as one of the main research themes, considering that data for this type of research is available more than ever, and the increasing number of academic and industrial labs employing AI for drug design (Sellwood et al. 2018). We expect to observe this topic as one of the leading research themes in the coming years, even with continuous growth. As a deadly disease

affecting more than 18 million people worldwide annually, cancer is a disease of the genes being caused by mutations in genomes (Vert 2020). The “cancer genomics” topic was one of the most prevalent topics, especially in the PubMed dataset in the final years. Hence, the result suggests AI is assisting researchers in performing a more concrete analysis on large quantities of genome data which may result in a better understanding of the disease while adapting the treatment to the molecular characteristic of each patient (Vert 2020).

The incorporation and increasing availability of clinical decision support systems (CDSS) in healthcare settings is major progress that supports clinicians with the decision-making process that would lead to improved quality of care while minimizing the costs (Mazo et al. 2020). The CDSS topic was identified as one of the most prevalent topics in our study. Considering the potential of AI for various types of cancer diagnosis, we expect that AI-powered CDSS research will continue to grow, resulting in a paradigm shift in cancer diagnosis and treatment, as AI is foreseen to keep helping scientists to overcome the challenges of cancer diagnosis (Huang et al. 2020). This development could be facilitated by the increased availability of digital medical data and the growth of medical data scientists. Finally, the appearance of “colorectal cancer” among the topics extracted from the PubMed dataset explains the importance of this type of cancer in the eyes of the researchers. According to a recent study, colorectal cancer is the second leading cause of cancer death and the third most common cancer in men and women (Siegel et al. 2020). Hence, special attention to this field of research and continuous support would be encouraged.

Although our results show that the importance of AI is being increasingly recognized in the medical domain, a tight collaboration between computer scientists and medical experts would play

a key role in ensuring the continued success of this interdisciplinary research. This motivated us to explore the collaboration pattern between authors in this field as the second research objective in the next section.

Chapter 5. Driving factors for co-authorship

We investigated the trend and evolution of AI techniques for cancer diagnosis and treatment in chapter 4. The successful employment of these techniques requires health care professionals to function as a multidisciplinary team such as medical experts and computer scientists (Bidassie et al. 2017; P. A. Morgan et al. 2012; Renshaw 2007; Sayed et al. 2013). Therefore, the collaboration of researchers and professionals with different disciplines, skills, and backgrounds is required for different cancer care processes, such as cancer diagnosis and treatment (Knoop et al. 2017; Mosallaie et al. 2021).

The importance of collaboration motivated us to investigate the co-authorship network of researchers with the main focus on AI applications for cancer diagnosis and treatment within 2000-2017. Our objective was to first discover different co-authorship patterns by predicting their associated links including the new, persistent, and discontinued co-authorship links. Predicting the co-authorship links would help us to gain valuable information by understanding the driving factors of collaboration and how different co-authorship networks evolve. Moreover, the predicted co-authorship links would serve as reasonable suggestions for potential collaboration to build strong research teams (Pavlov and Ichise 2007; Yu et al. 2014). Finally, we added an interpretability module to identify the most driving factors. Studying driving factors for research

collaboration is essential since it helps us understand how different factors drive collaboration behavior. These findings would help policymakers and research organizations look into drivers as well as constraints to improve the research collaboration (Cheng et al. 2013). Our proposed driving factors shed light on the dynamics of future collaboration by identifying various co-authorship patterns.

Section 5.1 first analyzes the trend of publications, number of authors and number of collaborations of the publications of AI in cancer area over the recent years. In section 5.2 we describe the methodologies used throughout this chapter. Finally, section 5.3 concludes the chapter with the results and discussion.

5.1 Data trend

As discussed in chapter 3, we used a group of keywords as a search query to extract publications with the main focus of AI applications in cancer diagnosis and treatment from Scopus. The result was 7,738 publications for the period of 2000 to 2017. Figure 6 shows that the publications trend has been growing at an increasing rate over recent years.

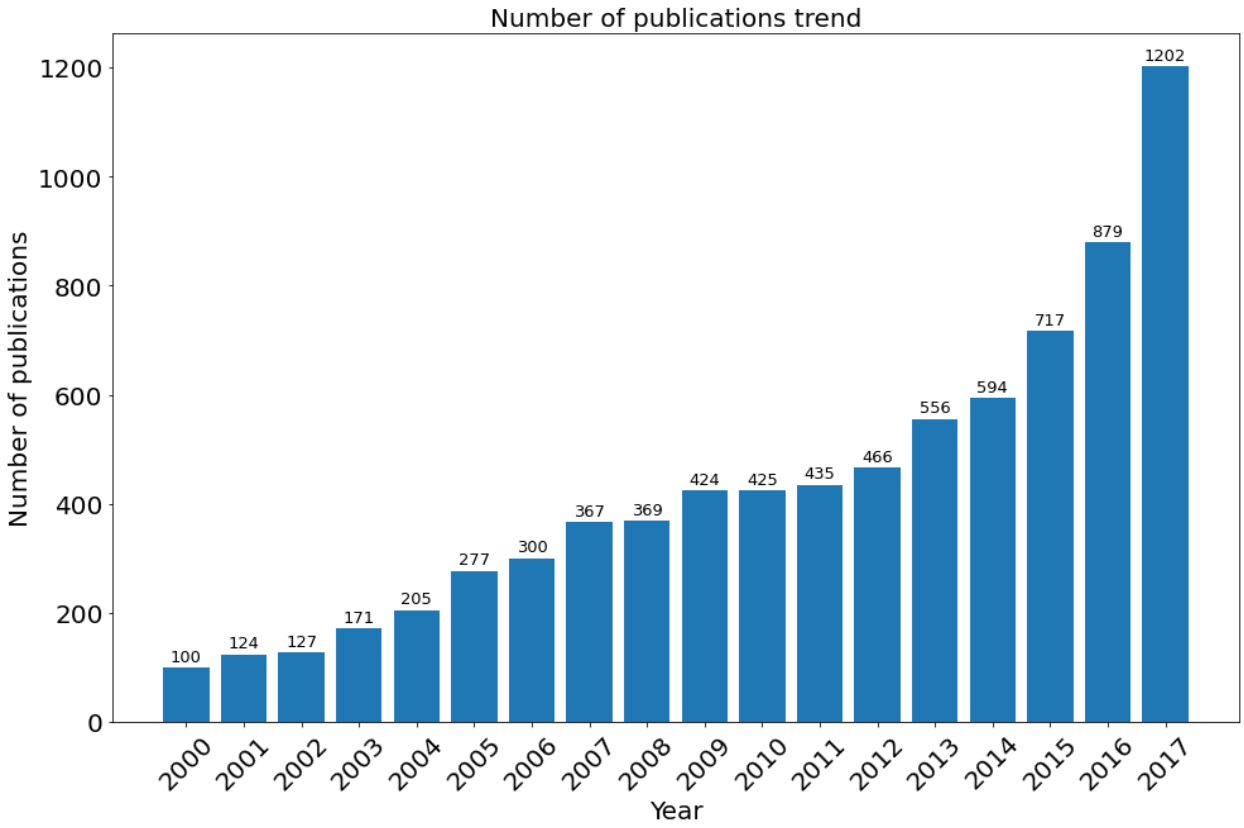


Figure 6. Scopus dataset, AI in cancer publications trend.

These publications include 46,644 authors involved in 123,054 collaborations, i.e., co-authorship relationships. Figure 7 shows the number of authors trend. Similar to the number of publications, the number of authors has been growing in recent years. In 2017, the number of authors had ~50% year over year (YoY) growth.

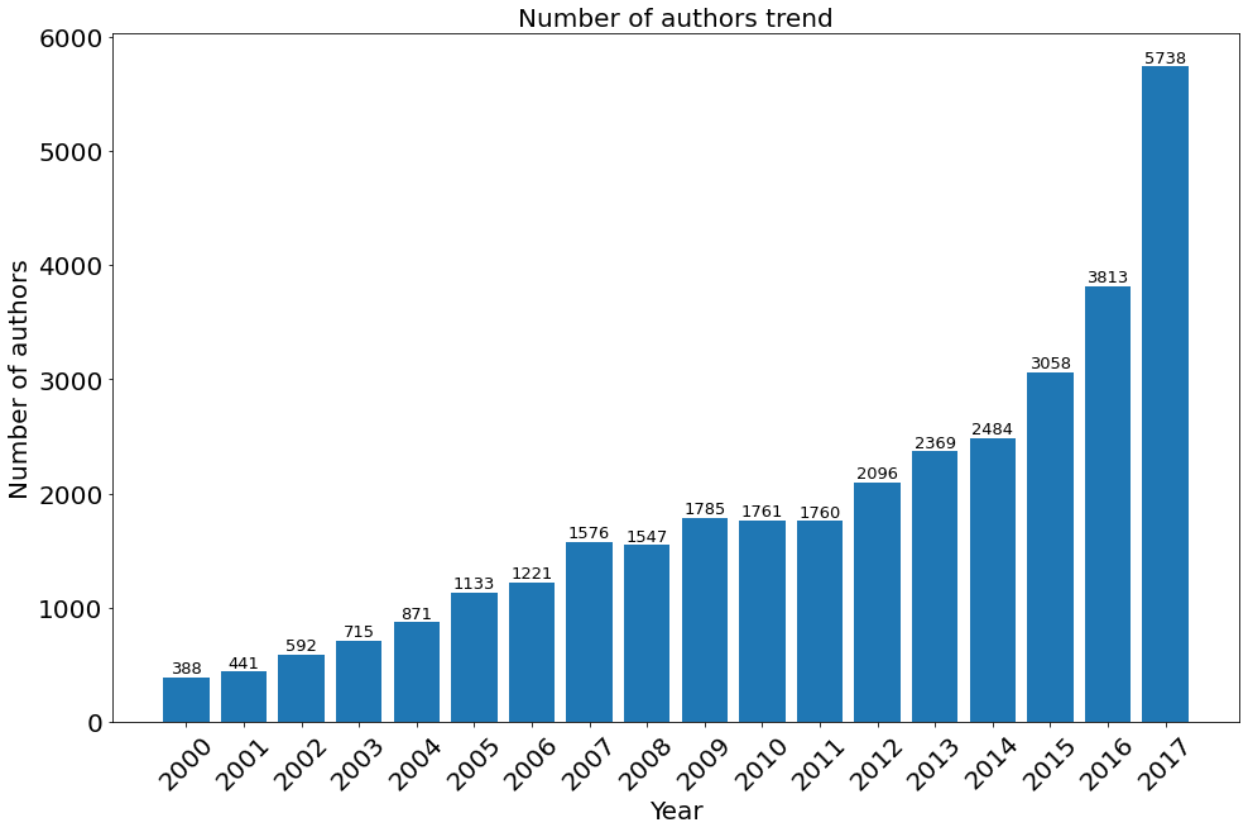


Figure 7. Scopus dataset, number of authors trend.

Figure 8 also shows the number of co-authorship links over time. The analysis result shows that the number of collaborations trend can be grouped into four periods. From 2000-2002 the number of collaborations increased and experienced a spike in 2002 with YoY growth of ~190%. Later on, in 2003, the number of collaborations dropped by ~44%. From 2003 to 2007, there was a growing trend with an increase in YoY collaborations by ~58% in 2007. In 2008, the number of collaborations experienced a drop of 21%, but later on, we saw a growing trend until 2010. In 2010, the number of collaborations increased by ~89%. However, in 2011, the number of collaborations dropped by ~47%. The reason for the spike in 2010 is a nationwide paper who had 102 number of collaborators including but not limited to medical doctors and statisticians (Buri et

al. 2010). Since 2011, the number of collaborations has been growing, with ~123% YoY growth in 2017.

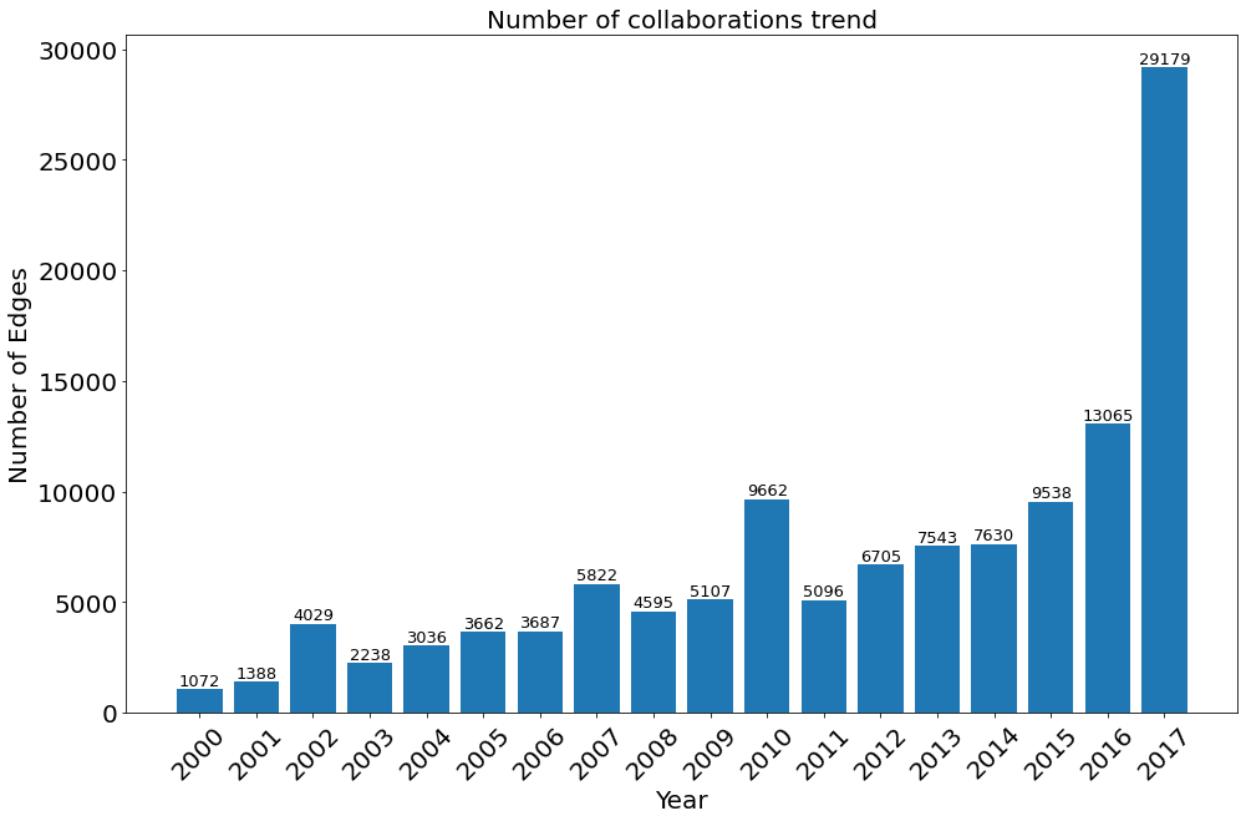


Figure 8. Scopus *dataset, n*umber of collaborations trend.

5.2 Methodology

In this section, we discuss the methodologies used to satisfy the second research objective of this thesis. As shown in Figure 9, after collecting publications data from Scopus and combining their titles and abstracts as explained in chapter 4, we followed three consecutive steps to explore different co-authorship patterns in AI in cancer research: 1) we carefully preprocessed publications

data to prepare it for analysis, 2) we performed ML model development to predict the co-authorship links of different co-authorship patterns, and 3) we included an interpretability module to the analytics pipeline to identify the most predictive features. These features proved to be effective and reasonable indicators for different co-authorship pattern and thus we consider them as driving factors for co-authorship.

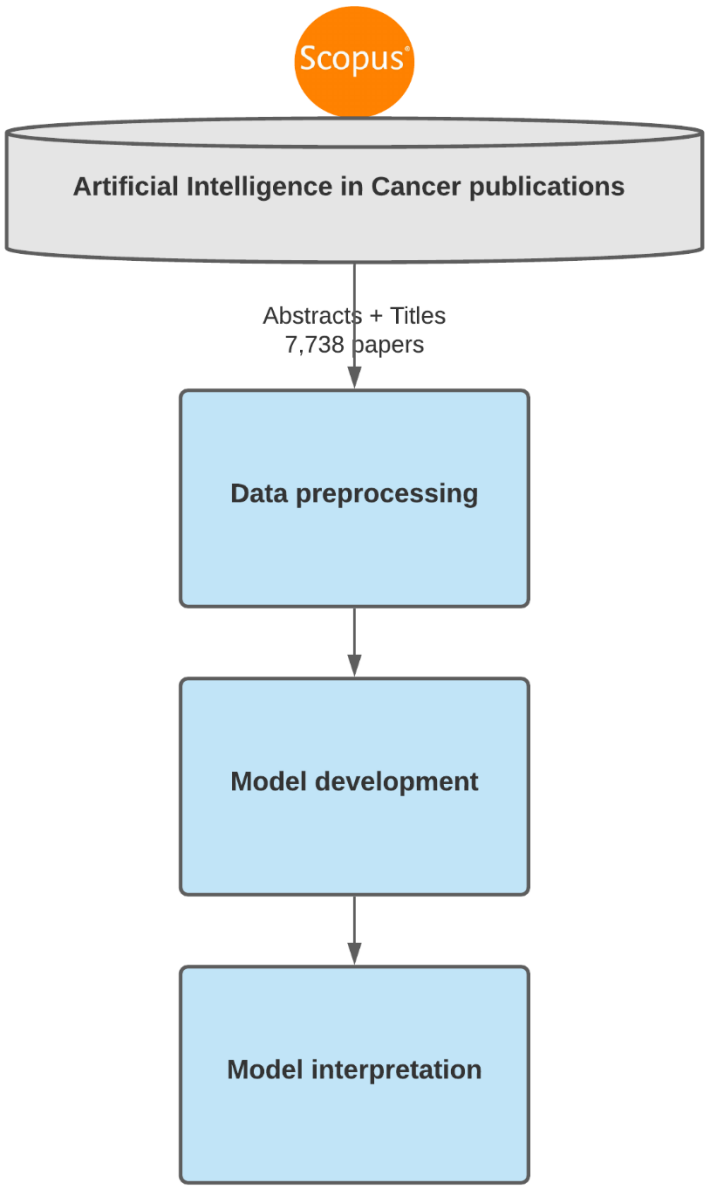


Figure 9. Methodology steps for identification of co-authorship driving factors

5.2.1 Data preprocessing

In this step, we preprocessed data to prepare it for the model development and model interpretation steps. The data preprocessing step consists of:

1) Textual data preprocessing
2) Data preparation
3) Feature extraction

5.2.1.1 Textual data preprocessing

We already discussed the textual data processing in section 4.1. We converted the textual data to lower case, we removed special characters (e.g., #, $, %), stop words (e.g., a, a, the), tags (e.g., , <body>,), and punctuation marks. We then tokenized the textual data and applied lemmatization. Lemmatization is the process of finding the normalized form of words. For instance, lemmatization reduces the words including builds, building and built to the normalized form of build. Lemmatization is used in text processing to improve the performance of topic identification (Skorkovská 2012). The output of the textual data preprocessing step was a document-term frequency matrix that was used to extract the "author's discipline" which we will explain in the feature extraction step in section 5.2.1.3.

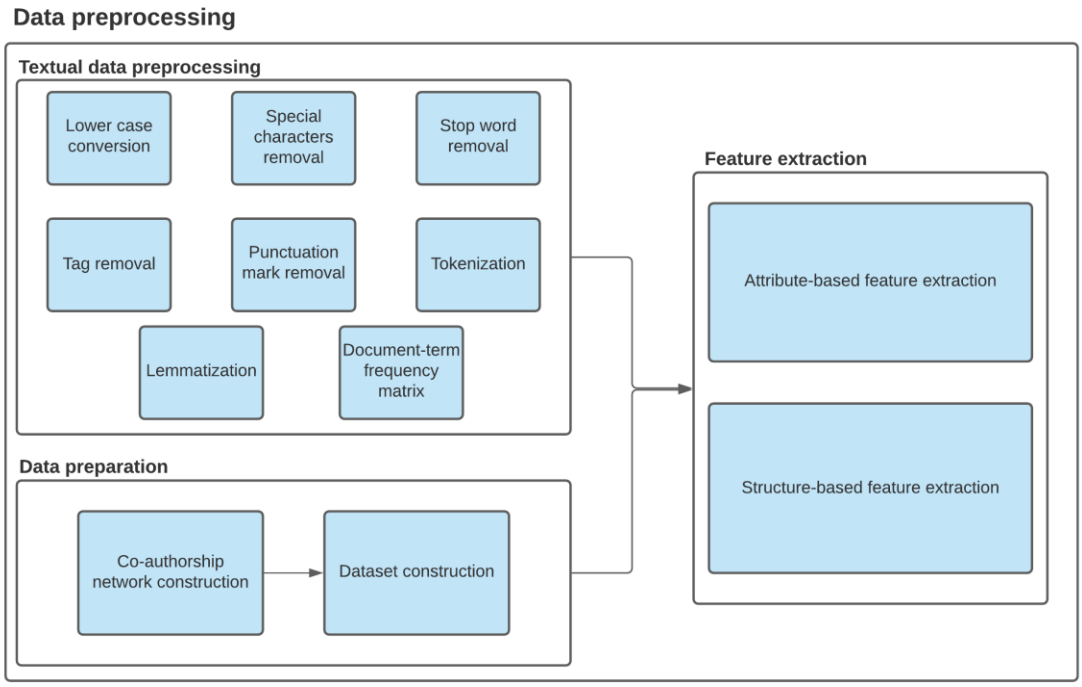


Figure 10. Data preprocessing steps

5.2.1.2 Data preparation

For the data preparation step, we constructed 12 overlapping co-authorship networks. As shown in Table 2, for each co-authorship network, the input window period length is four years and the length for the prediction window is three years, which are the same intervals as used in similar research, for example by Pavlov and Ichise (2007) or by Chuan et al. (2018). In other words, to predict the presence or absence of a co-authorship link for an author pair in future, we used their previous four years of co-authorship network to extract different features and then used the following three years to set a corresponding label by checking whether or not there was a new/persistent/discontinued co-authorship link between two given authors.

The objective of the data preparation step was to construct different datasets for co-authorship networks to be served as inputs for ML models. We formulated the problem as a supervised learning problem that is each dataset has a feature vector and a corresponding label. The feature vector consists of attribute-based and structure-based metrics and the label shows the presence of a new/persistent/discontinued co-authorship link according to different co-authorship patterns. We observed that some authors were only active during the input time window or only during the prediction time window. Therefore, to have a labelled dataset consisting of feature vectors and labels, we investigated the co-authorship link of authors who were active in both input and prediction time windows, as proposed by (Pavlov and Ichise 2007). Table 3 shows the number of authors (nodes), collaborations (edges), the density of co-authorship networks and the number of active authors for all 12 co-authorship networks.

For the dataset construction step, we constructed three groups of datasets for the three co-authorship pattens. Each group has twelve datasets corresponding to the twelve co-authorship networks shown in Table 2.

Table 2. Co-authorship network construction

Co-authorship network number	Co-authorship input window period	Co-authorship prediction window period
1	2000-2003	2004-2006
2	2001-2004	2005-2007
3	2002-2005	2006-2008
4	2003-2006	2007-2009
5	2004-2007	2008-2010
6	2005-2008	2009-2011
7	2006-2009	2010-2012
8	2007-2010	2011-2013
9	2008-2011	2012-2014
10	2009-2012	2013-2015
11	2010-2013	2014-2016
12	2011-2014	2015-2017

Table 3. Statistics of co-authorship networks

Co-authorship network number	Number of authors	Number of collaborations	Network density	Number of active authors
1	1803	7549	0.0046	319
2	2237	9409	0.0038	433
3	2840	11253	0.0028	539
4	3368	10586	0.0019	609
5	4055	13599	0.0017	684
6	4641	15185	0.0014	739
7	5208	16784	0.0012	819
8	5667	22632	0.0014	872
9	5851	22071	0.0013	875
10	6383	24239	0.0012	968
11	6926	26773	0.0011	1001
12	7612	24707	0.0009	1235

For the new co-authorship pattern, label 1 and label 0 correspond to the presence and the absence of a new co-authorship link in future, respectively. One of the common challenges of link prediction in co-authorship networks is the presence of imbalanced datasets for classification. The imbalanced dataset for a binary classification[6] task is a dataset, in which the number of observations in one label is significantly lower than the other label. For instance, in the new co-authorship pattern, the number of author pairs who did not collaborate in the prediction windows was significantly higher than the pairs of authors who collaborated. To reduce the imbalanced ratio of the datasets, we excluded co-authorship links among pairs of authors who had no path between them in their corresponding co-authorship networks. Moreover, we used the random under-sampling method to downsample pairs of authors from the majority class. The random under-sampling method was selected after experimenting with different techniques for dealing with imbalanced classification problems such as synthetic Minority Oversampling Techniques (SMOTE). Table 4 shows the statistics of 12 datasets, including the number of samples, the number of label 1 and label 0 samples and the imbalance ratio, which shows the proportion of samples of the number of majority class (label 0) to the number of minority class (label 1) for new co-authorship.

[6] Binary classification is the task of classifying the observations of a dataset into two different groups.

Table 4. New co-authorship datasets' statistics

Co-authorship network number	Number of samples	Number of label 1	Number of label 0	Imbalance ratio
1	681	23	658	0.03
2	794	153	641	0.2
3	390	84	306	0.3
4	661	92	569	0.2
5	2619	97	2522	0.04
6	2902	126	2776	0.05
7	2048	120	1928	0.06
8	1966	101	1865	0.05
9	1782	122	1660	0.07
10	1278	106	1172	0.09
11	1625	92	1533	0.06
12	2625	88	2537	0.03

Similar to the new co-authorship pattern, for persistent and discontinued co-authorship patterns we built 12 datasets. For the persistent co-authorship pattern, label 1 represents the continuation of collaboration between a pair of authors, and label 0 corresponds to the discontinuation of co-authorship. In contrast, for the discontinued co-authorship pattern, label 1 represents the discontinuity of collaboration, and label 0 shows the continuation of collaboration. The discontinued co-authorship pattern is the opposite case of the persistent co-authorship pattern. Table 5 shows the statistics of the datasets for the persistent co-authorship pattern. Since the only difference between the second and third co-authorship patterns is how we defined class 1 and class 0, we only provided the statistics of the dataset for the second co-authorship pattern which is persistent co-authorship pattern.

Table 5. Persistent co-authorship datasets' statistics

Co-authorship network number	Number of samples	Number of label 1	Number of label 0	Imbalance ratio
1	446	378	68	0.2
2	807	712	95	0.1
3	998	860	138	0.2
4	910	736	174	0.2
5	1105	909	196	0.2
6	1226	1007	219	0.2
7	1337	1079	258	0.2
8	1231	988	243	0.2
9	1203	929	274	0.3
10	1439	1179	260	0.2
11	1580	1286	294	0.2
12	1818	1500	318	0.2

5.2.1.3 Feature extraction

For the feature extraction step, we extracted and calculated two groups of features as follows:

- Attribute-based features
- Structure-based features

Tables 6 shows the attribute-based and structure-based features used in this study.

Table 6. Features for co-authorship link prediction

Attribute-based features	**Structure-based features**
Discipline similarity score	Common neighbors
Seniority similarity score	Jaccard coefficient
Seniority level of authors	Adamic-Adar index
Productivity similarity score	Preferential attachment
Productivity level of authors	Shortest path
Collaboration diversity similarity score	Weighted common neighbors
Collaboration diversity level of authors	Weighted Jaccard coefficient

Number of common source titles	Weighted Adamic-Adar index
	Weighted preferential attachment
	Collaboration weight

Attribute-based features are features at the individual level of authors such as: 1) author's discipline, 2) author's seniority level, 3) author's productivity level 4), author's collaboration diversity level, and 5) author's source titles[7].

First, we extracted the above features from individual authors and then we calculated the following features for author pairs:

- Discipline similarity score
- Seniority similarity score
- Seniority level of authors
- Productivity similarity score
- Productivity level of authors
- Collaboration diversity similarity score
- Collaboration diversity level of authors
- Number of common source titles

[7] Source titles of authors represent the names of the conferences and journals in which authors published their papers.

Author's discipline

Researchers analyzed textual data of publications to understand the discipline of authors for co-authorship link prediction. For instance, Chuan et al. (2018) proposed a hybrid content similarity metric that used LDA as a topic modeling approach to identify researchers' disciplines by extracting the research topics from their publications. This motivated us to leverage publication's textual data to extract the discipline of authors.

As shown in Figure 11, after extracting the publications from Scopus, we preprocessed data. We applied the same preprocessing steps for textual data, including lowercase conversion, special characters removal, stop word removal, tokenization, tag removal, lemmatization, and punctuation mark removal, and then we built a document-term frequency matrix. This matrix was used as an input for the LDA topic modeling algorithm.

We performed hyperparameter tuning on the LDA model. The LDA algorithm required the number of topics to be defined in advance. We found the best number of topics to be 6 based on the quality of the topics assessed quantitatively with *CV* coherence score (Röder et al. 2015), and qualitatively by verifying with three domain experts. The output of the LDA topic modeling process was a document-topic probability dataset for the publications. In the next step, we calculated the discipline of authors by taking the average of their publications' feature vectors, representing their concentration/contribution on/to the extracted topics.

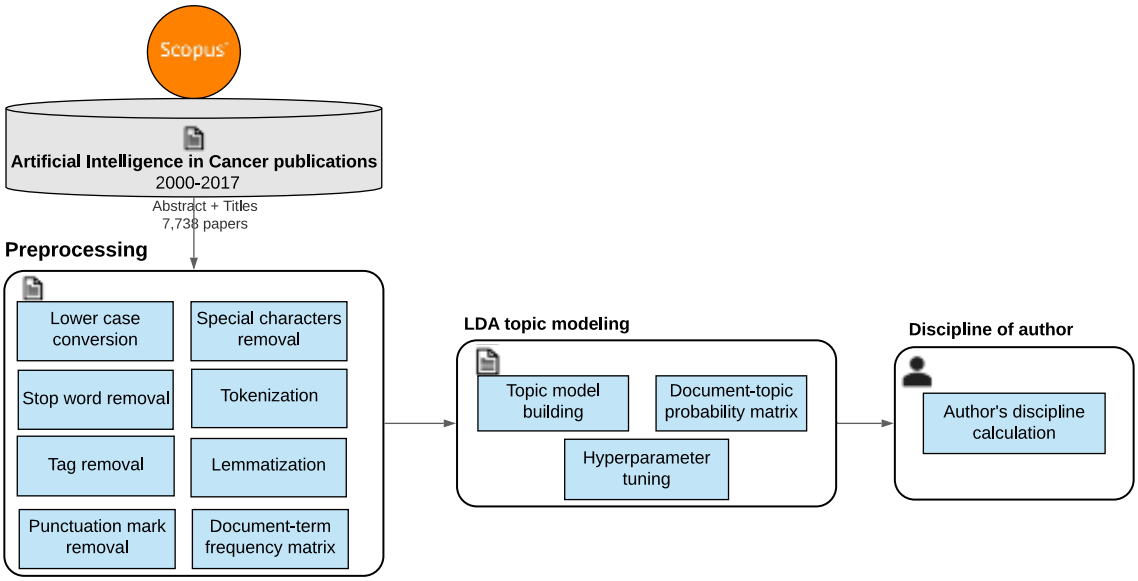


Figure 11. Conceptual flow of the author's discipline calculation

After extracting the discipline of individual authors and having a topic feature vector for each author as a result, we calculated the discipline similarity score for each author pair as shown in Equation (1). We used the discipline similarity score to understand how similar a pair of author's research interests are. The assumption behind this similarity score is that authors are more likely to collaborate with authors doing research within similar disciplines. For instance, two authors are more likely to collaborate if they are working in the similar fields such as medical image analytics for different cancer types. We used cosine similarity to calculate the discipline similarity score for an author pair. The cosine similarity score between two feature vectors X and Y (in this case, the topic feature vector for an author) is calculated based on the cosine angle between the vectors as follows:

$$\cos\theta = \frac{\vec{X}.\vec{Y}}{||\vec{X}||\,||\vec{Y}||}. \quad (1)$$

Discipline similarity score values range from 0 to 1, with higher values representing the higher similarity between the discipline of a pair of authors.

Author's seniority level

We calculated career age of authors as a proxy for their seniority level. To do so, we calculated the period between an author's first and last publications. After calculating the individual author's career age, we calculated two metrics for a pair of authors including: 1) seniority similarity score, and 2) seniority level of authors. Seniority similarity score represents how similar a pair of authors are in terms of their career age. It is calculated as the absolute value of a pair of authors' career age difference. For instance, for a pair of authors, who have the career age of 3 years, the seniority similarity score is zero meaning that they are similar in terms of seniority level whereas for two authors with the career age values of 10 and 2, the seniority similarity score is 8 showing that they are different in terms of seniority. The lower value of this metric shows higher seniority similarity score and vice versa. Seniority level of authors metric was used to show the seniority level of a pair of authors combined. For instance, for two authors with career age values of 5 and 3. The seniority level of authors is calculated to be 8. Unlike seniority similarity score metric, higher values of this metric show that the seniority level of a pair of authors is high and vice versa. It is calculated as the sum of the career age of a pair of authors.

Author's productivity level

We used an author’s number of publications as a proxy for his/her productivity level. As discussed in the literature review, the productivity level of authors is an essential metric for collaboration in

co-authorship networks (Glänzel and Schubert 2006). After calculating the productivity level of individual authors, we calculated two metrics for a pair of authors: 1) productivity similarity score, 2) and CO. Productivity similarity score metric calculates how similar a pair of authors are in terms of their productivity levels. The productivity similarity score is measured by calculating the absolute value of a pair of authors' numbers of publications (productivity levels) difference. The lower value of this feature shows higher productivity similarity score and vice versa. For instance, for a pair of authors, who have the productivity levels of 5 years, the productivity similarity score is zero meaning that they are similar in terms of productivity level whereas for two authors with the productivity levels of 4 and 9, the productivity similarity score is 5 showing that they are different in terms of productivity levels. Productivity level of authors metric was used to show the productivity level of a pair of authors combined, calculated as the sum of a pair of authors' number of publications. For instance, for two authors with productivity levels of 4 and 3, the productivity level of authors is calculated to be 7. Unlike productivity similarity score, higher values of this metric show that the productivity level of a pair of authors is high and vice versa.

Author's collaboration diversity level

We used authors' degree centrality[8] in a co-authorship network as a proxy for their collaboration diversity level. The collaboration diversity level shows how many collaborations with different persons an author has previously had. After calculating the collaboration diversity level of individual authors, we calculated two different metrics: 1) collaboration diversity similarity score, and 2) collaboration diversity level of authors. The collaboration diversity similarity score calculates how similar a pair of authors are based on their collaboration diversity levels. The collaboration diversity similarity score is measured by calculating the absolute value of a pair of authors' degree centrality values (collaboration diversity levels) difference. The lower value of this metric shows higher collaboration diversity similarity score and vice versa. Diversity similarity level of authors feature was used to show the diversity level of a pair of authors combined, i.e., calculated as the sum of a pair of authors' degree centrality values. Unlike collaboration diversity similarity score, higher values of this feature show that the collaboration diversity level of a pair of authors is high and vice versa.

Author's source titles

Source titles of authors represent the names of the conferences and journals in which authors published their papers. Even though we extracted the authors' disciplines, we also considered this feature as a general representation of an author's discipline. This feature calculates the number of common journals and conferences that a pair of authors published their papers.

[8] The degree centrality of a node in a network/graph is the fraction of nodes it is connected to. It is normalized by dividing the number of links that a node has divided by the number of possible links.

Structure-based feature extraction

This section involves the last step of data preprocessing which is extracting the structure-based features. They are called structure-based features since they are based on the structure of networks and extracted from co-authorship networks. We will introduce these features in detail, but first, we present the terminology and notation used throughout this section.

Terminology and notation

We followed the same terminology and notations as Martínez, Berzal, and Cubero (2016). We define network G as a tuple $G = (V, E)$ where V is a set of nodes/vertices and E is a set of edges/links between pairs of nodes in set V. $e_{x,y}$ represents an edge between the pair of node x and node y. $|V|$ shows the size of the network, i.e., the number of nodes in the network. $|E|$ represents the number of links in the network. Γ_x shows the set of nodes connected to node $x \in V$, also known as the neighborhood of node x. The degree of a node denoted as $|\Gamma_x|$ is the number of edges connected to a node. For showing the algorithmic complexity, we denote the total number of nodes and total number of edges by v and e respectively. We built undirected networks where all the edges are bidirectional.

In general, the structure-based metrics calculate a score for each pair of authors according to a function. These scores represent the similarity level between two authors in the network. Researchers categorize similarity-based metrics into three groups based on the amount of information they use for link prediction (Martínez et al 2016). These three subcategories are as follows:

1) Local similarity-based metrics
2) Global similarity-based metrics
3) Quasi-local similarity-based metrics

Local similarity-based link prediction features

Local similarity-based metrics consider only nodes with distance two to calculate the similarity scores. Therefore, the amount of information they use is less than global and quasi-local approaches, making them suitable for parallelized calculation (Martínez et al. 2016). This type of behavior is not always beneficial, especially in non-small-world networks where links form between nodes with a distance of more than two (Liben-Nowell and Kleinberg 2007; Martínez et al. 2016).

Global similarity-based link prediction approaches

Unlike local similarity-based approaches, global approaches consider the whole network structure to calculate the similarity score. Therefore, the parallelization of computation for these algorithms is considered challenging for large networks and distributed environments where each server (computation node) may not cover the network's whole structure (Martínez et al. 2016).

Quasi-local similarity-based link prediction approaches

Quasi-local approaches provide a balance between local and global similarity-based approaches by considering additional structural information of a network (Martínez et al. 2016). Therefore, their time complexity is lower than the time complexity of global similarity-based approaches.

In this thesis, we used local similarity-based link prediction metrics since they are proven to provide good performance for co-authorship link prediction (Chuan et al. 2018; Martínez et al. 2016; Pavlov and Ichise 2007). Moreover, the time complexity of these metrics is often the lowest which makes them suitable for distributed computation and makes these algorithms highly scalable in production (Martínez et al. 2016). Therefore, the local similarity-based link prediction metrics require less computation than global and quasi-local metrics. The time and spatial complexity of these metrics are $O\left(vk^2 f(k)\right)$ and $O(vk^2)$ respectively. $f(k)$ is calculated as the complexity of computing a particular similarity score for a pair of nodes and k is the maximum degree of a node in the network. The local similarity-based link prediction metrics are divided into two main groups such as:

1) Unweighted similarity-based local metrics
2) Weighted similarity-based local metrics

We used both "unweighted local similarity-based" and "weighted local similarity-based" link prediction metrics to have a better grasp of the driving factors for different co-authorship patterns and understand if adding weighted similarity-based link prediction metrics improves the performance of the classifiers for co-authorship link prediction. To avoid long names, we call these metrics unweighted and weighted similarity-based metrics.

Unweighted similarity-based metrics calculate a similarity score for two given nodes in the network, but they do not incorporate the weight of the edges into their metrics. For example, in a co-authorship network, for two authors who might have collaborated several times on a joint paper

in the past, these co-authorship links are only considered as one collaboration by unweighted similarity-based metrics. In contrast, weighted similarity-based metrics include weight of co-authorship links in their calculations and two authors with higher number of joint collaborations are considered to have stronger ties than author pairs with fewer number of previous collaborations. In this thesis, we used the following unweighted similarity-based metrics. These metrics were successfully adopted by researchers to predict new co-authorship links (Chuan et al. 2018; Hasan et al. 2006; Pavlov and Ichise 2007; Yu et al. 2014).

1) Common neighbors (CN)
2) Jaccard coefficient (JC)
3) Adamic-Adar index (AA)
4) Preferential attachment index (PA)
5) Shortest path

Common neighbors (CN)

Common neighbors metric assumes that two nodes with a higher number of common neighbors are more likely to form a link in the network. This assumption stems from observing the correlation between the number of common neighbors and the probability of link formation (Newman 2001). Common neighbors metric (Liben-Nowell and Kleinberg 2007) defines the similarity score as the number of shared neighbors as in Equation (2):

$$CN = | \Gamma_x \cap \Gamma_y |. \quad (2)$$

In co-authorship networks in which nodes represent authors and edges represent co-authorship links, common neighbors metric calculates the number of common co-authors for each pair of authors in the co-authorship network.

Jaccard coefficient (JC)

Jaccard coefficient metric has roots in information retrieval systems and was first introduced by Paul Jaccard (Jaccard 1901). As shown in Equation (3), Jaccard coefficient metric defines the similarity score as the ratio of the number of common neighbors to the set of all neighbors of the two nodes.

$$JC = \frac{|\Gamma_x \cap \Gamma_y|}{|\Gamma_x \cup \Gamma_y|}. \quad (3)$$

In co-authorship networks, the Jaccard coefficient index calculates the ratio of common co-authors to the total number of possible co-authors in the co-authorship network for a pair of authors.

Adamic-Adar index (AA)

Adamic and Adar (2003) introduced this approach to assign a similarity score to a pair of nodes that considers the features of their common neighbors. Adamic-Adar index is a derived version of common neighbors where each common neighbor is penalized by its degree as shown in Equation (4). This characteristic was expected particularly for social networks in which the amount of resources that a node can allocate between its resources decreases as the number of neighbors of that node (node degree) increases (Martínez et al. 2016).

$$AA = \sum_{z \in \Gamma_x \cap \Gamma_y} \frac{1}{\log |\Gamma_z|}. \quad (4)$$

In the co-authorship network, Adamic-Adar Index metric assigns greater weights to common co-authors that have fewer co-authors (Yu et al. 2014).

Preferential attachment (PA)

The preferential attachment metric is based on the power-law distribution and scale-free network, which was introduced by the Barabasi-Albert network model (Barabási and Albert 1999). They proposed that as the degree of two unconnected nodes increases the probability of forming a link between them increases (Martínez et al. 2016). Based on this model, the similarity score is defined in Equation (5).

$$PA = |\Gamma_x| \left|\Gamma_y\right|. \quad (5)$$

In a co-authorship network, preferential attachment metric is calculated as the product of an author pairs' degrees in the network.

Shortest path (SP)

Shortest path is a network measure that calculates the length of the shortest path between two nodes in a network. Pair of nodes with shortest path distance equal to 1 are linked to each other by an edge.

$$SP = \left|shortest\ path_{x,y}\right|. \quad (6)$$

In co-authorships networks, this metric calculates the distance between two authors in a network. This metric has been used as a metric for link prediction (Liben-Nowell and Kleinberg 2003).

Weighted similarity-based metrics

Unlike unweighted similarity-based features, weighted similarity-based features, as shown in Equation (7), (8), (9) and (10), consider extra level of information to calculate the similarity score by incorporating different types of collaboration weights into their formulas. The collaboration weight denoted by $\omega(a, b)$ calculates the weight of the link between nodes a and b. This study defines the collaboration weight for weighted similarity-based metrics as the number of joint papers a pair of authors have previously collaborated on (Chuan et al. 2018). For instance, weighted common neighbors, as shown in Equation (7), assumes that author pairs with higher number of common co-authors with whom these author pairs published higher number of joint publications are more likely to collaborate in future (Murata and Moriyasu 2007). Other weighted similarity-based metrics such as weighted Jaccard coefficient, weighted Adamic-Adar index and weighted preferential attachment hold the similar assumptions as the weighted common neighbors.

We used both weighted and unweighted similarity-based metrics to have a comprehensive view of the driving factors for different co-authorship patterns since weighted similarity-based features were successfully used by researchers for co-authorship link prediction (Chuan et al. 2018; Murata and Moriyasu 2007). As discussed in last paragraph, weighted similarity-based metrics use weights or the "strength" of relationships between authors in a network, this has been shown to improve

the performance of the models to predict future co-authorship links in a network in some co-authorship networks (De Sá and Prudêncio 2011).

In these formulas, x and y are node pairs of interest. the weighted similarity-based metrics calculate different weighted similarity-based metrics for these nodes. z denotes the nodes that are elements of common neighbors set of nodes x and y. $\omega(x,z)$ calculates the weight of the link between nodes x and its common neighbor z. Similarly, $\omega(y,z)$ shows the weight of the link between nodes y and its common neighbor z.

Weighted common neighbors (WCN)

$$WCN = \sum_{z \in \Gamma_x \cap \Gamma_y} \frac{\omega(x,z) + \omega(y,\ z)}{2}. \quad (7)$$

As the name suggests, weighted common neighbors is a weighted version of common neighbors which incorporates the weight of links between nodes x and y and their common neighbors (Chuan et al. 2018)

Weighted Jaccard coefficient (WJC)

$$WJC = \frac{\sum_{z \in \Gamma_x \cap \Gamma_y} \frac{\omega(x,z) + \omega(y,\ z)}{2}}{\sum_{u \in \Gamma_x} \omega(x\ ,u) + \sum_{v \in \Gamma_y} \omega(y,v)}. \quad (8)$$

Similarly, Weighted Jaccard coefficient is a weighted version of Jaccard coefficient (Chuan et al. 2018). u and v denote the nodes that are neighbors of nodes x and y respectively. (x, u) calculates the weight of the link between nodes x and its neighbor u. Similarly, $\omega(y, v)$ measures the weight of the link between nodes y and its neighbor v.

Weighted Adamic-Adar index (WAA)

$$WAA = \sum_{z \in \Gamma_x \cap \Gamma_y} \frac{\omega(x,z) + \omega(y,\ z)}{2} \times \frac{1}{Log(\sum_{t \in \Gamma_z} \omega(z,\ t))}. \quad (9)$$

This metric is a weighted version of the Adamic-Adar metric which considers the weight of links between nodes and their common neighbors as well as the links between common neighbors and their neighbors (Chuan et al. 2018). t denotes the nodes that are elements of neighbors set of nodes z which themselves are elements of common neighbors set of nodes x and y.

Weighted preferential attachment (WPA)

$$WPA = \sum_{u \in \Gamma_x} \omega(x,u) \times \sum_{v \in \Gamma_y} \omega(y,v). \quad (10)$$

This metric is a weighted version of preferential attachment. (x, u) shows the weight of the link between nodes x and its neighbor u (Chuan et al. 2018). Similarly, $\omega(y, v)$ shows the weight of the link between nodes y and its neighbor v.

Collaboration weight (CW)

We also considered collaboration weight as a separate metric that shows the strength of collaboration between two authors in a network (Parimi and Caragea 2011). Not only does this metric consider the number of collaborations for a pair of authors but also it considers the number of authors in their joint publications. As shown in Equation (11), the collaboration weight for a pair of authors is penalized by the number of authors in the joint publications.

$$CW = \sum_{i=1}^{p} \frac{\delta_x^i * \delta_y^i}{n_i - 1} . \quad (11)$$

In this formula, δ_x^i = 1 when x is an author of paper i and δ_x^i = 0 otherwise. n_i is the number of authors in paper. p is the total number of papers.

We used the NetworkX Python package to construct the co-authorship networks and extract the structure-based metrics (Hagberg et al. 2008). We used Pandas and NumPy packages for working with dataframes and multi-dimensional arrays, respectively. Pandas is a Python library that consists of tools and data structures for manipulating datasets (McKinney 2011). NumPy is a package for creating and working with multi-dimensional arrays and matrices (Harris et al. 2020). We used Concordia University High-Performance Computing (HPC) facility named Speed[9] to extract and compute the structure-based metrics from the co-authorship networks.

[9] https://www.concordia.ca/ginacody/aits/speed.html

5.2.2 Model development

This step deals with developing several ML classifiers on the datasets discussed in section 5.2.1.2 to predict different co-authorship links such as new, persistent, and discontinued links as a function of attribute-based and structure-based metrics.

Model construction

We chose four ML classifiers . The four classifiers are as follows: 1) logistic regression, 2) decision trees, 3) random forests, and 4) extreme gradient boosting (XGBoost). As we will explain in the model evaluation section, we used the average precision metric which is used to calculate the area under the precision-recall curve for evaluating our models. This metric is recommended for binary classification problems especially when dealing with imbalanced classification tasks (Davis and Goadrich 2006; Saito and Rehmsmeier 2015). We also compared the models against each other and two baselines including the logistic regression model and the prevalence metric. The prevalence metric is the ratio of positive samples. In other words, the baseline for a random classification would be equal to the prevalence.

It is crucial to realize that identifying which classifier to use in order to achieve the highest performance depends on the conditions and structures of a dataset. Therefore it is essential to run experiments with different models and choose the best performing one. For instance, we included logistic regression since sometimes due to the various conditions and structures of some datasets, logistic regression might outperform more complicated algorithms such as random forest (Kirasich et al. 2018).

Logistic regression

Logistic regression is a supervised classification ML algorithm. According to Equation (12), logistic regression models the natural log odds of the target variable using a linear combination of features (Cramer 2005). The output of the linear combination of features is passed through a sigmoid function as follows:

$$S(x) = \frac{1}{1 + e^{-x}}. \quad (12)$$

Logistic regression has two groups of parameters including weights and biases that need to be optimized. To optimize these parameters, we need to define a cost function. The cross-entropy function is used to optimize the parameters as follows:

$$Cross\ entropy\ cost\ function = -\frac{1}{n}\sum_{i}^{n} y_i \,.\log \widehat{y_i} + (1 - y_i)\,.\,\log(1 - \widehat{y_i}). \quad (13)$$

Decision tree

Decision tree is a non-parametric supervised ML algorithm used for regression and classification tasks (Apté and Weiss 1997). A decision tree includes decision nodes, branches (decisions) and leaf nodes. Each node presents a feature, and each branch represents a decision, and leaves represent outcomes. Each branch of the decision tree should end with a leaf node. Decision trees are constructed by recursively splitting the training samples for classification or regressions tasks. The recursion finishes when the classes of all data samples are the same or when partitioning the data does not further improve the performance of the prediction tasks. Evaluation metrics, including but not limited to entropy and the Gini index, are used to measure the node impurity for determining the best split. As shown in Equation (14), entropy is the amount of information

required to describe a sample accurately. $p(i)$ is the fraction of examples for class i and n is the number of classes.

$$Entropy = -\sum_{i=1}^{n} -p_i \log_2 p_i. \quad (14)$$

Gini index, as shown in Equation (15), measures how often a randomly selected element from a set would be incorrectly clasified. $p(i)$ is the probability that a sample being classified to class i and n is the number of classes (Farris 2010).

$$Gini\ index = 1 - \sum_{i=1}^{n} (p_i)^2. \quad (15)$$

Random forest

Random forests is an ensemble learning technique used for both regression and classification (Ho 1995). Random forests constructs many classification and regression trees. Each decision tree is trained on a sample of the training data (bootstrapping) and across a randomly selected subset of the input features. For classification, the class with the most votes becomes the model prediction. For the regression task, the final model prediction is the average of individual trees' predictions. Comparing to decision tree, random forests is less prone to overfitting (Hastie et al. 2009). Random forests uses different evaluation metrics including the Gini index to determine the split with the lowest impurity at every node.

Extreme gradient boosting (XGBoost)

XGBoost is used for both regression and classification tasks. It combines the predictions of weak classifiers to achieve a powerful classifier. XGBoost has demonstrated to be an efficient and

reliable algorithm (Chen and Guestrin 2016) in many application such as in CAD systems to diagnose cancer (Liew et al. 2021), recommender systems to suggest products (Shahbazi et al. 2020), cyber security systems to detect cyber-attacks (Chen et al. 2018) and financial services for credit evaluation (Li et al. 2020). The objective of this algorithm is to minimize the loss function. The loss function formula is as follows:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{t-1} + f_t(x_i)\right) + \Omega\left(f_t\right). \quad (16)$$

n is the total number of the based tree models, i is the instance, and t is each iteration step. y_i is the true value and $\hat{y}_i^{t-1}$ is the previously generated tree model at iteration t and $f_t(x_i)$ is the new generated tree model, $l\left(y_i, \hat{y}_i^{t-1} + f_t(x_i)\right)$ is the loss function and $\Omega\left(f_t\right)$ is the term for regularization. The second order optimization is used to optimize the loss function as each step t as follows:

$$\tilde{\mathcal{L}}^{(t)} = \sum_{i=1}^{n} \left[g_i\, f_t\,(x_i) + \frac{1}{2}\, h_i f_t^2(x_i)\right] + \Omega\left(f_t\right). \quad (17)$$

We used Scikit-learn and XGBoost Python packages to build the model. Scikit-learn is a Python package that includes a large number of state-of-the-art ML algorithms for supervised and unsupervised tasks (Pedregosa et al. 2011). XGBoost is an optimized distributed gradient boosting library used for different ML tasks (Chen and Guestrin 2016). We split the dataset into training and test datasets. The training dataset was 80% of the total dataset size, and the testing dataset was 20% of the total dataset size. We tuned the hyperparameters of the classifiers using random search. Repeated stratified 5-fold cross-validation (number of repeats = 3) was used to validate the ML models and find their best hyperparameters.

Model Evaluation

In a binary classification problem, in which the outcomes are labelled either as positive (p) or negative (n), there are four possible outcomes:

- True positive (TP)
- False positive (FP)
- True negative (TN)
- False negative (FN)

If the prediction outcome is p and the true value is also p, then it is counted as TP, whereas, if the true value is n, it is counted as a FP. When both the prediction and true values are n, it is counted as TN. FN is when the true value is p but the prediction is n. Several metrics can be calculated based on TP, FP, TN and FN and are used to evaluate ML classifiers. The evaluation metrics are as follows:

Precision

As shown in Equation (18), precision, also called positive predictive value, is defined as the proportion of true positive predictions out of all positive predictions. It is considered as an appropriate evaluation metric, especially when FP's cost is high (e.g., email spam detection model).

$$Precision = \frac{TP}{TP + FP}. \quad (18)$$

Recall

Recall represents how well a ML model predicted the actual positive class (Powers 2020). As shown in Equation (19), recall is the proportion of TP predictions out of all actual positive examples. It is considered as a proper metric when the cost of FN's is high (e.g., cancer detection model).

$$Recall = \frac{TP}{TP + FN} . \quad (19)$$

F1 score

F1 score, also known as F-score or F-measure, calculates the weighted average of precision and recall. As shown in Equation (20), the F1 score is the harmonic mean of precision and recall. The range of F1 score is [0, 1].

$$F1 = 2.\frac{Precision * Recall}{Precision + Recall} . \quad (20)$$

Area under the curve (AUC)

AUC is the area under the Receiver Operating Characteristic (ROC) curve. ROC curve presents the performance of a binary classifier for various threshold settings. A ROC curve is created by plotting TP rate as shown in Equation (21) vs. FP rate as shown in Equation (22) at various thresholds. AUC ranges between 0 and 1. A model that misidentified all samples has an AUC of 0, and a model whose predictions are 100% correct has an AUC of 1. The AUC of a random predictor is 0.5.

$$TPR = \frac{TP}{TP + FN} . \quad (21)$$

$$FPR = \frac{FP}{FP + TN} . \quad (22)$$

Average precision (AP)

The precision-recall curve is used for evaluating the performance of binary classifications. It plots the precision against recall across various thresholds. A single value named, average precision, is used to calculate the area under the precision-recall curve. As shown in Equation (23), AP calculates the weighted mean of precision achieved at each threshold, with the weight calculated as the increase in recall from the previous threshold. R_n and P_n are the recall and precision at the nth threshold. The baseline metric for the AP is the prevalence metric which is the fraction of positive samples. A good classifier has an AP between the baseline value and the value of 1.

$$AP = \sum_n (R_n - R_{n-1}) P_n . \quad (23)$$

This metric is recommended for binary classification problems especially when dealing with imbalanced classification tasks (Davis and Goadrich 2006; Saito and Rehmsmeier 2015). Since in this thesis, our classifications tasks are binary as well as imbalanced, we used the AP metric as an evaluation metric to compare different models and choose the best performing models.

5.2.3 Model interpretability

We identified the driving factors for different co-authorship patterns using the SHapley Additive exPlanations (SHAP) approach (Lundberg and Lee 2017). SHAP is a unified approach to interpret ML models based on the game theory approach. The objective of the SHAP is to explain a

prediction of any data point as a sum of contributions of each feature. Therefore, it assigns contribution values named SHAP values to different features. SHAP values are calculated according to the following formula:

$$\phi_i = \sum_{S \subseteq M \setminus \{i\}} \frac{|S|!\,(M - |S| - 1)!}{M!} [f_x(S \cup \{i\}) - f_x(S)]\,. \quad (24)$$

Generally, the above formula measures the weighted average of the difference between the prediction of the model with and without the desired feature. The weight considers all possible permutation sets of features excluding the desired feature. $|M|$ is the total number of features. S presents any subset of features that does not have the feature i. $|S|$ is the size of the subset of features. f is the function that predicts the target for the subset of features (Lundberg and Lee 2017). Lundberg and Lee (2017) showed that compared to other interpretability approaches, using the new method of feature importance measure incorporated in SHAP approach not only improves the computational performance but also, they stated it is more aligned with human intuition.

5.3 Results and discussion

This section presents the results and discussion of the classifiers for the three different co-authorship patterns. We built four classifiers for each of the twelve co-authorship networks of different co-authorship patterns (the total of 48 classifiers). We divide this section into three subsections. The first subsection deals with the results and the discussion for the first co-authorship pattern, which is new co-authorship pattern. The second and third subsection deals with the persistent and discontinued co-authorship patterns, respectively.

In each section, we first compare the performance results from the decision tree, random forests, and XGBoost classifiers against each other and our two baselines including logistic regression as explained and the prevalence metric. Later, we discuss the factors that may be good indicators or driving factors for each co-authorship pattern.

5.3.1 New co-authorship

This section investigates the performance results of the classifiers and identifies the driving factors for the new co-authorship pattern. The objective is to predict the new co-authorship links and interpret the driving factors for collaboration between author pairs who have not collaborated on a joint paper before.

5.3.1.1 Machine learning models' performance results

We trained four different classifiers for each of the twelve co-authorship networks for the new co-authorship pattern. Tables 7, 8, 9, and 10 show the performance results for logistic regression, decision tree, random forests, and XGBoost classifiers, respectively.

Table 7. Logistic regression models' performance results for new co-authorship pattern

Co-authorship network number	F1 score	Recall	Precision	AUC	AP
1	0.32	1.00	0.19	0.94	0.44
2	0.68	0.68	0.68	0.91	0.76
3	0.42	0.41	0.44	0.74	0.54

4	0.48	0.63	0.39	0.79	0.58
5	0.32	1.00	0.19	0.96	0.49
6	0.28	0.76	0.17	0.91	0.44
7	0.26	0.71	0.16	0.86	0.29
8	0.22	0.85	0.13	0.80	0.22
9	0.33	0.83	0.21	0.81	0.30
10	0.31	0.95	0.18	0.87	0.35
11	0.25	0.94	0.15	0.91	0.38
12	0.23	0.50	0.15	0.81	0.23

Table 8. Decision tree models' performance results for new co-authorship pattern

Co-authorship network number	F1 score	Recall	Precision	AUC	AP
1	0.37	1.00	0.23	0.94	0.23
2	0.73	0.90	0.61	0.88	0.55
3	0.49	0.76	0.36	0.76	0.38
4	0.56	0.79	0.43	0.81	0.38
5	0.33	0.95	0.20	0.92	0.25
6	0.34	0.80	0.22	0.84	0.24
7	0.34	0.88	0.21	0.85	0.22
8	0.28	0.60	0.18	0.80	0.17
9	0.31	0.88	0.19	0.84	0.21
10	0.37	0.76	0.24	0.82	0.24
11	0.27	0.78	0.16	0.85	0.25
12	0.14	0.67	0.08	0.77	0.09

Table 9. Random forest models' performance results for new co-authorship pattern

Co-authorship network number	F1 score	Recall	Precision	AUC	AP
1	0.31	1.00	0.19	0.96	0.52
2	0.76	0.87	0.68	0.92	0.73

3	0.56	0.94	0.40	0.87	0.61
4	0.57	0.89	0.41	0.89	0.59
5	0.33	0.95	0.20	0.97	0.74
6	0.33	0.84	0.21	0.92	0.49
7	0.38	0.79	0.25	0.89	0.36
8	0.26	0.75	0.15	0.85	0.42
9	0.34	0.88	0.21	0.90	0.53
10	0.38	1.00	0.23	0.91	0.49
11	0.28	0.94	0.16	0.89	0.42
12	0.15	0.67	0.09	0.82	0.33

Table 10. XGBoost models' performance results for new co-authorship pattern

Co-authorship network number	F1 score	Recall	Precision	AUC	AP
1	0.18	1.00	0.10	0.85	0.30
2	0.67	0.81	0.57	0.88	0.69
3	0.59	0.88	0.44	0.83	0.52
4	0.56	0.74	0.45	0.88	0.52
5	0.30	1.00	0.18	0.96	0.45
6	0.33	0.80	0.21	0.92	0.33
7	0.37	0.83	0.24	0.89	0.31
8	0.33	0.90	0.20	0.87	0.40
9	0.36	0.88	0.22	0.85	0.29
10	0.36	0.95	0.22	0.89	0.41
11	0.27	1.00	0.16	0.89	0.34
12	0.15	0.61	0.09	0.79	0.28

Figure 12 compares AP performance of classifiers against the two baselines for the new co-authorship pattern. As seen in this figure, all classifiers performed better than the baseline-prevalence, which shows the high performance of these classifiers to predict new co-authorship links compared to the random predictor. We also compared XGBoost, random forest and decision tree against logistic regression as a baseline. Random forest classifier performed better than the logistic regression baseline in all co-authorship networks except the second co-authorship network in which logistic regression performed slightly better than random forest (~4%).

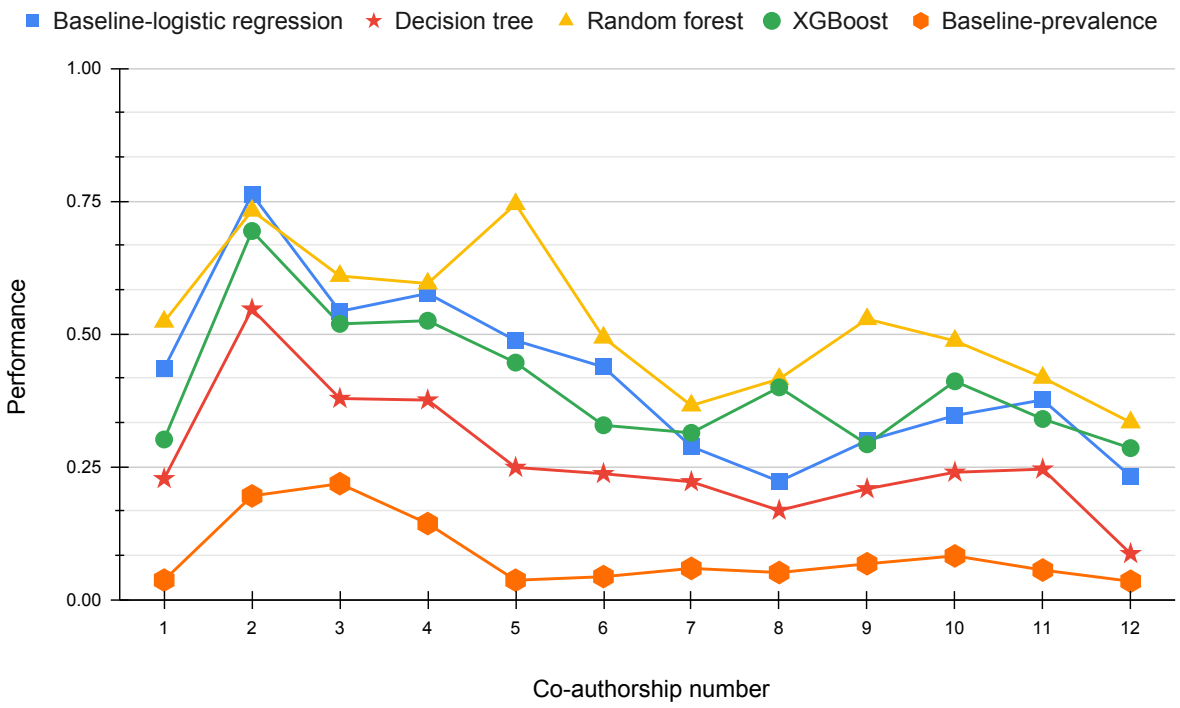


Figure 12. Comparison of models' AP performances against each other and the baselines for the new co-authorship pattern

Random forests model had the highest AP value of 0.74 for the fifth co-authorship network, while the prevalence baseline performance for a random predictor was 0.04 and the logistic regression baseline performance was 0.49. Therefore, the random forests AP performance was approximately 95% higher than the random predictor and also 26% higher than the logistic regression. Logistic regression and XGBoost were subsequently ranked as the second best algorithms for the new co-authorship link prediction. For the co-authorship networks #1, #2, #3, #4, #5, #6, and #11, logistic regression outperformed XGBoost. For all other co-authorship networks, XGBoost provided better performance than logistic regression. The decision tree classifier had the lowest performance for all co-authorship networks. However, it always performed better than the prevalence baseline metric or random predictor.

The average recall (the ability of the model to predict new co-authorship links) for the logistic regression, decision tree, random forests, and XGBoost classifiers was 0.77, 0.81, 0.88 and 0.87, respectively. Therefore, the random forests classifier represented the highest average recall. In other words, on average, random forests predicted 88% of the new co-authorship links. This is encouraging since these models can be used in co-author recommender systems to predict new collaboration links with a good performance.

The average precision for the logistic regression, decision tree, random forests, and XGBoost classifiers was 0.25, 0.26, 0.26 and 0.26, respectively. The lower values of precision was expected in the link prediction problem as mentioned by Chuan et al. (2018). However, when using these classifiers, one might increase either the recall or precision at the expense of the other metric to

have a predictive model that is well tailored to the desired requirements. For instance, in a recommender system for potential co-authors, one might probably require a recall near 1.0 or to find all possible co-authors while accepting low precision if the cost of false positive or recommending wrong co-authors is not significant.

The average AUC for the logistic regression, decision tree, random forests, and XGBoost classifiers was 0.86, 0.84, 0.90, and 0.88, respectively. This shows all classifiers performed well in discriminating between those authors who would have new co-authorship links and those who would not.

5.3.1.2 Driving factors

We used the SHAP approach to find the driving factors or good indicators for the new co-authorship pattern. Figures 13, 14, 15 provide the most predictive factors for the new co-authorship pattern in co-authorship networks #1, #2 and #5. We chose these co-authorship networks since in these networks we saw a high performance with respect to AP and recall in predicting the new co-authorship links. The two colors in these figures represent the positive impact (red) or negative impact (blue) of these factors on the presence of new co-authorship links.

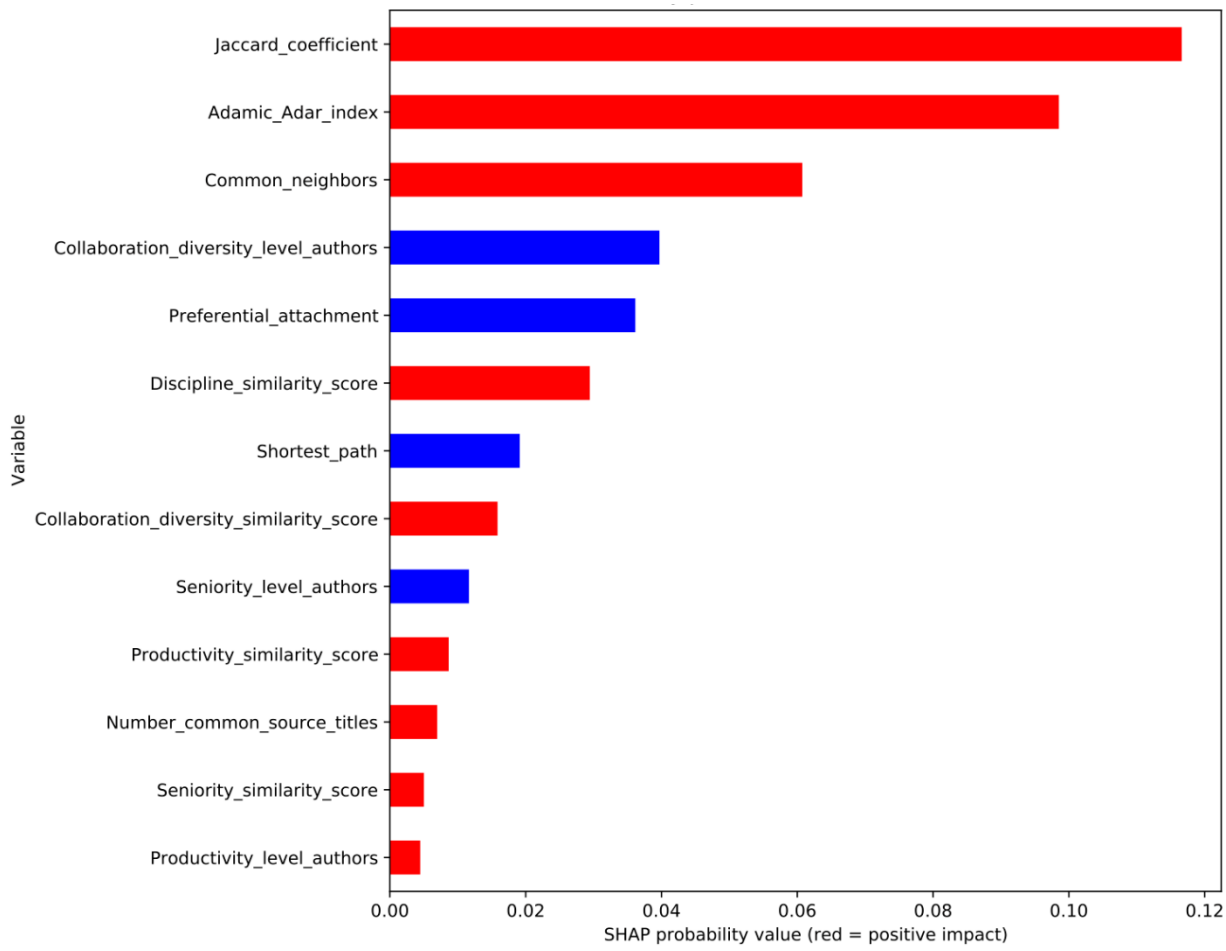


Figure 13. Co-authorship network #1: Driving factors for new co-authorship pattern

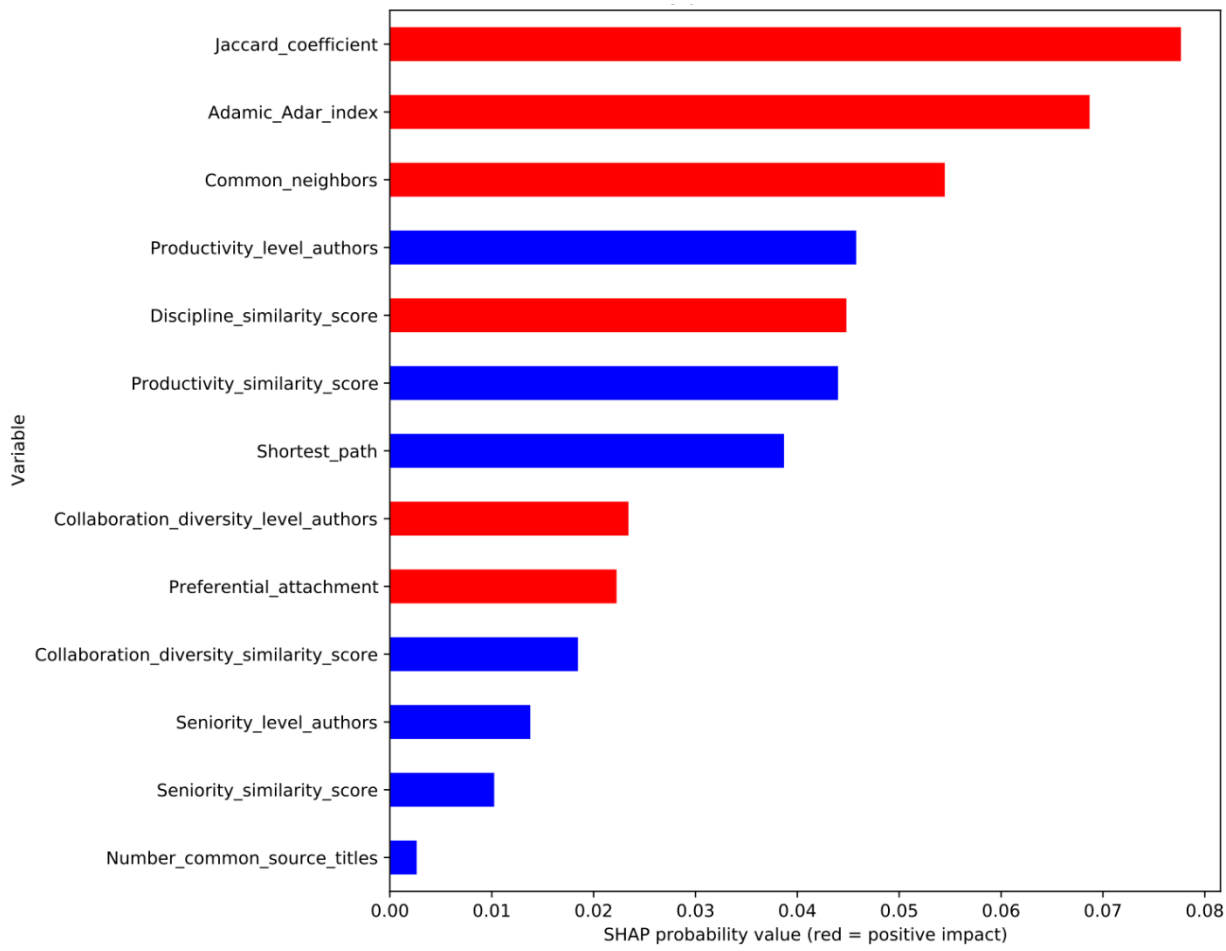


Figure 14. Co-authorship network #2: Driving factors for new co-authorship pattern

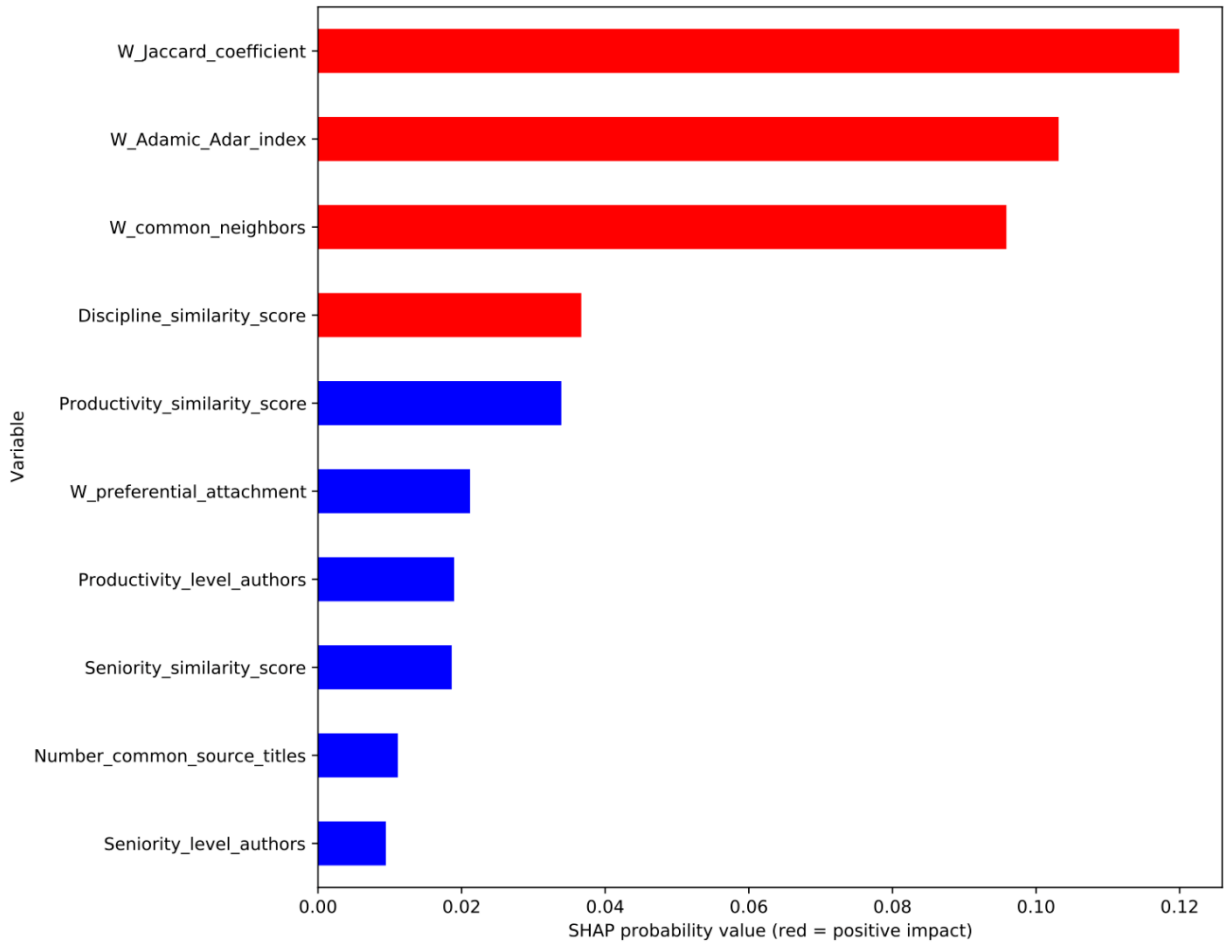


Figure 15. Co-authorship network #5: Driving factors for new co-authorship pattern

As seen in these figures, all types of metrics, including unweighted and weighted similarity-based metrics and attribute-based metrics, appeared among the most predictive factors for the new co-authorship pattern. This shows the importance of considering both structure (unweighted and weighted similarity-based metrics) and attribute-based metrics to have a better performance for the new co-authorship link prediction. In these figures, the horizontal axis shows the contribution value of these metrics in predicting the new co-authorship links for the corresponding co-authorship network. Features are ranked in the descending order based on their contributions, i.e., top-ranked features have the most contribution, whereas the bottom-ranked ones are considered to

have less contribution in the new co-authorship link prediction. We chose the following metrics as the driving factors for the new co-authorship pattern, since these features were among the most predictive features in all 12 co-authorship networks. However, other metrics such as shortest path have shown to be less contributive than the discussed metrics to predict the new co-authorship links. Thus, we did not consider them as driving factors.

- Jaccard coefficient and weighted Jaccard coefficient
- Adamic Adar index and weighted Adamic Adar index
- Common neighbors and weighted common neighbors
- Preferential attachment and weighted preferential attachment
- Discipline similarity score
- Productivity level of authors
- Productivity similarity score

The above metrics can be divided into structure and attribute-based factors. The structural-based factors include Jaccard coefficient, weighted Jaccard coefficient, Adamic-Adar index, weighted Adamic-Adar index, common neighbors, weighted common neighbors, preferential attachment, and weighted preferential attachment metrics. These metrics were extracted from the co-authorship networks and computed based on the structure of the network. It is important to note that, as discussed, the weighted similarity-based metrics also incorporate authors' number of publications. These structure-based factors were found to have the greatest impact on the new co-authorship pattern, which is a result also supported by other researchers (Liben-Nowell and Kleinberg 2007; Pavlov and Ichise 2007; Yu et al. 2014). All these structure-based factors except the weighted and

unweighted preferential attachment metrics are variants of common neighbors metrics. As seen in Figures 13, 14 and 15, these common neighbors-based metrics have a positive impact on the appearance of new co-authorship links. Common neighbors metric calculates the number of common co-authors for a pair of authors. Therefore, in these co-authorship networks, authors with a higher number of co-authors are more likely to collaborate in future than those with fewer co-authors. Jaccard coefficient is another common neighbors-based metric and is defined as the ratio of common co-authors to the total number of possible co-authors. Similar to the common neighbors metric, the positive impact of this metric on the new co-authorship pattern shows that authors with a higher ratio of common co-authors are more likely to collaborate in future. Adamic-Adar index is another common neighbors-based metric that takes into account the number of common co-authors. However, it assigns greater weights to common co-authors who have fewer co-authors. The positive impact of the Adamic-Adar index on the appearance of a new co-authorship link shows that for two author pairs who have not collaborated before, with the same number of common co-authors, the author pair, for whom, their common co-authors have fewer co-authors are more likely to form a co-authorship link in future than the author pair, for whom, their co-authors have higher number of co-authors.

The importance of the common neighbors-based metrics such as Adamic-Adar index and common neighbors for the new co-authorship pattern is in line with the most recent link prediction leaderboard on the ogbl-collab[12] co-authorship benchmark dataset named open co-authorship benchmark graph dataset (Wang et al. 2020). according to this leaderboard, the Adamic-Adar

[12] "The ogbl-collab dataset is an undirected graph, representing a subset of the collaboration network between authors indexed by Microsoft Academic Graph. Each node represents an author and edges indicate the collaboration between authors" https://ogb.stanford.edu/docs/linkprop/#ogbl-collab

index and common neighbors metric are ranked the third and the fourth among the top performing link prediction algorithms respectively. It is essential to note that all the aforementioned common neighbors-based metrics have zero parameters which makes them highly suitable for parallelization and fast computation in distributed environments where the structure of a graph/network might not be included in every computational node (Martínez et al. 2016). In contrast, most of the other advanced algorithms for link prediction in the leaderboard have many parameters. For instance, according to the ogbl-collab link prediction leaderboard[14], Node2vec algorithm has 30 million parameters while their performance on ogbl-collab is approximately 23% less than the performance of the Adamic-Adar index. The higher number of parameters requires heavier computation and subsequently more infrastructure cost which might be considered as a weakness for these algorithms.

Furthermore, weighted common neighbors-based metrics such as weighted Jaccard coefficient, weighted Adamic Adar index and weighted common neighbors appeared among the driving factors of new co-authorship pattern along with their positive impact on this pattern. This indicates the importance of weighted similarity-based metrics and considering the authors' number of publications into the similarity-based link prediction metrics for the new co-authorship link prediction. This is in line with De Sá and Prudêncio (2011) study who showed that adding weight or the strength of collaboration to the similiarity-based link prediction metrics improves the co-authorship link prediction performance.

[14] The link prediction leaderboard on the ogbl-collab dataset shows researchers' state-of-the-art methods and their performances on the ogbl-collab benchmark dataset.
https://ogb.stanford.edu/docs/leader_linkprop/#ogbl-collab

Preferential attachment and weighted preferential attachment are other structure-based metrics (but not common neighbors-based metrics) that appeared among the driving factors for the new co-authorship pattern. However, they do not always have a positive or a negative impact on the new co-authorship pattern. In other words, the higher product of an author pairs' degrees in the co-authorship network does not necessarily indicate that this author pair is more likely to collaborate in future or vice versa. For instance, as seen in Figure 13, for co-authorship #1, preferential attachment has a negative impact on the new co-authorship pattern whereas in Figure 14, for co-authorship #2, it has a positive impact.

Attribute-based metrics including discipline similarity score, productivity level of authors and productivity similarity score are among the driving factors for the new co-authorship pattern. The attribute-based metrics, as the name suggests, are solely based on the attributes of authors, and are extracted from the individual authors. The discipline similarity score compares the discipline of authors and assigns a score representing how similar a pair of authors are in terms of their discipline and research fields. As seen in Figures 13, 14 and 15, the discipline similarity score has a positive impact on the new co-authorship pattern. In other words, authors doing research within similar disciplines and research fields will be more likely to collaborate than authors from different disciplines. Productivity similarity score and productivity level of authors are another attribute-based driving factors that appeared among driving factors for the new co-authorship pattern. These productivity metrics use an authors' number of publications as a proxy for their productivity levels. These productivity metrics are found to have a negative impact on the new co-authorship pattern in most of the co-authorship networks. In other words, author pairs with similar productivity levels

and lower combined number of publications are more likely to collaborate for the first time in future.

Lastly, other metrics were also among the predictive features for the new co-authorship pattern, but since they showed to be less contributive than the discussed metrics to predict the new co-authorship links, we did not consider them as driving factors. For instance, the distance between a pair of authors in the network has an overall negative impact on the new co-authorship pattern meaning that two authors who are far from each other in the co-authorship network are less likely to collaborate for the first time than authors who are close to each other in the network. However, the shortest path metric has less contribution in new co-authorship link prediction compared to other discussed driving factors.

The importance of the discussed metrics for the new co-authorship link prediction reinforces the fact that there are hidden patterns in co-authorship networks that may indicate signals of new collaboration between researchers such that two scientists are more likely to collaborate if they have higher number of common co-authors and are similar to each other. To put it differently, the positive impact of the discussed metrics on the new co-authorship pattern, such as a higher number of common co-authors and similarity scores will eventually increase the probability of two authors communicating and subsequently collaborating.

5.3.2 Persistent co-authorship

This section investigates the performance of the classifiers and identifies the driving factors for the persistent co-authorship pattern. The objective is to predict the continuity of collaboration and interpret the driving factors for the appearance of future co-authorship links between author pairs who have collaborated on at least one joint paper before.

5.3.2.1 Machine learning models' performance results

The performance results for logistic regression, decision tree, random forests, and XGBoost models are shown in Tables 11, 12, 13, and 14.

Table 11. Logistic regression models' performance results for persistent co-authorship pattern

Co-authorship network number	F1 score	Recall	Precision	AUC	AP
1	0.86	0.79	0.95	0.82	0.97
2	0.86	0.81	0.91	0.69	0.95
3	0.83	0.74	0.93	0.76	0.95
4	0.82	0.74	0.92	0.77	0.92
5	0.80	0.73	0.90	0.74	0.92
6	0.73	0.61	0.91	0.70	0.91
7	0.85	0.82	0.87	0.77	0.93
8	0.84	0.78	0.90	0.79	0.92
9	0.84	0.79	0.90	0.82	0.93
10	0.80	0.72	0.90	0.74	0.92
11	0.78	0.67	0.93	0.79	0.93
12	0.82	0.74	0.92	0.79	0.94

Table 12. Decision tree models' performance results for persistent co-authorship pattern

Co-authorship network number	F1 score	Recall	Precision	AUC	AP
1	0.80	0.67	0.98	0.87	0.97
2	0.80	0.68	0.97	0.79	0.95
3	0.82	0.71	0.97	0.86	0.96
4	0.72	0.60	0.92	0.73	0.90
5	0.76	0.64	0.93	0.78	0.92
6	0.75	0.63	0.92	0.72	0.91
7	0.81	0.71	0.94	0.78	0.91
8	0.79	0.70	0.90	0.76	0.91
9	0.77	0.70	0.87	0.74	0.89
10	0.77	0.67	0.92	0.78	0.93
11	0.79	0.68	0.94	0.79	0.93
12	0.81	0.70	0.96	0.81	0.94

Table 13. Random forest models' performance results for persistent co-authorship pattern

Co-authorship network number	F1 score	Recall	Precision	AUC	AP
1	0.86	0.78	0.97	0.90	0.98
2	0.84	0.73	0.97	0.84	0.98
3	0.83	0.72	0.98	0.87	0.97
4	0.80	0.71	0.92	0.80	0.94
5	0.82	0.73	0.93	0.84	0.96
6	0.78	0.67	0.94	0.79	0.95
7	0.84	0.77	0.93	0.85	0.96
8	0.79	0.69	0.92	0.81	0.94
9	0.83	0.77	0.91	0.85	0.95
10	0.82	0.73	0.93	0.81	0.95
11	0.81	0.71	0.94	0.86	0.97
12	0.83	0.74	0.95	0.85	0.96

Table 14. XGBoost models' performance results for persistent co-authorship pattern

Co-authorship network number	F1 score	Recall	Precision	AUC	AP
1	0.84	0.76	0.94	0.88	0.97
2	0.82	0.71	0.95	0.82	0.97
3	0.84	0.74	0.97	0.84	0.96
4	0.78	0.68	0.92	0.78	0.93
5	0.81	0.71	0.96	0.85	0.97
6	0.78	0.67	0.92	0.75	0.93
7	0.86	0.81	0.91	0.82	0.95
8	0.80	0.73	0.89	0.78	0.94
9	0.85	0.80	0.90	0.84	0.95
10	0.84	0.76	0.93	0.85	0.96
11	0.81	0.71	0.95	0.85	0.96
12	0.88	0.82	0.95	0.86	0.96

We compared the AP performance of these classifiers against each other and the two baselines in Figure 16. As seen in this figure, the performance of all the classifiers was higher than the prevalence baseline, which shows the higher performance of these classifiers compared to the random predictor to predict the persistent co-authorship links. Random forest and XGBoost classifiers performed better than the logistic regression baseline in all co-authorship networks. This shows the higher ability of these two models to predict the persistent co-authorship links. However, the decision tree classifier did not always outperform the logistic regression baseline. In co-authorship networks #1, #4, #6, #7, #8, #9, #11 and #12, logistic regression classifier outperformed decision tree.

As seen in Figure 16, all these classifiers had AP performance close to each other ranging from 0.89 to 0.98. Similar to the new co-authorship pattern, random forests provided the best

performance for all co-authorship networks except for the co-authorships #5 and #10, in which XGBoost outperformed it.

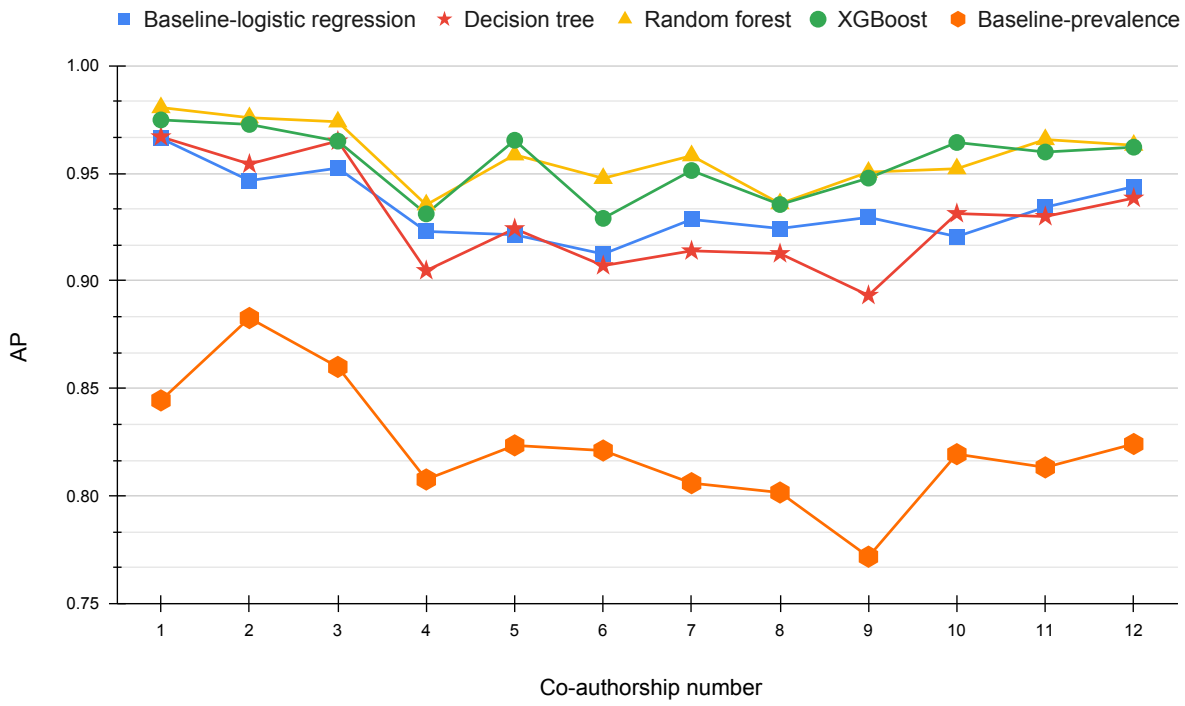


Figure 16. Comparison of models' AP performances against each other and the baselines for the persistent co-authorship pattern

The average recall for logistic regression, decision tree, random forests and XGBoost was 0.75, 0.67, 0.73 and 0.74, respectively. In other words, these classifiers were able to predict approximately 70% of the persistent co-authorship links. This is encouraging since these models can be used in co-author recommender systems to predict persistent collaboration links with a good performance.

The average precision for logistic regression, decision tree, random forests, and XGBoost classifiers was 0.91, 0.93, 0.94 and 0.93, respectively. The higher precision of these classifiers indicates that a high proportion of the predicted persistent co-authorship links actually occurred, which again supports the ability of these classifiers to predict persistent co-authorship links. We also saw a high performance based on F1 scores that indicate that the models greatly fit the data. Moreover, the high AUC of the classifiers confirms the ability of the models to distinguish well between the presence and the absence of the persistent co-authorship links.

5.3.2.2 Driving factors

This section discusses the driving factors for the persistent co-authorship pattern. We selected three co-authorship networks (#7, #9 and #11) whose desired co-authorship links were predicted by classifiers with the highest performance concerning the AP and recall metrics. In this case, the desired co-authorship link is the reappearance of a co-authorship link among pairs of authors who have collaborated at least on one joint paper before. As seen in Figures 17, 18 and 19, and similarly to the new co-authorship pattern, both structure-based and attribute-based metrics appeared among the predictive features for the persistent co-authorship pattern. It is thus important to consider both types of metrics to predict the persistent co-authorship links.

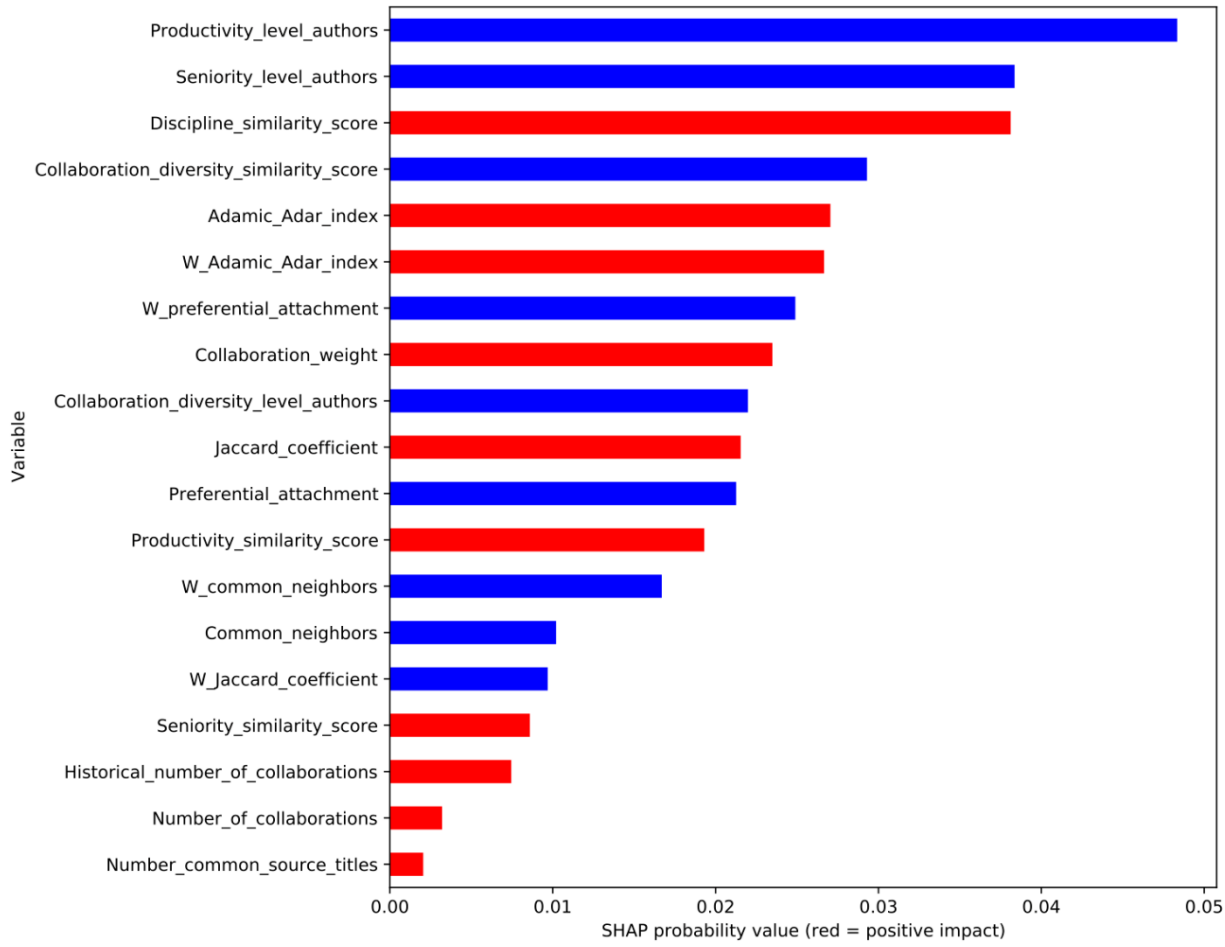


Figure 17. Co-authorship network #7: Driving factors for persistent co-authorship pattern

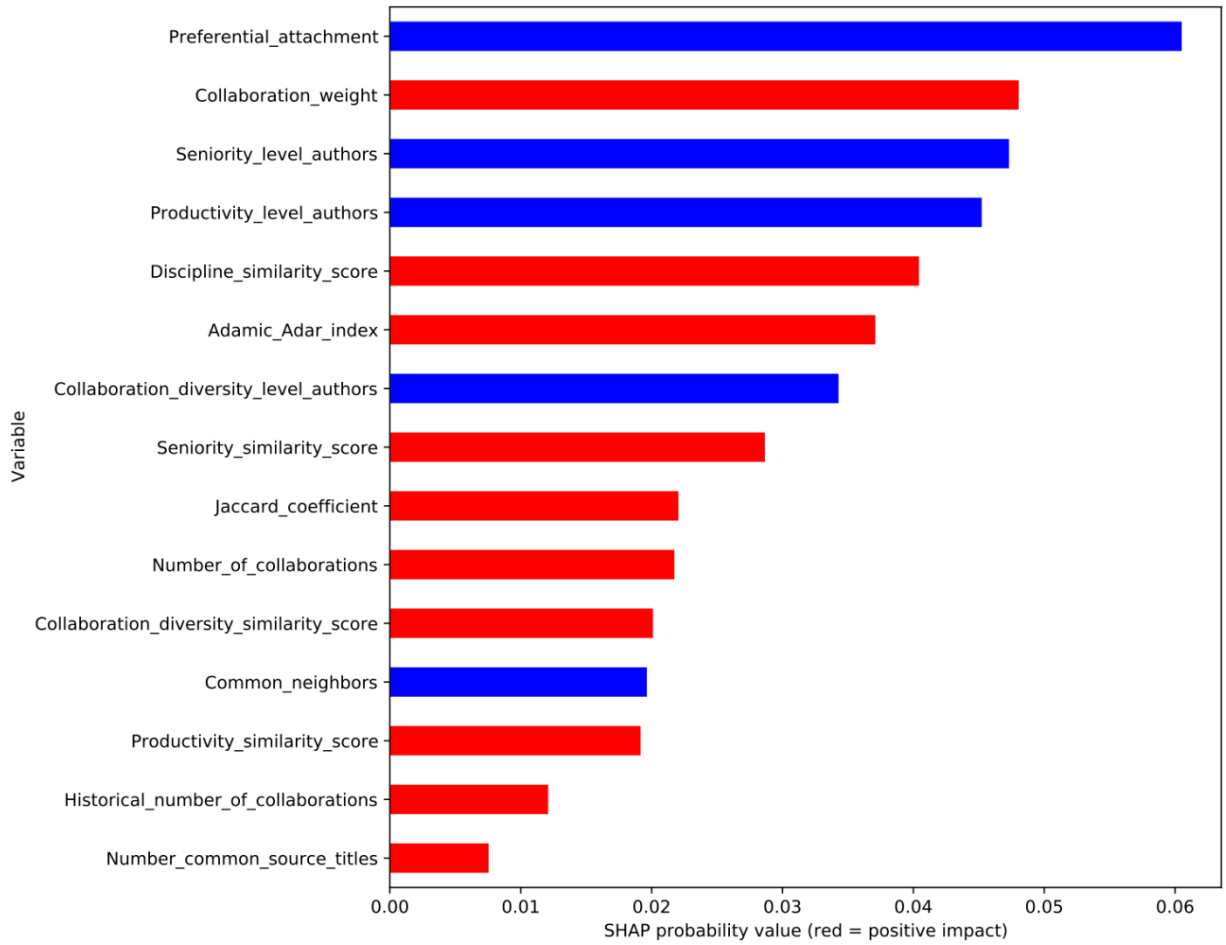


Figure 18. Co-authorship network #9: Driving factors for persistent co-authorship pattern

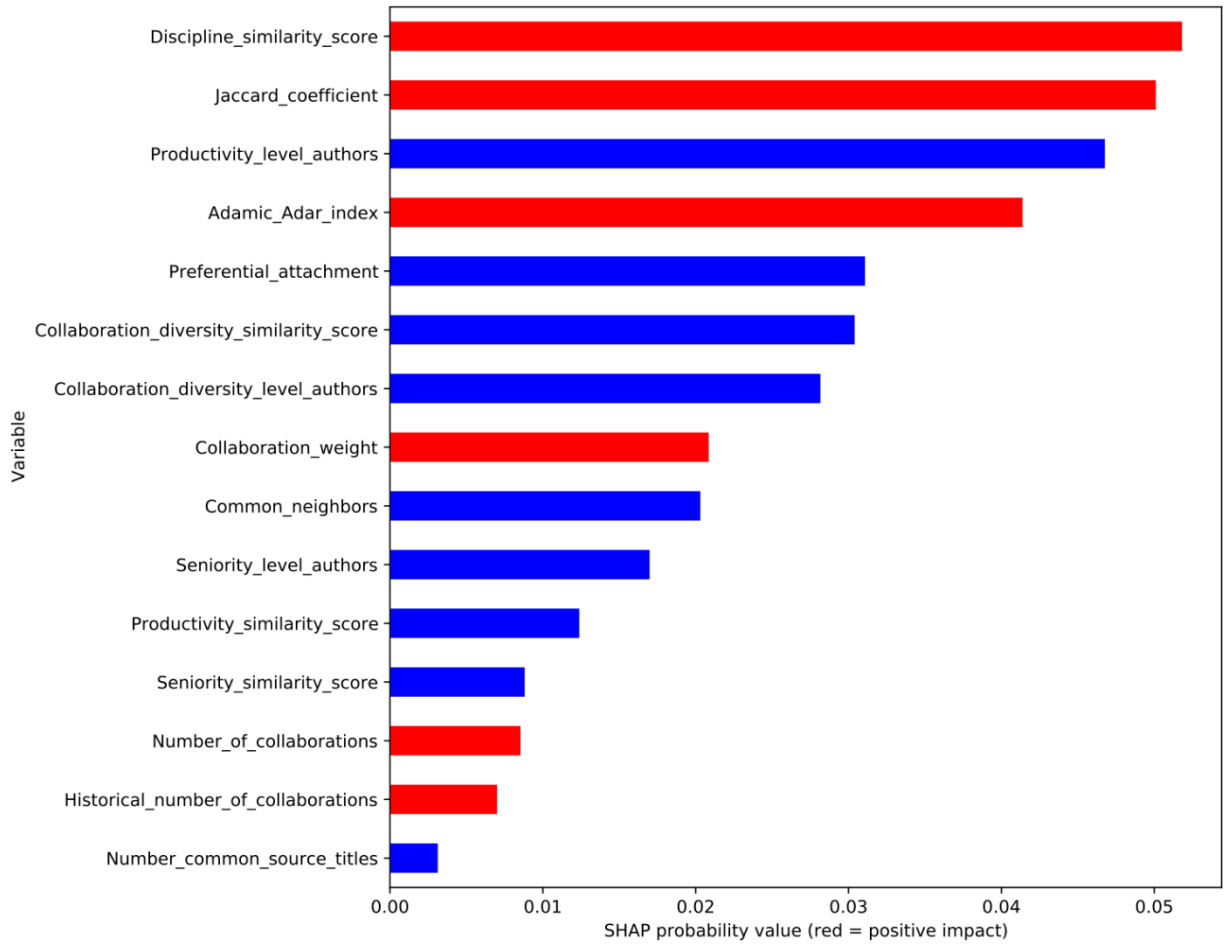


Figure 19. Co-authorship network #11: Driving factors for persistent co-authorship pattern

Similar to the new co-authorship pattern, we chose the following metrics as the driving factors for the persistent co-authorship pattern, since these features were among the most predictive features in all 12 co-authorship networks. However, other metrics such as the number of common source titles have shown to be less contributive than the discussed metrics to predict the persistent co-authorship links. Thus, we did not consider them as driving factors.

- The productivity level of authors
- Discipline similarity score
- Seniority level of authors
- Preferential attachment, weighted and unweighted
- Collaboration weight
- Collaboration diversity similarity score
- Jaccard coefficient
- Adamic-Adar index

The attribute-based factors include the productivity level of authors, the discipline similarity score, and the seniority level of authors. The productivity level of authors appeared among the driving factors for the persistent co-authorship pattern. As discussed, it is defined as the sum of the number of publications for an author pair. As shown in Figures 17, 18 and 19, it has a negative impact on the persistent co-authorship link. In other words, author pairs who are more productive together and have at least published one joint paper together before are less likely to collaborate again in future. Similar to the productivity level of authors, preferential attachment and weighted preferential attachment as structure-based metrics negatively impact the presence of continuous

co-authorship links. The negative impact of the productivity level of authors and unweighted and weighted preferential attachment on the persistent co-authorship link might be due to the productive author pairs who are more visible in the research community. Consequently, other authors can easily find them and collaborate with them, which results in a reduced probability of collaborating again with their old peers (Yu et al. 2014).

The discipline similarity score has a positive impact on the persistent co-authorship pattern and author pairs doing research within similar disciplines are more likely to collaborate again in future. The seniority level of authors is another attribute-based metric that has an overall negative impact on the persistent co-authorship pattern. In other words, author pairs with longer career ages are less likely to collaborate again in future than young author pairs. Preferential attachment and weighted preferential attachment as structure-based metrics negatively impact the presence of continuous co-authorship links.

Collaboration weight is an engineered feature that we added to take into account the number of joint publications for author pairs and the number of authors in their publications. The collaboration weight metric assigns greater weights to publications with a fewer number of authors. As seen in Figures 17, 18 and 19, collaboration weight has a positive impact on the persistent co-authorship pattern. In other words, author pairs who have had a higher number of joint publications with fewer authors are more likely to collaborate again in future than author pairs who collaborated in publications in which a higher number of authors was involved.

Collaboration diversity similarity score, a structure-based metric, also appeared among the persistent co-authorship pattern's driving factors. Collaboration diversity similarity score measures how similar author pairs are in terms of the diversity of their collaboration with their co-authors. Based on our results, this metric sometimes has a positive and sometimes negative impact on the persistent co-authorship pattern. In some co-authorship networks such as co-authorships #7 and #11, as shown in Figure 17 and 19, the collaboration diversity similarity score has a negative impact meaning that in these networks, authors are more likely to collaborate again with authors who have similar collaboration diversity levels (the number of collaborations with different co-authors) whereas the opposite case is true for the co-authorship network #9 in Figure 18.

Similar to the driving factors for the new co-authorship pattern, Jaccard coefficient and Adamic-Adar index are among the driving factors for the persistent co-authorship pattern with a positive impact. Therefore, author pairs with a higher ratio/number of common co-authors are more likely to collaborate again in future than author pairs with fewer common co-authors. As discussed, other metrics are also predictive of the persistent co-authorship links, however, they contribute less than the discussed metrics. Like the new co-authorship pattern, we observed that two scientists are more likely to collaborate again in future if they have had a higher number of joint publications in the past, had a higher number of common co-authors and had similar research profiles.

5.3.3 Discontinued co-authorship

This section discusses the performance of the classifiers in the discontinued co-authorship pattern. The objective was to predict the discontinued co-authorship links and shed light on the driving factors.

5.3.3.1 Machine learning models' performance results

As discussed, we trained four classifiers for each of the twelve co-authorship networks to predict the discontinued co-authorship links. Tables 15, 16, 17 and 18 provide the performance results of logistic regression, decision tree, random forests, and XGBoost, respectively.

Table 15. Logistic regression models' performance results for discontinued co-authorship pattern

Co-authorship network number	F1 score	Recall	Precision	AUC	AP
1	0.40	0.71	0.28	0.67	0.40
2	0.45	0.58	0.37	0.77	0.44
3	0.46	0.68	0.35	0.80	0.48
4	0.39	0.54	0.30	0.68	0.40
5	0.29	0.38	0.23	0.60	0.38
6	0.39	0.66	0.28	0.68	0.41
7	0.40	0.54	0.31	0.73	0.51
8	0.53	0.78	0.40	0.86	0.68
9	0.53	0.55	0.52	0.76	0.62
10	0.46	0.62	0.37	0.77	0.49
11	0.53	0.80	0.40	0.82	0.54
12	0.52	0.72	0.41	0.79	0.46

Table 16. Decision tree models' performance results for discontinued co-authorship pattern

Co-authorship network number	F1 score	Recall	Precision	AUC	AP
1	0.52	0.93	0.36	0.89	0.48
2	0.34	0.58	0.24	0.74	0.27
3	0.56	0.75	0.45	0.82	0.48
4	0.42	0.63	0.31	0.66	0.30
5	0.48	0.69	0.36	0.75	0.33
6	0.45	0.80	0.32	0.72	0.30
7	0.56	0.75	0.45	0.78	0.41
8	0.54	0.76	0.42	0.79	0.42
9	0.54	0.67	0.45	0.76	0.45
10	0.51	0.67	0.41	0.80	0.46
11	0.60	0.78	0.48	0.82	0.48
12	0.48	0.58	0.42	0.75	0.36

Table 17. Random forest models' performance results for discontinued co-authorship pattern

Co-authorship network number	F1 score	Recall	Precision	AUC	AP
1	0.48	0.79	0.34	0.86	0.64
2	0.45	0.68	0.33	0.83	0.46
3	0.55	0.82	0.42	0.89	0.73
4	0.46	0.80	0.32	0.70	0.37
5	0.48	0.67	0.37	0.79	0.50
6	0.40	0.70	0.28	0.75	0.43
7	0.53	0.77	0.41	0.82	0.56
8	0.57	0.84	0.43	0.87	0.64
9	0.58	0.65	0.52	0.84	0.69
10	0.51	0.73	0.39	0.84	0.61
11	0.55	0.81	0.42	0.86	0.61
12	0.50	0.70	0.39	0.81	0.55

Table 18. XGBoost models' performance results for discontinued co-authorship pattern

Co-authorship network number	F1 score	Recall	Precision	AUC	AP
1	0.48	0.79	0.34	0.86	0.65
2	0.38	0.63	0.27	0.77	0.43
3	0.50	0.82	0.36	0.89	0.61
4	0.35	0.51	0.27	0.64	0.37
5	0.45	0.64	0.35	0.76	0.47
6	0.44	0.68	0.32	0.77	0.46
7	0.53	0.79	0.39	0.83	0.59
8	0.58	0.88	0.43	0.86	0.60
9	0.59	0.69	0.51	0.79	0.61
10	0.52	0.81	0.38	0.85	0.57
11	0.58	0.85	0.45	0.88	0.64
12	0.48	0.66	0.38	0.79	0.50

Figure 20 compares the AP performance results of these classifiers against each other and the two baselines. As seen in this figure, all classifiers outperformed the prevalence baseline which shows the higher ability of these models compared to the random predictor to forecast future discontinued co-authorship links. Overall, random forests and XGBoost classifiers provided the best performance in predicting the discontinued co-authorship links. Similar to the persistent co-authorship pattern, random forest and XGBoost outperformed the logistic regression model in almost all co-authorship networks. The decision tree classifier was outperformed by logistic regression in all co-authorship networks except co-authorship #1 and #3.

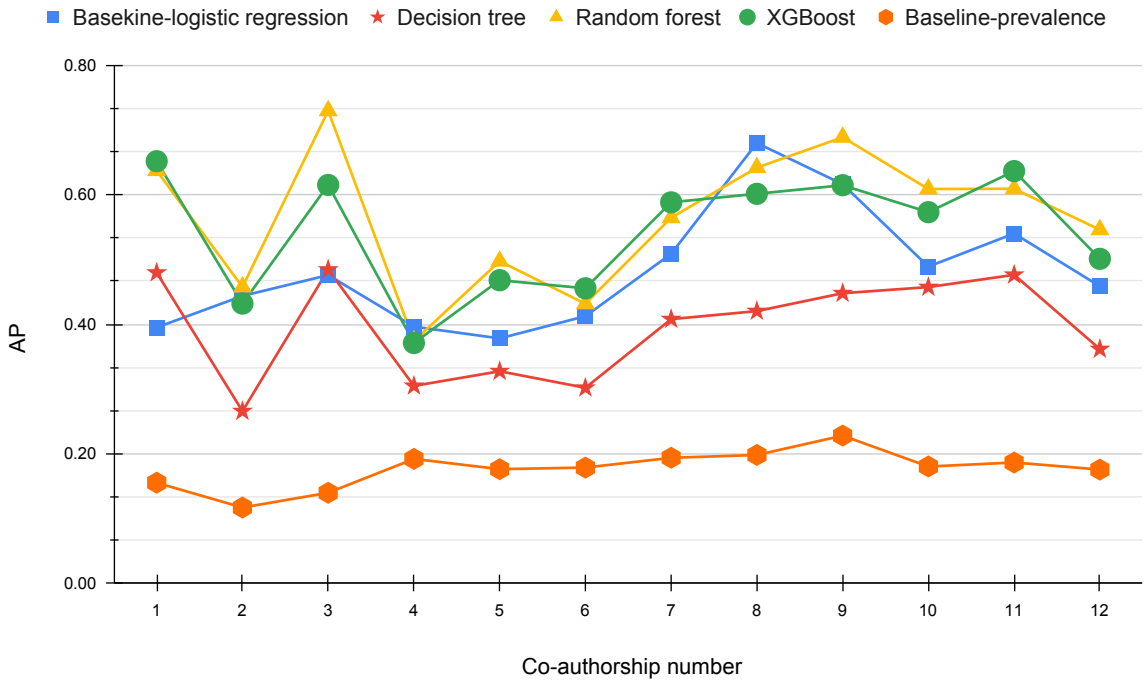


Figure 20. Comparison of models' AP performances against each other and the baselines for discontinued co-authorship pattern

The average recall (the ability of the model to predict the discontinuity of collaboration for author pairs) for the logistic regression, decision tree, random forests, and XGBoost classifiers was 0.63, 0.72, 0.75 and 0.73, respectively. Therefore, the random forests classifier provided the highest average recall and on average was able to predict 75% of discontinued co-authorship links. This is encouraging since these models can be used to predict which authors pairs are more likely to stop collaborating again in future.

5.3.3.2 Driving factors for the discontinued co-authorship pattern

This section discusses the driving factors for the discontinued co-authorship pattern. Like the other two co-authorship patterns, we investigated the classifiers with the highest performance concerning the AP and recall metrics and selected their corresponding co-authorship networks (#1, #3 and #8). As seen in Figures 21, 22 and 23, different types of metrics, including structure-based and attribute-based metrics, appeared among predictive features for the discontinued co-authorship pattern. Similar to the other co-authorship patterns, this shows the importance of using both the structure-based and attribute-based factors for co-authorship prediction.

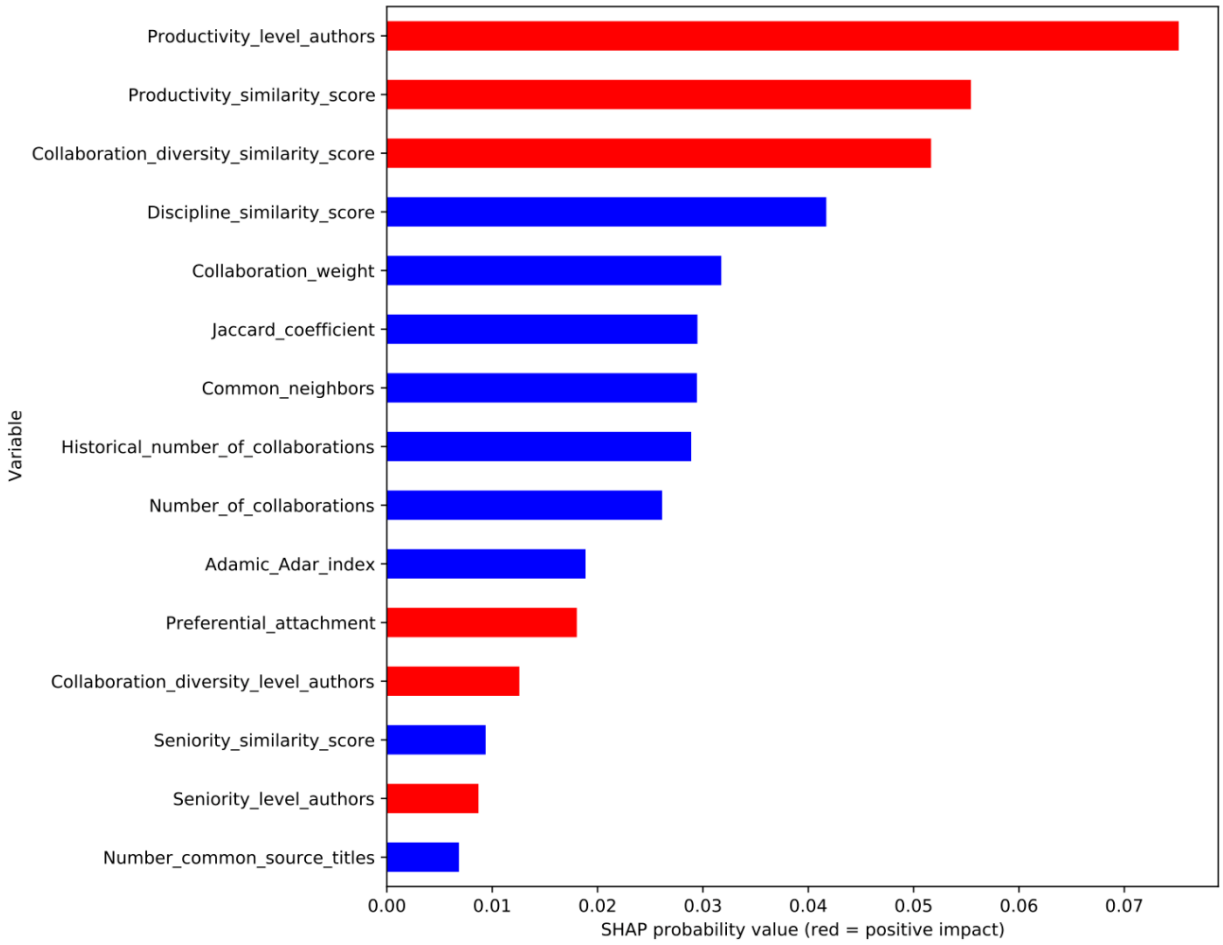


Figure 21. Co-authorship network #1: Driving factors for discontinued co-authorship pattern

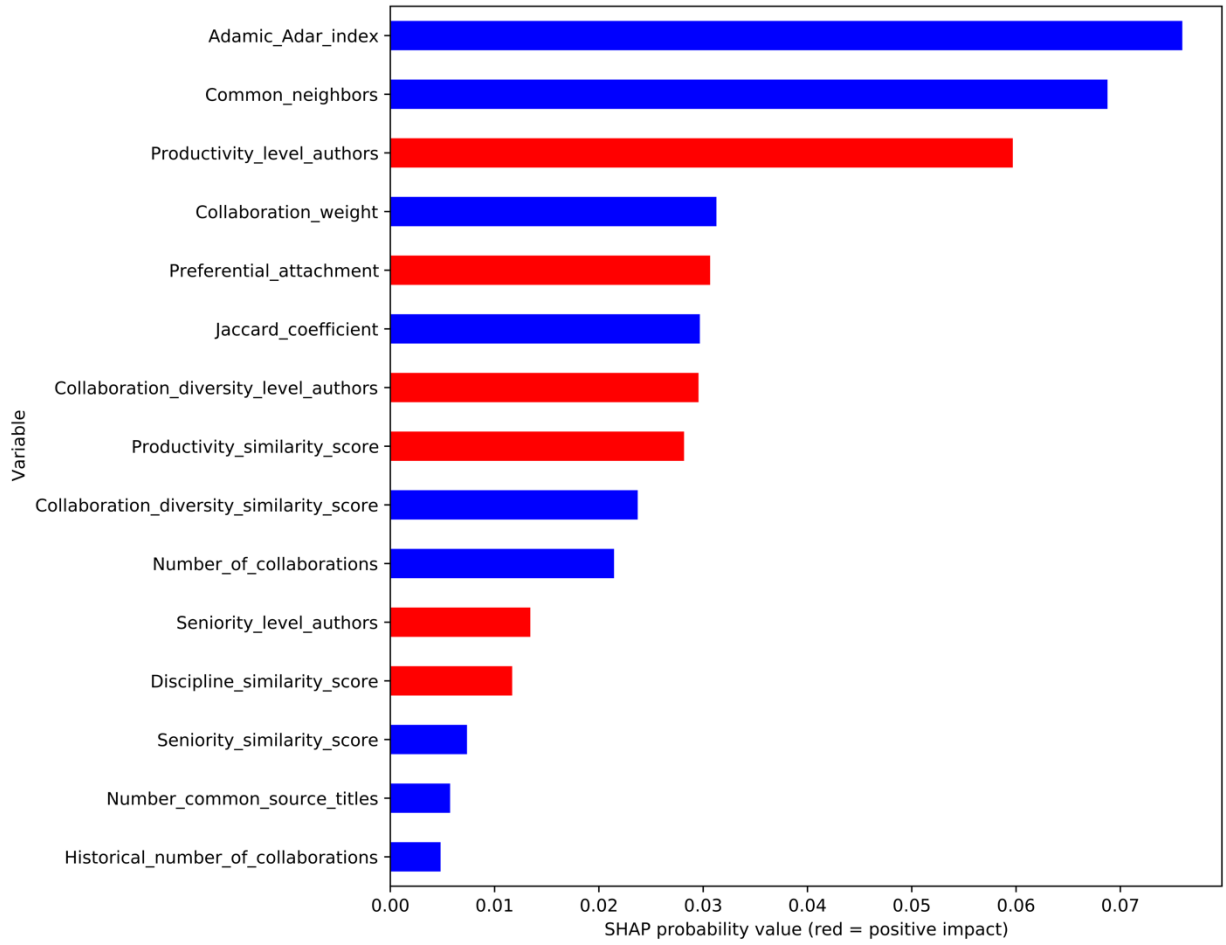


Figure 22. Co-authorship network #3: Driving factors for discontinued co-authorship pattern

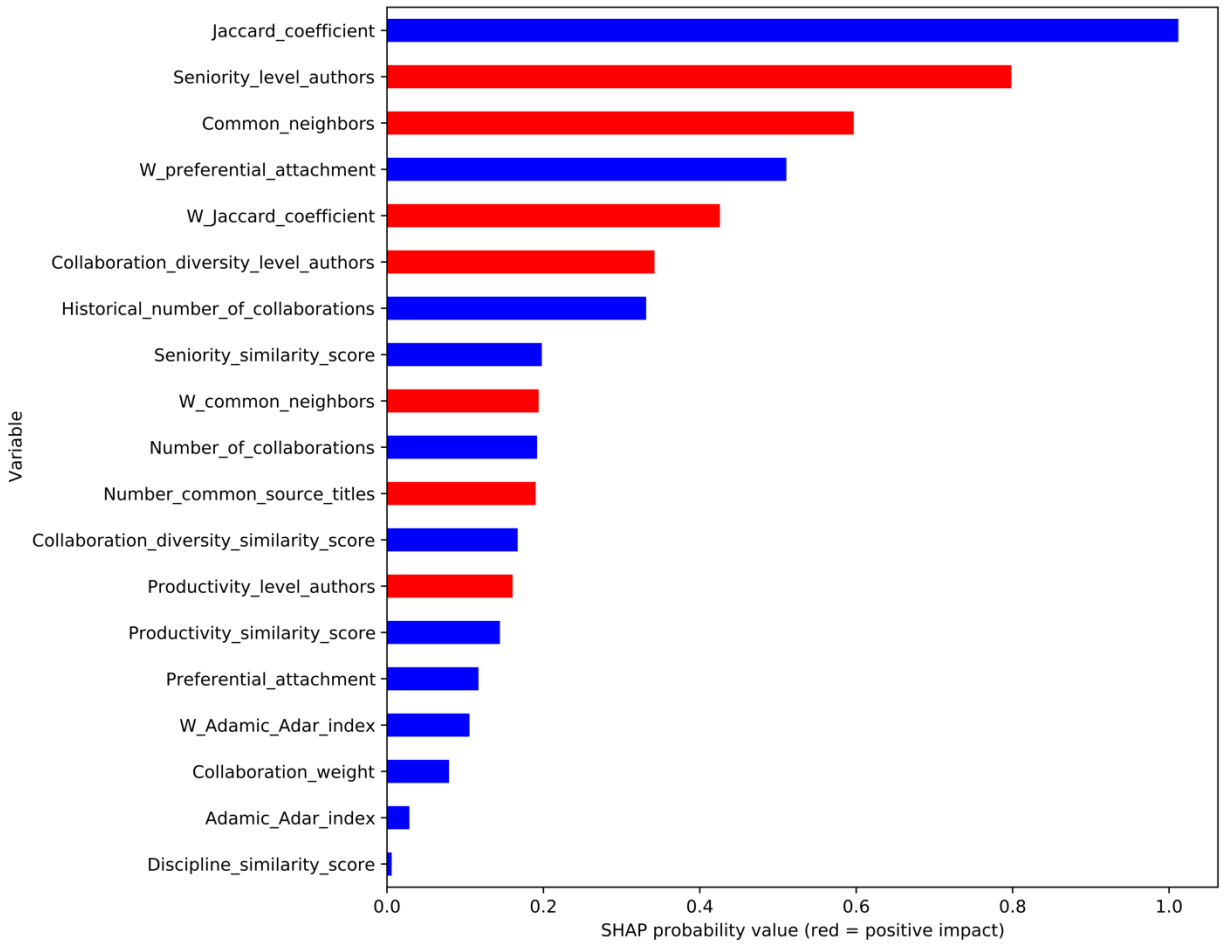


Figure 23. Co-authorship network #8: Driving factors for discontinued co-authorship pattern

We selected the most predictive features as the driving factors for the discontinued co-authorship pattern since these metrics were among the most predictive features in all 12 co-authorship networks however other metrics such as number of common source titles have shown to be less predictive of discontinued co-authorship links. The driving factors are as follows:

- Productivity level of authors
- Productivity similarity score
- Collaboration diversity similarity score

- Discipline similarity score
- Seniority level of authors
- Collaboration weight
- Jaccard coefficient, weighted and unweighted
- Adamic-Adar index
- Common neighbors
- Preferential attachment, weighted and unweighted

As discussed, the productivity level of authors metric indicates the number of publications of authors pairs and is calculated as the sum of their number of publications. As seen in Figure 21, 22, and 23, in contrast to the persistent co-authorship pattern, the productivity level of authors has a positive impact on the discontinued co-authorship links. To put it differently, author pairs with higher productivity levels are more likely to stop collaborating in future. The reason could be author pairs who have collaborated once. Still, the higher productivity level of these authors might lead them to collaborate on more papers that might involve different researchers, thus increasing the probability of them not collaborating again with their previous collaborators. The impact of some similarity scores including productivity similarity score and collaboration diversity similarity score, which measure how similar two authors are in terms of the number of publications and the collaboration diversity levels, respectively, are not always positive. The discipline similarity score has a negative impact on discontinued co-authorship links. In other words, authors doing research within less similar disciplines are most likely to stop collaborating in future. The seniority level of authors, calculated as the sum of the career age of authors in pairs, appeared among the driving factors for the discontinued co-authorship pattern and has a positive impact. In other words, author

pairs whose career age combined are higher are more likely to stop collaborating in future than author pairs with shorter career ages. Collaboration weight, which considers the number of joint publications of author pairs and the number of co-authors in their publications, has a negative impact on discontinued co-authorship links. In other words, authors pairs with higher number of joint publications in which fewer authors were involved are less likely to stop collaborating in future. This behaviour might result from strong collaboration ties that these author pairs have shown in the past, which might encourage their further collaboration and prevent co-authorship discontinuation in future.

Structure-based metrics such as unweighted and weighted Jaccard coefficient, Adamic-Adar index, common neighbors, weighted and unweighted preferential attachment are among the driving factors influencing the discontinued co-authorship pattern. However, we saw that the impact of these metrics is not always positive or consistently negative for the discontinued co-authorship pattern.

Similar to the previous co-authorship patterns, other metrics such as number of common source titles metric are also predictive of the discontinued co-authorship links, but they have shown lower contribution than the already discussed metrics. Like the new and persistent co-authorship patterns, we saw similar behaviours in discontinued co-authorship pattern that resemble the behaviours in social networks. For example, collaboration weight is similar to the strength of relationship tie between two entities in a social network. For instance in a social networking application, two users might stop having relationships with each other because their relationship ties have become weaker over time.

Chapter 6: Conclusions and limitations

6.1 Conclusions

This study explored the interdisciplinary research field of AI for cancer diagnosis and treatment and achieved two main research objectives. For the first research objective, the characterization of AI applications for cancer diagnosis and treatment, we proposed an approach to better understand the AI in cancer research landscape and investigated how researchers contributed to this innovation ecosystem and used advanced techniques to improve patients' outcomes by extracting the main research themes and assessing their temporal evolution. The DTM algorithm allowed us to dynamically calculate topic proportions over time. Our findings confirm the growth of using machine/deep learning techniques in analyzing healthcare data. Especially, medical image analytics and deep convolutional neural networks were found to be promising directions in recent years in analyzing enormous sets of medical imaging. The appearance of the “clinical decision support systems” as one of the prevalent topics was also noticeable. Clinical decision support systems take advantage of different techniques in multidisciplinary areas such as AI, machine learning, and statistical pattern recognition to support and improve decision-making in the healthcare ecosystem. The significant and increasing investment in AI in recent years is reflected in the main research themes as well as their trends, moving from conventional to advanced learning techniques.

The second research objective was to predict different co-authorship links and identify the driving factors for different co-authorship patterns of researchers with the main focus on the topic of AI in cancer research. We extracted and calculated different structure-based and attribute-based

metrics from the co-authorship networks and individual authors. These features served as inputs for ML classifiers used to predict the co-authorship links of three different co-authorship patterns including new, persistent, and discontinued co-authorship patterns. The ML classifiers provided an encouraging performance to predict future co-authorship links of different patterns. Our results confirmed the importance of considering both structure-based and attribute-based metrics for co-authorship link prediction. Moreover, we observed a high similarity of co-authorship networks to social networks. For instance, in social network applications such as Instagram[15], users are more likely to follow similar and like-minded people. Also, a higher number of mutual friends between two users might increase their chance of following each other. We also shed light on the driving factors for different co-authorship patterns. For the new and persistent co-authorship patterns, our results showed that common neighbors-based factors have a positive impact on the appearance of co-authorship links. In other words, authors with a higher number of common co-authors are more likely to collaborate for the first time and again in future. Another driving factor for all co-authorship patterns was found to be the discipline similarity score having a positive impact on the new and persistent co-authorship patterns and a negative impact on the discontinued co-authorship pattern. In other words, authors doing research within similar fields are more likely to collaborate again in future or for the first time and less likely to stop collaborating. The encouraging co-authorship prediction performance of the discussed classifiers and their interpretable results for different co-authorship patterns could enable scientists worldwide to locate potential collaborators and help research organizations organize and build strong research teams.

[15] Instagram is a photo and video sharing social networking service. https://www.instagram.com/

6.2 Limitations and future research

We were exposed to some limitations in this study. In terms of the data sources, for the first research objective, which is the characterization of AI for cancer diagnosis and treatment, we extracted publications from two data sources including Scopus and PubMed. For the second research objective, we only considered Scopus for constructing co-authorship networks due to the better coverage of Scopus publications than PubMed in the field of AI in cancer. Even though the discussed databases are considered comprehensive and optimal tools in the medical domain (Falagas et al. 2008), to have the most comprehensive view, other data sources could be also considered. We also excluded non-English publications in this study. The findings of this research may only shed light on the research themes of the target researchers at a very high level. A future research direction could be performing a similar analysis at a different level of granularity. We extracted uni-grams and used them as the input to the model. Future research can examine n-grams (n>1). We used the DTM algorithm to extract the "AI in cancer" research themes. Our proposed methodology could be applied to other innovation ecosystems, with proper tuning of the model based on the target ecosystem. We analyzed the prevalence of the extracted topics and their temporal evolution. Future research can investigate the existence of topics fusion and/or division phenomenon over time. Another future direction would be analyzing the entire body of publications and/or the methods sections as it may better reveal methodological evolution. Last but not least, the increasing trend of "AI in cancer" publications as well as the highly interdisciplinary nature of the field, suggest performing such analysis on a regular and frequent basis.

For the second research objective, it is important to note that co-authorship and collaboration are not always equivalent and two authors collaborating does not necessarily result in co-authorship

and also not all co-authored papers are the results of collaboration (Yu et al. 2014). In this study, we extracted structure-based metrics from homogeneous co-authorship networks as inputs for ML classifiers and SHAP approach for co-authorship link prediction and identification of driving factors of collaboration respectively. In these homogeneous co-authorship networks, there is one type of node (author) and one type of edge (co-authorship link). In future research, heterogeneous co-authorship networks can be constructed. In these types of networks, there are various type of nodes (e.g., author, institute, paper, field and venue) and various types of edges (e.g., author is_(first/last/other)_author_of paper, author is_affiliated_with institute, paper is_published_(conference/journal)_at venue, paper has_filed_of field, paper has_citation_to paper). These heterogeneous co-authorship networks can be constructed to extract more features that can be used by more advanced techniques for co-authorship link prediction and subsequently a much better understanding of driving factors for co-authorship (Hu et al. 2020).

7. References

Abbod, M. F., Catto, J. W. F., Linkens, D. A., & Hamdy, F. C. (2007). Application of Artificial Intelligence to the Management of Urological Cancer. *Journal of Urology*, *178*(4), 1150–1156. https://doi.org/10.1016/j.juro.2007.05.122

About | Elsevier Scopus Blog. (2021). https://blog.scopus.com/about. Accessed 17 November 2021

Acedo, F. J., Barroso, C., Casanueva, C., & Galán, J. L. (2006). Co-authorship in management and organizational studies: An empirical and network analysis. *Journal of Management Studies*, *43*(5), 957–983. https://doi.org/10.1111/j.1467-6486.2006.00625.x

Alakwaa, W., Nassef, M., & Badr, A. (2017). Lung cancer detection and classification with 3D convolutional neural network (3D-CNN). *International Journal of Biology and Biomedical Engineering*, *11*(8), 66–73. https://doi.org/10.14569/ijacsa.2017.080853

Anwar, S. M., Majid, M., Qayyum, A., Awais, M., Alnowami, M., & Khan, M. K. (2018). Medical Image Analysis using Convolutional Neural Networks: A Review. *Journal of Medical Systems*, *42*(11), 1–13. https://doi.org/10.1007/s10916-018-1088-1

Apté, C., & Weiss, S. (1997). Data mining with decision trees and decision rules. *Future Generation Computer Systems*, *13*(2–3), 197–210. https://doi.org/10.1016/s0167-739x(97)00021-6

Aslan, S., & Kaya, M. (2019). A Hybrid recommendation system in co-Authorship networks. *2019 International Conference on Artificial Intelligence and Data Processing Symposium, IDAP 2019*. https://doi.org/10.1109/IDAP.2019.8875989

Ayele, W. Y., & Juell-Skielse, G. (2020). Eliciting Evolving Topics, Trends and Foresight about Self-driving Cars Using Dynamic Topic Modeling. In *Advances in Intelligent Systems and*

Computing (Vol. 1129 AISC, pp. 488–509). Springer. https://doi.org/10.1007/978-3-030-39445-5_37

Bandodkar, N. R., & Grover, V. (2016). Factors influencing the extent of co-authorship in IS research: An empirical investigation. *Communications of the Association for Information Systems*, *38*(1), 84–105. https://doi.org/10.17705/1cais.03803

Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*(5439), 509–512. https://doi.org/10.1126/science.286.5439.509

Bashiri, A., Ghazisaeedi, M., Safdari, R., Shahmoradi, L., & Ehtesham, H. (2017). *Improving the Prediction of Survival in Cancer Patients by Using Machine Learning Techniques: Experience of Gene Expression Data: A Narrative Review*. *Iran J Public Health* (Vol. 46). http://ijph.tums.ac.ir

Beaver, D. de B., & Rosen, R. (1978). Studies in scientific collaboration - Part I. The professional origins of scientific co-authorship. *Scientometrics*, *1*(1), 65–84. https://doi.org/10.1007/BF02016840

Bidassie, B., Hoffman-Hōgg, L., Eapen, S., Aggarwal, A., Park, Y.-H. A., Keller, A., & Kelley, M. J. (2017). Cancer Care Collaborative Approach to Optimize Clinical Care. *Federal practitioner : for the health care professionals of the VA, DoD, and PHS*, *34*(Suppl 3), S42–S49. http://www.ncbi.nlm.nih.gov/pubmed/31089321. Accessed 22 May 2021

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. *ACM International Conference Proceeding Series*, *148*, 113–120. https://doi.org/10.1145/1143844.1143859

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). *Latent Dirichlet Allocation Michael I. Jordan*. *Journal of Machine Learning Research* (Vol. 3). https://doi.org/10.5555/944919.944937

Bosch, B., & Mansell, H. (2015). Interprofessional collaboration in health care: Lessons to be

learned from competitive sports. *Canadian Pharmacists Journal*, *148*(4), 176–179. https://doi.org/10.1177/1715163515588106

Bougrain, F., & Haudeville, B. (2002). Innovation, collaboration and SMEs internai research capacities. *Research Policy*, *31*(5), 735–747. https://doi.org/10.1016/S0048-7333(01)00144-5

Brazil, K., Whelan, T., O'Brien, M. A., Sussman, J., Pyette, N., & Bainbridge, D. (2004). Towards improving the co-ordination of supportive cancer care services in the community. *Health Policy*, *70*(1), 125–131. https://doi.org/10.1016/j.healthpol.2004.02.007

Breiger, R. L. (2004). The Analysis of Social Networks In: Handbook of Data Analysis. https://doi.org/10.4135/9781848608184

Buri, L., Hassan, C., Bersani, G., Anti, M., Bianco, M. A., Cipolletta, L., et al. (2010). Appropriateness guidelines and predictive rules to select patients for upper endoscopy: A nationwide multicenter study. *American Journal of Gastroenterology*, *105*(6), 1327–1337. https://doi.org/10.1038/ajg.2009.675

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-Augu*, 785–794. https://doi.org/10.1145/2939672.2939785

Chen, Z., Jiang, F., Cheng, Y., Gu, X., Liu, W., & Peng, J. (2018). XGBoost Classifier for DDoS Attack Detection and Analysis in SDN-Based Cloud. *Proceedings - 2018 IEEE International Conference on Big Data and Smart Computing, BigComp 2018*, 251–256. https://doi.org/10.1109/BigComp.2018.00044

Cheng, M. Y., Hen, K. W., Tan, H. P., & Fok, K. F. (2013). Patterns of co-authorship and research collaboration in Malaysia. *Aslib Proceedings: New Information Perspectives*,

65(6), 659–674. https://doi.org/10.1108/AP-12-2012-0094

Cho, H., Mariotto, A. B., Schwartz, L. M., Luo, J., & Woloshin, S. (2014). When do changes in cancer survival mean progress? The insight from population incidence and mortality. *Journal of the National Cancer Institute - Monographs*, *2014*(49), 187–197. https://doi.org/10.1093/jncimonographs/lgu014

Choudhary, A. K., Oluikpe, P. I., Harding, J. A., & Carrillo, P. M. (2009). The needs and benefits of Text Mining applications on Post-Project Reviews. *Computers in Industry*, *60*(9), 728–740. https://doi.org/10.1016/j.compind.2009.05.006

Chuan, P. M., Son, L. H., Ali, M., Khang, T. D., Huong, L. T., & Dey, N. (2018). Link prediction in co-authorship networks based on hybrid content similarity metric. *Applied Intelligence*, *48*(8), 2470–2486. https://doi.org/10.1007/s10489-017-1086-x

Cramer, J. S. (2005). The Origins of Logistic Regression. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.360300

Cronin, B., Shaw, D., & La Barre, K. (2003). A cast of thousands: Coauthorship and subauthorship collaboration in the 20th century as manifested in the scholarly journal literature of psychology and philosophy. *Journal of the American Society for Information Science and Technology*, *54*(9), 855–871. https://doi.org/10.1002/asi.10278

Cummings, J. N., & Kiesler, S. (2005). Collaborative research across disciplinary and organizational boundaries. *Social Studies of Science*, *35*(5), 703–722. https://doi.org/10.1177/0306312705055535

Davis, J., & Goadrich, M. (2006). The Relationship Between Precision-Recall and ROC Curves.

De Sá, H. R., & Prudêncio, R. B. C. (2011). Supervised link prediction in weighted networks. *Proceedings of the International Joint Conference on Neural Networks*, 2281–2288.

https://doi.org/10.1109/IJCNN.2011.6033513

Dias, R., & Torkamani, A. (2019). Artificial intelligence in clinical and genomic diagnostics. *Genome Medicine*, *11*(1), 1–12. https://doi.org/10.1186/s13073-019-0689-8

Dilsizian, S. E., & Siegel, E. L. (2014). Artificial intelligence in medicine and cardiac imaging: Harnessing big data and advanced computing to provide personalized medical diagnosis and treatment. *Current Cardiology Reports*, *16*(1). https://doi.org/10.1007/s11886-013-0441-8

E Fonseca, B. de P. F., Sampaio, R. B., Fonseca, M. V. de A., & Zicker, F. (2016). Co-authorship network analysis in health research: Method and potential use. *Health Research Policy and Systems*, *14*(1), 1–10. https://doi.org/10.1186/s12961-016-0104-5

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

ElShawi, R., Sherif, Y., Al-Mallah, M., & Sakr, S. (2020). Interpretability in healthcare: A comparative study of local machine learning interpretability techniques. *Computational Intelligence*, (September 2019), 1–18. https://doi.org/10.1111/coin.12410

Falagas, M. E., Pitsouni, E. I., Malietzis, G. A., & Pappas, G. (2008). Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. *The FASEB Journal*, *22*(2), 338–342. https://doi.org/10.1096/fj.07-9492lsf

Fanelli, D., & Larivière, V. (2016). Researchers' individual publication rate has not increased in a century. *PLoS ONE*, *11*(3), 1–12. https://doi.org/10.1371/journal.pone.0149504

Farris, F. A. (2010). The gini index and measures of inequality. *American Mathematical Monthly*, *117*(10), 851–864. https://doi.org/10.4169/000298910X523344

Ferroni, P., Zanzotto, F. M., Riondino, S., Scarpato, N., Guadagni, F., & Roselli, M. (2019).

Breast cancer prognosis using a machine learning approach. *Cancers*, *11*(3), 1–9. https://doi.org/10.3390/cancers11030328

Fleming, N. (2018). How artificial intelligence is changing drug discovery. *Nature*, *557*(7707), S55–S57. https://doi.org/10.1038/d41586-018-05267-x

Gallivan, M., & Ahuja, M. (2015). Co-authorship, homophily, and scholarly influence in information systems research. *Journal of the Association for Information Systems*, *16*(12), 980–1015. https://doi.org/10.17705/1jais.00416

Glänzel, W., & Schubert, A. (2006). Analysing Scientific Networks Through Co-Authorship. In *Handbook of Quantitative Science and Technology Research* (pp. 257–276). Kluwer Academic Publishers. https://doi.org/10.1007/1-4020-2755-9_12

Greene, D., & Cross, J. P. (2017). Exploring the political agenda of the european parliament using a dynamic topic modeling approach. *Political Analysis*, *25*(1), 77–94. https://doi.org/10.1017/pan.2016.7

Hagberg, A. A., Schult, D. A., & Swart, P. J. (2008). Exploring network structure, dynamics, and function using NetworkX. *7th Python in Science Conference (SciPy 2008)*, (SciPy), 11–15.

Hannigan, T. R., Haan, R. F. J., Vakili, K., Tchalian, H., Glaser, V. L., Wang, M. S., et al. (2019). Topic modeling in management research: Rendering new theory from textual data. *Academy of Management Annals*, *13*(2), 586–632. https://doi.org/10.5465/annals.2017.0099

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., et al. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hasan, M. Al, Chaoji, V., Salem, S., Zaki, M., & York, N. (2006). Link Prediction using Supervised Learning, (April 2020).

Hastie, T., Tibshirani, R., & Friedman, J. (2009). Springer Series in Statistics. *The Elements of Statistical Learning*, *27*(2), 83–85. http://www.springerlink.com/index/D7X7KX6772HQ2135.pdf. Accessed 11 July 2021

Hearst, M. A. (1999). Untangling text data mining, 3–10. https://doi.org/10.3115/1034678.1034679

Henriksen, D. (2016). The rise in co-authorship in the social sciences (1980–2013). *Scientometrics*, *107*(2), 455–476. https://doi.org/10.1007/s11192-016-1849-x

Ho, T. K. (1995). Random decision forests. *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, *1*, 278–282. https://doi.org/10.1109/ICDAR.1995.598994

homophily, n.2 : Oxford English Dictionary. (n.d.). https://www.oed.com/view/Entry/68972848?rskey=YrTgkc&result=2&isAdvanced=false#eid. Accessed 7 August 2021

Hu, Z., Dong, Y., Wang, K., & Sun, Y. (2020). Heterogeneous Graph Transformer. *The Web Conference 2020 - Proceedings of the World Wide Web Conference, WWW 2020*, 2704–2710. https://doi.org/10.1145/3366423.3380027

Huang, S., Yang, J., Fong, S., & Zhao, Q. (2020, February 28). Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges. *Cancer Letters*. Elsevier Ireland Ltd. https://doi.org/10.1016/j.canlet.2019.12.007

Husain, A., Barbera, L., Howell, D., Moineddin, R., Bezjak, A., & Sussman, J. (2013). Advanced lung cancer patients' experience with continuity of care and supportive care needs. *Supportive Care in Cancer*, *21*(5), 1351–1358. https://doi.org/10.1007/s00520-012-1673-7

Hwang, K. (2008). Periphery in the Globalization of Science, 101–133.

Ioannidis, J. P. A. (2008). Measuring co-authorship and networking-adjusted scientific impact. *PLoS ONE*, *3*(7), 1–8. https://doi.org/10.1371/journal.pone.0002778

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et du Jura. https://doi.org/10.5169/SEALS-266450

Jalalian, A., Mashohor, S. B. T., Mahmud, H. R., Saripan, M. I. B., Ramli, A. R. B., & Karasfi, B. (2013). Computer-aided detection/diagnosis of breast cancer in mammography and ultrasound: A review. *Clinical Imaging*, *37*(3), 420–426. https://doi.org/10.1016/j.clinimag.2012.09.024

Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., et al. (2017). Artificial intelligence in healthcare: Past, present and future. *Stroke and Vascular Neurology*, *2*(4), 230–243. https://doi.org/10.1136/svn-2017-000101

Jiang, H., Qiang, M., & Lin, P. (2016). A topic modeling based bibliometric exploration of hydropower research. *Renewable and Sustainable Energy Reviews*, *57*, 226–237. https://doi.org/10.1016/j.rser.2015.12.194

Jiang, N., & Xu, X. (2019). Exploring the survival prognosis of lung adenocarcinoma based on the cancer genome atlas database using artificial neural network. *Medicine*, *98*(20), e15642. https://doi.org/10.1097/MD.0000000000015642

Johri, A., Wang, G. A., Liu, X., & Madhavan, K. (2011). Utilizing topic modeling techniques to identify the emergence and growth of research topics in engineering education. *Proceedings - Frontiers in Education Conference, FIE*, 1–6. https://doi.org/10.1109/FIE.2011.6142770

Katuwal, G. J., & Chen, R. (2016). Machine Learning Model Interpretability for Precision Medicine. http://arxiv.org/abs/1610.09045

Katz, J. S., & Martin, B. R. (1997). What is research collaboration? *Research Policy*, *26*(1), 1–

18. https://doi.org/10.1016/S0048-7333(96)00917-1

Kirasich, K., Smith, T., & Sadler, B. (2018). Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets. *Data Science Review*, *1*(3), 9. https://scholar.smu.edu/datasciencereviewhttp://digitalrepository.smu.edu.Availableat:https://scholar.smu.edu/datasciencereview/vol1/iss3/9

Knoop, T., Wujcik, D., & Wujcik, K. (2017). Emerging Models of Interprofessional Collaboration in Cancer Care. *Seminars in Oncology Nursing*, *33*(4), 459–463. https://doi.org/10.1016/j.soncn.2017.08.009

Kostoff, R. N., Eberhart, H. J., & Toothman, D. R. (1999). Hypersonic and supersonic flow roadmaps using bibliometrics and database tomography. *Journal of the American Society for Information Science*, *50*(5), 427–447. https://doi.org/10.1002/(SICI)1097-4571(1999)50:5<427::AID-ASI5>3.0.CO;2-P

Kostoff, Ronald N., Braun, T., Schubert, A., Toothman, D. R., & Humenik, J. A. (2000). Fullerene Data Mining Using Bibliometrics and Database Tomography. *Journal of Chemical Information and Computer Sciences*, *40*(1), 19–39. https://doi.org/10.1021/ci990045n

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, *13*, 8–17. https://doi.org/10.1016/j.csbj.2014.11.005

Le, M. H., Chen, J., Wang, L., Wang, Z., Liu, W., Cheng, K. T., & Yang, X. (2017). Automated diagnosis of prostate cancer in multi-parametric MRI based on multimodal convolutional neural networks. *Physics in Medicine and Biology*, *62*(16), 6497–6514. https://doi.org/10.1088/1361-6560/aa7731

Li, H., Cao, Y., Li, S., Zhao, J., & Sun, Y. (2020). XGBoost Model and Its Application to Personal Credit Evaluation. *IEEE Intelligent Systems*, *35*(3), 52–61. https://doi.org/10.1109/MIS.2020.2972533

Li, X., Zhang, S., Zhang, Q., Wei, X., Pan, Y., Zhao, J., et al. (2019). Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study. *The Lancet Oncology*, *20*(2), 193–201. https://doi.org/10.1016/S1470-2045(18)30762-9

Liben-Nowell, D., & Kleinberg, J. (2003). The link prediction problem for social networks, 556. https://doi.org/10.1145/956863.956972

Liben-Nowell, D., & Kleinberg, J. (2007). The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, *58*(7), 1019–1031. https://doi.org/10.1002/asi.20591

Liew, X. Y., Hameed, N., & Clos, J. (2021). An investigation of XGBoost-based algorithm for breast cancer classification. *Machine Learning with Applications*, *6*(August), 100154. https://doi.org/10.1016/j.mlwa.2021.100154

Lisboa, P. J., & Taktak, A. F. G. (2006). The use of artificial neural networks in decision support in cancer: A systematic review. *Neural Networks*, *19*(4), 408–415. https://doi.org/10.1016/j.neunet.2005.10.007

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *2017-Decem*(Section 2), 4766–4775.

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., et al. (2018). Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology. *Communication Methods and Measures*, *12*(2–3), 93–118.

https://doi.org/10.1080/19312458.2018.1430754

Mak, K. K., & Pichika, M. R. (2019). Artificial intelligence in drug development: present status and future prospects. *Drug Discovery Today*, *24*(3), 773–780. https://doi.org/10.1016/j.drudis.2018.11.014

Martínez, V., Berzal, F., & Cubero, J. C. (2016). A survey of link prediction in complex networks. *ACM Computing Surveys*, *49*(4). https://doi.org/10.1145/3012704

Mazo, C., Kearns, C., Mooney, C., & Gallagher, W. M. (2020). Clinical decision support systems in breast cancer: A systematic review. *Cancers*, *12*(2), 1–15. https://doi.org/10.3390/cancers12020369

McKinney, W. (2011). pandas: a Foundational Python Library for Data Analysis and Statistics. *Python for High Performance and Scientific Computing*, (January 2011), 1–9.

Mehta, N., Pandit, A., & Shukla, S. (2019). Transforming healthcare with big data analytics and artificial intelligence: A systematic mapping study. *Journal of Biomedical Informatics*, *100*, 103311. https://doi.org/10.1016/j.jbi.2019.103311

Miller, K. D., Nogueira, L., Mariotto, A. B., Rowland, J. H., Yabroff, K. R., Alfano, C. M., et al. (2019). Cancer treatment and survivorship statistics, 2019. *CA: A Cancer Journal for Clinicians*, *69*(5), 363–385. https://doi.org/10.3322/caac.21565

Morgan, K. H., Barroso, C. S., Bateman, S., Dixson, M., & Brown, K. C. (2020). Patients' Experiences of Interprofessional Collaborative Practice in Primary Care: A Scoping Review of the Literature. *Journal of Patient Experience*, *7*(6), 1466–1475. https://doi.org/10.1177/2374373520925725

Morgan, P. A., Murray, S., Moffatt, C. J., & Honnor, A. (2012). The challenges of managing complex lymphoedema/chronic oedema in the UK and Canada. *International Wound*

Journal, *9*(1), 54–69. https://doi.org/10.1111/j.1742-481X.2011.00845.x

Morley, L., & Cashell, A. (2017). Collaboration in Health Care. *Journal of Medical Imaging and Radiation Sciences*, *48*(2), 207–216. https://doi.org/10.1016/j.jmir.2017.02.071

Murata, T., & Moriyasu, S. (2007). Link Prediction of Social Networks Based on Weighted Proximity Measures. https://doi.org/10.1109/WI.2007.52

Newman, M. E. J. (2001). Clustering and preferential attachment in growing networks. *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, *64*(2), 4. https://doi.org/10.1103/PhysRevE.64.025102

Nie, B., & Sun, S. (2017). Using text mining techniques to identify research trends: A case study of design research. *Applied Sciences (Switzerland)*, *7*(4). https://doi.org/10.3390/app7040401

Parimi, R., & Caragea, D. (2011). Predicting friendship links in social networks using a topic modeling approach. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *6635 LNAI*(PART 2), 75–86. https://doi.org/10.1007/978-3-642-20847-8_7

Parish, A. J., Boyack, K. W., & Ioannidis, J. P. A. (2018). Dynamics of co-authorship and productivity across different fields of scientific research. *PLoS ONE*, *13*(1), 1–12. https://doi.org/10.1371/journal.pone.0189742

Pavlov, M., & Ichise, R. (2007). Finding experts by link prediction in co-authorship networks. *CEUR Workshop Proceedings*, *290*, 42–55.

Pedregosa FABIANPEDREGOSA, F., Michel, V., Grisel OLIVIERGRISEL, O., Blondel, M., Prettenhofer, P., Weiss, R., et al. (2011). Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA,

VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot. *Journal of Machine Learning Research*, *12*, 2825–2830. http://scikit-learn.sourceforge.net. Accessed 5 September 2021

Perez-Iratxeta, C., Andrade-Navarro, M. A., & Wren, J. D. (2007). Evolving research trends in bioinformatics. *Briefings in Bioinformatics*, *8*(2), 88–95. https://doi.org/10.1093/bib/bbl035

Ponomariov, B., & Boardman, C. (2016). What is co-authorship? *Scientometrics*, *109*(3), 1939–1963. https://doi.org/10.1007/s11192-016-2127-7

Powers, D. M. W. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation, (January 2008). http://arxiv.org/abs/2010.16061

Pramanik, M. I., Lau, R. Y. K., Demirkan, H., & Azad, M. A. K. (2017). Smart health: Big data enabled health paradigm within smart cities. *Expert Systems with Applications*, *87*, 370–383. https://doi.org/10.1016/j.eswa.2017.06.027

Reisenbichler, M., & Reutterer, T. (2019). Topic modeling in marketing: recent advances and research opportunities. *Journal of Business Economics*, *89*(3), 327–356. https://doi.org/10.1007/s11573-018-0915-7

Renshaw, M. (2007, April). Lymphorrhoea: “leaky legs” are not just the nurse’s problem. *British journal of community nursing*. Br J Community Nurs. https://doi.org/10.12968/bjcn.2007.12.sup2.23261

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. *WSDM 2015 - Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, 399–408. https://doi.org/10.1145/2684822.2685324

S. Mosallaie, M. Rad, A. Schiffauerova, A. E. (2021). Characterization of the Application of Artificial Intelligence in Cancer Diagnosis and Treatment using Dynamic Topic Modeling. *COLLNET Journal of Scientometrics and Information Management*, 1–19.

Sadoughi, F., Kazemy, Z., Hamedan, F., Owji, L., Rahmanikatigari, M., & Azadboni, T. T. (2018). Artificial intelligence methods for the diagnosis of breast cancer by image processing: A review. *Breast Cancer: Targets and Therapy*, *10*, 219–230. https://doi.org/10.2147/BCTT.S175311

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*, *10*(3), 1–21. https://doi.org/10.1371/journal.pone.0118432

Sayed, S., Moloo, Z., Bird, P., Wasike, R., Njoroge, W., Karanu, J., et al. (2013). Breast cancer diagnosis in a resource poor environment through a collaborative multidisciplinary approach: The Kenyan experience. *Journal of Clinical Pathology*, *66*(4), 307–311. https://doi.org/10.1136/jclinpath-2012-201404

Sellwood, M. A., Ahmed, M., Segler, M. H. S., & Brown, N. (2018). Artificial intelligence in drug discovery. *Future Medicinal Chemistry*, *10*(17), 2025–2028. https://doi.org/10.4155/fmc-2018-0212

Shahbazi, Z., Byun, Y., & Byun, Y.-C. (2020). Product Recommendation Based on Content-based Filtering Using XGBoost Classifier. *International Journal of Advanced Science and Technology*, *29*(04), 6979–6988. https://www.researchgate.net/publication/342864588

Shen, T. L., & Fu, X. L. (2018). Application and prospect of artificial intelligence in cancer diagnosis and treatment. *Zhonghua zhong liu za zhi [Chinese journal of oncology]*, *40*(12), 881–884. https://doi.org/10.3760/cma.j.issn.0253-3766.2018.12.001

Siegel, R. L., Miller, K. D., Goding Sauer, A., Fedewa, S. A., Butterly, L. F., Anderson, J. C., et al. (2020). Colorectal cancer statistics, 2020. *CA: A Cancer Journal for Clinicians*, *70*(3), 145–164. https://doi.org/10.3322/caac.21601

Skorkovská, L. (2012). Application of lemmatization and summarization methods in topic identification module for large scale language modeling data filtering. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *7499 LNAI*(September 2012), 191–198. https://doi.org/10.1007/978-3-642-32790-2_23

Smith, C., & Sotala, K. (2011). *Knowledge , networks and nations Global scientific collaboration in the 21st century*. *Networks* (Vol. 03/11). http://royalsociety.org/uploadedFiles/Royal_Society_Content/Influencing_Policy/Reports/2011-03-28-Knowledge-networks-nations.pdf

Sonnenwald, D. H. (2007a). Scientific collaboration. *Annual Review of Information Science and Technology*, *41*, 643–681. https://doi.org/10.1002/aris.2007.1440410121

Sonnenwald, D. H. (2007b). Scientific collaboration. *Annual Review of Information Science and Technology*, *41*, 643–681. https://doi.org/10.1002/aris.2007.1440410121

Spelt, L., Andersson, B., Nilsson, J., & Andersson, R. (2012). Prognostic models for outcome following liver resection for colorectal cancer metastases: A systematic review. *European Journal of Surgical Oncology*, *38*(1), 16–24. https://doi.org/10.1016/j.ejso.2011.10.013

Subramanyam, K. (1983). Bibliometric studies of research collaboration: A review. *Journal of Information Science*, *6*(1), 33–38. https://doi.org/10.1177/016555158300600105

Sun, L., & Yin, Y. (2017). Discovering themes and trends in transportation research using topic modeling. *Transportation Research Part C: Emerging Technologies*, *77*, 49–66. https://doi.org/10.1016/j.trc.2017.01.013

Sutton, R. T., Pincock, D., Baumgart, D. C., Sadowski, D. C., Fedorak, R. N., & Kroeker, K. I. (2020). An overview of clinical decision support systems: benefits, risks, and strategies for

success. *npj Digital Medicine*, *3*(1), 1–10. https://doi.org/10.1038/s41746-020-0221-y

Thakur, D., Wang, J., & Cozzens, S. (2011). What does international co-authorship measure? *2011 Atlanta Conference on Science and Innovation Policy: Building Capacity for Scientific Innovation and Outcomes, ACSIP 2011, Proceedings*. https://doi.org/10.1109/ACSIP.2011.6064489

Tran, B., Vu, G., Ha, G., Vuong, Q.-H., Ho, M.-T., Vuong, T.-T., et al. (2019). Global Evolution of Research in Artificial Intelligence in Health and Medicine: A Bibliometric Study. *Journal of Clinical Medicine*, *8*(3), 360. https://doi.org/10.3390/jcm8030360

Tran, B. X., Latkin, C. A., Sharafeldin, N., Nguyen, K., Vu, G. T., Tam, W. W. S., et al. (2019). Characterizing Artificial Intelligence Applications in Cancer Research: A Latent Dirichlet Allocation Analysis. *JMIR Medical Informatics*, *7*(4), e14401. https://doi.org/10.2196/14401

Vayansky, I., & Kumar, S. A. P. (2020). A review of topic modeling methods. *Information Systems*, *94*, 101582. https://doi.org/10.1016/j.is.2020.101582

Vert, J. P. (2020). Artificial intelligence and cancer genomics. In *Healthcare and Artificial Intelligence* (pp. 165–174). Springer International Publishing. https://doi.org/10.1007/978-3-030-32161-1_22

Viator, J. A., & Pestorius, F. M. (2001). Investigating trends in acoustics research from 1970–1999. *The Journal of the Acoustical Society of America*, *109*(5), 1779–1783. https://doi.org/10.1121/1.1366711

Wang, K., Shen, Z., Huang, C., Wu, C.-H., Dong, Y., & Kanakia, A. (2020). Microsoft Academic Graph: When experts are not enough. *Quantitative Science Studies*, *1*(1), 396–413. https://doi.org/10.1162/qss_a_00021

Wardle, J., Robb, K., Vernon, S., & Waller, J. (2015). Screening for prevention and early diagnosis of cancer. *American Psychologist*, *70*(2), 119–133. https://doi.org/10.1037/a0037357

Workman, P., Antolin, A. A., & Al-Lazikani, B. (2019). Transforming cancer drug discovery with Big Data and AI. *Expert Opinion on Drug Discovery*, *14*(11), 1089–1095. https://doi.org/10.1080/17460441.2019.1637414

World Health Organization. (2018). *World Health Organization*. https://www.who.int/news-room/fact-sheets/detail/cancer. Accessed 21 March 2021

Xuan Tran, B., Thu Vu, G., Hai Ha, G., Vuong, Q.-H., Ho, M.-T., Vuong, T.-T., et al. (2019a). Clinical Medicine Global Evolution of Research in Artificial Intelligence in Health and Medicine: A Bibliometric Study. *J. Clin. Med*, *8*, 360. https://doi.org/10.3390/jcm8030360

Xuan Tran, B., Thu Vu, G., Hai Ha, G., Vuong, Q.-H., Ho, M.-T., Vuong, T.-T., et al. (2019b). Clinical Medicine Global Evolution of Research in Artificial Intelligence in Health and Medicine: A Bibliometric Study. *J. Clin. Med*, *8*, 360. https://doi.org/10.3390/jcm8030360

Yang, H. L., Chang, T. W., & Choi, Y. (2018). Exploring the research trend of smart factory with topic modeling. *Sustainability (Switzerland)*, *10*(8). https://doi.org/10.3390/su10082779

Yu, K. H., Beam, A. L., & Kohane, I. S. (2018). Artificial intelligence in healthcare. *Nature Biomedical Engineering*, *2*(10), 719–731. https://doi.org/10.1038/s41551-018-0305-z

Yu, Q., Long, C., Lv, Y., Shao, H., & He, P. (2014). Predicting Co-Author Relationship in Medical Co-Authorship Networks. *PLoS ONE*, *9*(7), 101214. https://doi.org/10.1371/journal.pone.0101214.t001

Zhu, Y., Wang, Q. C., Xu, M. D., Zhang, Z., Cheng, J., Zhong, Y. S., et al. (2019). Application

of convolutional neural network in the diagnosis of the invasion depth of gastric cancer based on conventional endoscopy. *Gastrointestinal Endoscopy*, *89*(4), 806-815.e1. https://doi.org/10.1016/j.gie.2018.11.011

Zou, L., Yu, S., Meng, T., Zhang, Z., Liang, X., & Xie, Y. (2019). A Technical Review of Convolutional Neural Network-Based Mammographic Breast Cancer Diagnosis. *Computational and Mathematical Methods in Medicine*, *2019*(Dm). https://doi.org/10.1155/2019/6509357