What Can We Learn from The Authors' Contribution Section?

Itai Assraf[,1].

[a]Department of Data Engineering, Ben Gurion University, Israel.

Abstract**—***This study investigates author contributions in collaborative research, focusing on the relationship between authors' positions and their tasks, recurring task patterns, and the correlation between task types and author positions. Using data from OnePlus and Nature databases, we standardized author contributions and analyzed patterns using machine learning. Key findings include the ubiquitous nature of the writing task across all author positions, the diverse contribution patterns of first and last authors, and the effectiveness of our classification model in predicting task types based on author position and contribution metrics. These findings offer insights into authorship dynamics and have implications for research assessment and collaboration practices.*

1. Introduction

One of the current prerequisites for drafting and publishing an article involves composing a paragraph delineating the individual contributions of each author to the collaborative efforts, the article itself, and the various stages of the project. This practice prompts several inquiries regarding the authorship order and the nature of contributions, particularly in larger groups. We aim to address several key questions: Firstly, do authors who contribute later undertake fewer tasks? Is there a discernible pattern where the first and last authors tend to handle the most and some tasks respectively? Secondly, as the group size increases, do we observe a more equitable distribution of tasks among members? Or is there a skewed allocation, with first authors shouldering more responsibilities than last authors? Are these trends consistent across different group sizes? Thirdly, are there recurring patterns of collaborative tasks among authors? If so, how prevalent are these patterns, and what is their typical size? Lastly, is there a correlation between the type of task and the author's position in the authorship order?

To tackle these questions comprehensively, we will employ a multifaceted approach that encompasses various stages of data analysis and utilization of advanced machine learning techniques.

2. Related Work

Indeed, the field of general information extraction from scientific literature has witnessed significant advancements over the years, with various approaches and tools emerging to facilitate this process. Notable examples include CERMINE [1], GROBID [2], PDFX [3] and ParsCit [4]. These tools have proven invaluable in extracting diverse types of information from scientific papers, ranging from bibliographic data to structured metadata.

However, despite the abundance of tools available, none of the existing systems are specifically designed to extract information related to the contributions of individual authors directly from the content of the paper. This represents a notable gap in the current landscape of information extraction tools, as understanding the contributions of authors is essential for assessing their scholarly impact and assigning appropriate credit.

One of the challenges in this regard is the lack of standardized criteria for authorship across the scientific community. While authorship guidelines exist, such as those published by the International Committee of Medical Journal Editors, adherence to these guidelines remains voluntary, and there is no universal consensus on what constitutes significant author contributions. Consequently, inconsistencies in authorship practices persist across different journals and research fields.

In addition to the challenges posed by inconsistent authorship criteria, another significant hurdle in assessing author contributions lies in the variability of how authors' initials are represented. Authors may use different combinations or formats for their initials, such as including middle initials or using different punctuation conventions, making it challenging to automatically match initials with the full names of authors and infer their contributions based on their location within the author list.

To address these challenges, some databases, such as OnePlus and Nature, have established standardized contribution types that authors can specify when submitting articles[1].

These contribution categories encompass a wide range of tasks, including writing, methodology, investigation, conceptualization, data analysis, and project administration, among others. By categorizing contributions in this manner, databases aim to provide clarity and transparency regarding each author's role in the research process, thereby facilitating more accurate assessments of individual contributions.

Despite these efforts, there remains a dearth of research focusing specifically on the relationship between authors' locations within the author list and their contributions. However, a few notable studies have attempted to tackle this issue using different approaches. One approach [5] involves developing rules to assign tasks to predefined categories, thereby reducing the complexity of categorizing diverse types of contributions. This method enables the creation of automated tools that can identify patterns in author contributions and associate them with specific task categories, ultimately streamlining the assessment process.

Another study [6] aimed to explore the relationship between authors' positions and their contributions, albeit with a limited dataset comprising approximately 576 articles. While the findings of this study provide valuable insights, the small size of the dataset raises questions about the generalizability of the results to a broader context. Therefore, it remains challenging to ascertain whether these findings accurately reflect trends in author contributions across a wider range of publications. Further research with larger and more diverse datasets is needed to elucidate the complex dynamics underlying authorship and contribution patterns in scientific literature.

3. Methodology

3.1 Dataset

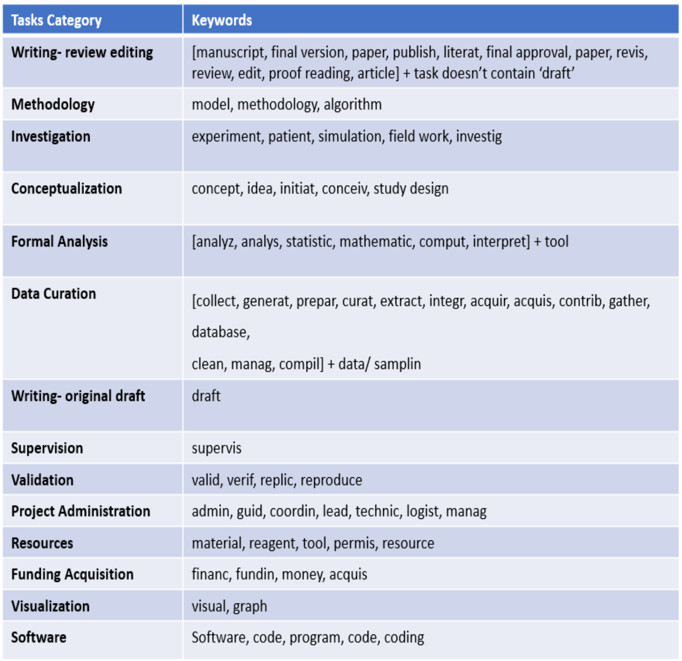
In this article, our focus was on extracting articles from OnePlus and Nature databases to investigate author contributions. OnePlus provides a corpus of articles represented in XML format, from which we extracted key features including the article's publication year, author names, article title, and author contributions. Similarly, we obtained articles from Nature, extracting information from HTML files to acquire the same set of features. Given the large volume of

https://journals.plos.org/plosone/s/authorship [1] data we were dealing with, we initially saved the extracted data in JSON files, resulting in a dataset comprising 353 thousand documents.

3.2 Pre-Processing

Upon completing the data extraction process, our next step involved data cleaning and standardization to ensure uniform representation.

One major aspect of this was standardizing the representation of authors' contributions, as there is no specific and uniform requirement for describing authors' tasks in scholarly articles. This led to a myriad of different possibilities, making it challenging to analyze and categorize author contributions effectively. To address this challenge, we identified 14 distinct "task categories" based on common types of contributions observed in scholarly literature. These categories include 'writing - review editing', 'methodology', 'investigation', 'conceptualization', 'formal analysis', 'data curation', 'writing - original draft', 'supervision', 'validation', 'project administration', 'resources', 'funding acquisition', 'visualization', and 'software'. Leveraging the descriptions provided for each task category, we devised rules and keyword-based algorithms to automatically assign each author's contribution to the most appropriate category. This approach enabled us to streamline the categorization process and ensure consistency in how author contributions were represented and analyzed, as illustrated in Figure 1.



Tasks Category	Keywords
Writing- review editing	[manuscript, final version, paper, publish, literat, final approval, paper, revis, review, edit, proof reading, article] + task doesn't contain 'draft'
Methodology	model, methodology, algorithm
Investigation	experiment, patient, simulation, field work, investig
Conceptualization	concept, idea, initiat, conceiv, study design
Formal Analysis	[analyz, analys, statistic, mathematic, comput, interpret] + tool
Data Curation	[collect, generat, prepar, curat, extract, integr, acquir, acquis, contrib, gather, database, clean, manag, compil] + data/ samplin
Writing- original draft	draft
Supervision	supervis
Validation	valid, verif, replic, reproduce
Project Administration	admin, guid, coordin, lead, technic, logist, manag
Resources	material, reagent, tool, permis, resource
Funding Acquisition	financ, fundin, money, acquis
Visualization	visual, graph
Software	Software, code, program, code, coding

Figure1: The keywords for each category of task are established, and if any of these keywords appear, we will assign the task to the corresponding category. Additionally, we have implemented two more rules: firstly, for "writing review editing," if any of the keywords appear in the researcher's contribution and the word "draft" does not appear, we will assign the task to this category. Secondly, for "data curation," if any of the keywords appear in the researcher's contribution, in addition to them, the task must contain "data" or "sampling" to be categorized as such.

Another significant challenge we encountered during the data processing phase was the inconsistency in the representation of authors' names, especially the use of initials. In the author contribution paragraph, authors' initials are typically used instead of their full names, unlike in the list of authors' names where full names are usually provided. This discrepancy was an obstacle in matching the names of the authors with their initials, something essential for performing the data analysis necessary to address the questions raised above.

The variation in initials added to the complexity, as initials could be represented in many combinations. For example, authors' initials may consist of the first letter of their first name followed by the second letter of their last name, or vice versa. Alternatively, authors' initials can contain their full first name together with the first letter of their last name, or vice versa. Additionally, authors may choose to use their full name without initials altogether. This complexity is further compounded in cases where authors have multiple names. To address this challenge, we systematically tested all possible combinations of the initials and matched them to their corresponding full names. By standardizing the representation of author names, we were able to streamline subsequent analyses and enable meaningful insights into author contributions.

3.3 Authors' Contribution Analysis

To commence our investigation, we embarked on identifying the most and least common types of tasks undertaken by authors. As illustrated in Figure 2, "writing review editing" emerged as the predominant type of author contribution to an article, while "software development" ranked as the least common. Our initial assumption posited that during the article-writing process, each member of the group would assume a task, leveraging their specialized knowledge. Consequently, we anticipated that the majority of participants in the group would contribute to writing the article.

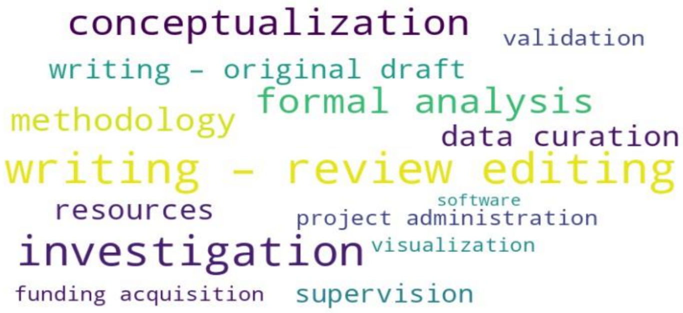


Figure 2: Word Clouds graph for Authors' Contribution

Subsequently, we aimed to uncover intriguing patterns among task types, focusing on the types of tasks that authors frequently undertake together. To achieve this, we employed an Apriori algorithm, a methodology renowned for its ability to mine frequent item sets and extract association rules from relational databases. We set a threshold to identify patterns with a frequency exceeding 0.2, thereby filtering out less frequent associations.

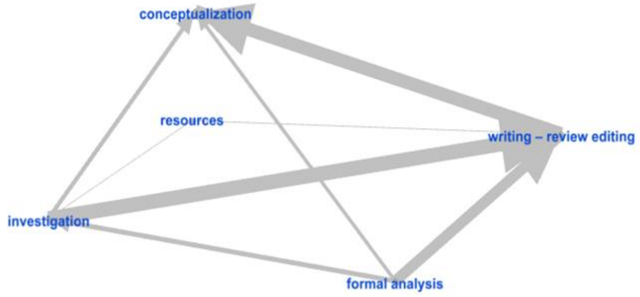


Figure 3: A network graph illustrating the most prevalent patterns is depicted. Pairs associated with a specific pattern are connected by an edge. The size of the edge corresponds to the frequency of occurrence, with larger edges indicating higher frequencies.

As we can see in figure 3, the results yielded eight distinct patterns of task pairs that met our predefined frequency criterion. Notably, the task of "writing review editing" exhibited a strong association with "formal analysis," "conceptualization," "resources," and "investigation." Furthermore, we observed that "formal analysis" and "conceptualization," as well as "investigation" and "formal analysis," are very common. This finding underscores a strong correlation between these task types in the research process. We conjecture that the frequent appearance of the writing task in five of the eight most common patterns is attributed to our earlier assumption. Given its universal relevance to all group members, the writing task naturally accompanies various other tasks in the research paper.

3.4 Authors' Position in Authors list.

Upon investigating the potential correlation between an author's position in the list of authors and the number of tasks they undertake, our analysis revealed intriguing insights. As illustrated in figure 4, authors positioned later in the list tend to engage in fewer tasks compared to those listed earlier. This observation underscores the significance of an author's position, suggesting that individuals occupying prominent positions at the beginning of the list contribute significantly to the group's endeavors. The placement of an author in the early positions signifies their substantial importance within the group and their substantial contributions to its collective output.

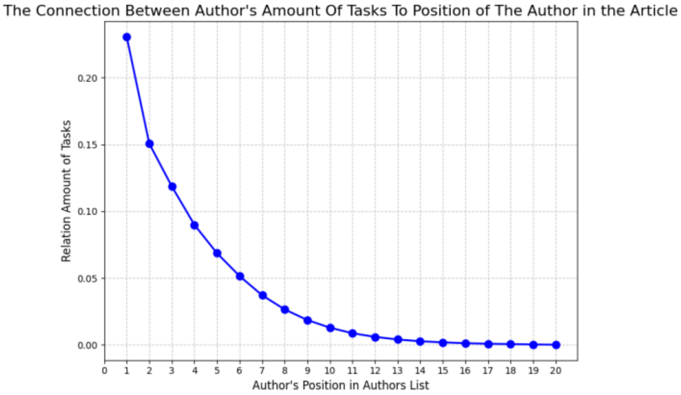


Figure 4: The connection between authors' position in author list to the number of tasks the author performed.

Expanding our investigation further, we sought to explore the distribution of author contributions across varying positions within the author list, taking into account the total number of authors involved in a publication. This inquiry aimed to elucidate whether there exists a discernible pattern in the contribution percentages of authors based on their positional ranking relative to total number of authors listed.

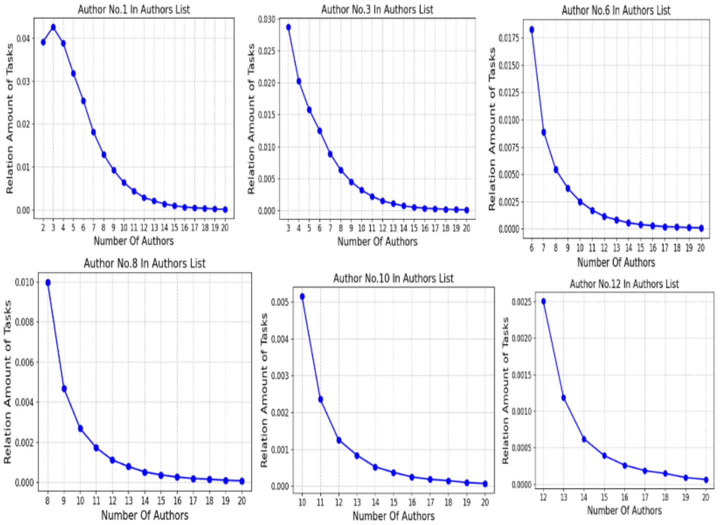


Figure 5: The connection between authors' position in author list to the number of tasks the author performed across varying author counts.

For instance, we posited the question: Would an author occupying the third position demonstrate a lower contribution percentage when part of a larger group with ten authors compared to a scenario where only three authors are involved? Our hypothesis stemmed from the intuitive notion that as the size of the author group increases, the individual contribution of each member may diminish proportionally. Conversely, in smaller author groups, where each member's contribution carries relatively more weight, we anticipated observing a higher percentage of contributions from authors in similar positional ranks.

As depicted in Figure 5, a discernible trend emerges when examining the relationship between author involvement and group size. Notably, there is a noticeable decrease in author contributions as the number of authors involved in a publication increases. This observation raises intriguing questions regarding the distribution of responsibilities and the dynamics of collaborative research efforts across varying group sizes.
For instance, consider the scenario where an author occupies the first position in the list of contributors. In smaller groups comprising a limited number of authors, this individual typically assumes a more substantial share of responsibilities, contributing significantly to the group's collective output. However, as the size of the author group expands, we observe a relative decrease in the contribution percentage of authors occupying similar positional ranks.
This phenomenon suggests that in larger author groups, the individual contributions of each member may diminish proportionally, reflecting a more distributed and diversified allocation of tasks and responsibilities. Authors occupying prominent positions in the author list may find their relative contributions attenuated as the group size increases, potentially due to a broader distribution of tasks among a larger pool of contributors.

3.5 Connection Between Authors' Position to Task Type

The subsequent inquiry we sought to address pertains to the potential correlation between the author's position within the list and the type of contribution made. Are certain contribution types more prevalent for authors occupying specific positions? Irrespective of the author's position within the list, the task of "writing the article" emerged as the most frequent contribution type. Additionally, a consistent pattern emerges across various author positions, wherein certain tasks such as conceptualization, formal analysis, and investigation exhibit notable recurrence in terms of prevalence percentages.

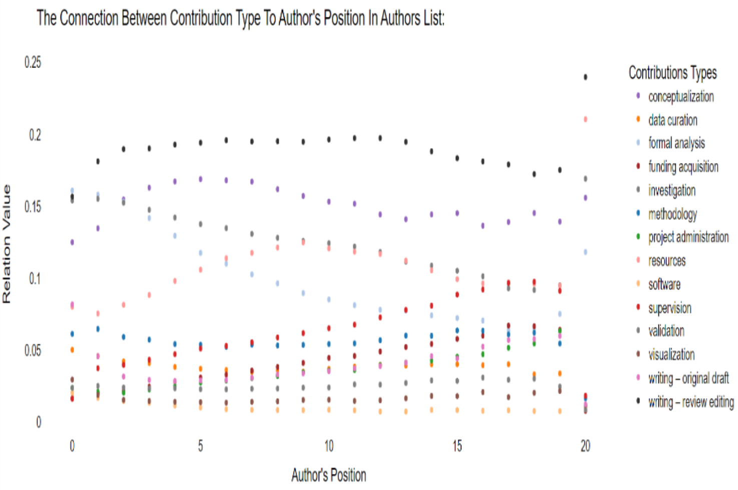
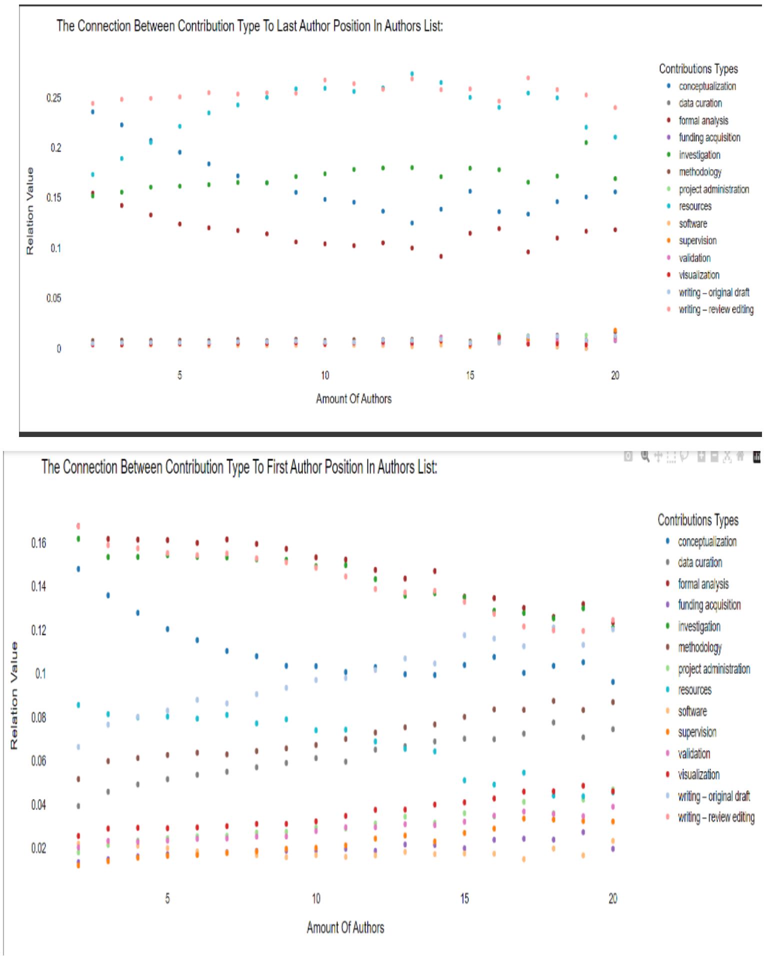


Figure 6: The connection between task type to authors' position in author list.

This suggests a degree of uniformity or consensus in the distribution of tasks among authors, regardless of their placement within the authorship hierarchy.
These findings confirm our initial assumption regarding the ubiquitous nature of the writing task across all authors. Assembling an article entails synthesizing the collective findings and contributions of each author, thereby necessitating the integration of diverse insights and perspectives. Consequently, the task of writing the article emerges as the most prevalent among all potential configurations of authorship. This underscores the fundamental role of the writing task in consolidating and presenting the research outcomes comprehensively.

In this aspect, we aimed to delve deeper into the nuanced dynamics of authorship responsibilities, particularly focusing on authors occupying the first and last positions. Our curiosity led us to explore whether specific tasks exhibit repetition or if there exists diversity and variation in the types of contributions across different author positions.
Upon scrutinizing the contribution patterns of the last author, a discernible trend emerges: there appears to be a limited range of tasks performed by the final author, with only a handful of tasks showing notable occurrence frequencies. Conversely, when analyzing the contribution profile of the first author, a contrasting picture unfolds. Here, we observe a greater diversity and uniformity in the distribution of tasks. This discrepancy can be attributed to the pivotal role typically assumed by the first author, who often shoulders a larger share of responsibilities compared to other authors. Consequently, the first author tends to undertake a broader spectrum of tasks, reflecting their central role in the research endeavor.



Figures 7+8: The connection between contribution type to last and first authors in authors list. There are 5 tasks that are most common for the last author: resources, writing review editing, investigation, formal analysis and conceptualization.

4. Model

After uncovering intriguing insights from our analysis, we embarked on the development of a classification model aimed at predicting whether an author would undertake specific types of tasks. This endeavor entailed predicting 14 distinct classifications, each corresponding to a different task type. Leveraging various features extracted from the dataset, including the author's position in the list, the number of authors, the publication year, and the author's contribution metrics relative to the entire group, we sought to construct a robust predictive framework.
In our approach, we evaluated two prominent machine learning algorithms, namely XGBoost and logistic regression, to identify the most effective model for our classification task. The evaluation metrics revealed that the XGBoost algorithm outperformed logistic regression, yielding an average accuracy of 0.84 across all 14 classification tasks, whereas logistic regression achieved a slightly lower average accuracy of 0.81.

Notably, certain tasks such as writing the article, conceptualization, and formal analysis exhibited lower accuracy percentages in classification. This trend stemmed from the ubiquitous nature of these tasks across various author positions, making them challenging to predict solely based on the provided features. Conversely, tasks such as software development, visualization, and validation demonstrated higher accuracy percentages in classification. This can be attributed to their prevalence among specific author positions, with minimal overlap across different author locations. Consequently, the predictive model was more adept at discerning patterns and accurately classifying authors based on these specialized tasks.

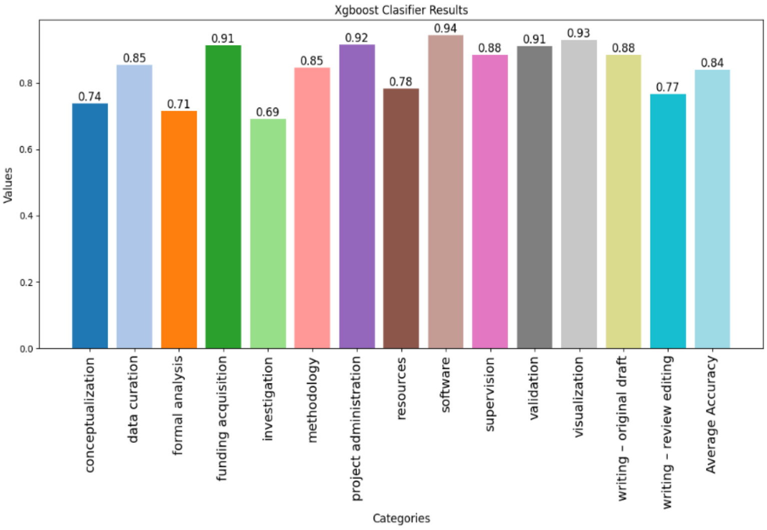


Figure 9: XGboost classifier model results for each contribution type task. The model with the best performance.

5. Future Work

While this study provides valuable insights into author contributions and their relationship with author positions, there are several avenues for future research to explore and enhance our understanding of collaborative research dynamics:

First, fine-grained task classification. One potential area for future investigation involves refining the task classification framework to capture more nuanced variations in author contributions. Currently, our study categorizes tasks into broad categories; however, developing a more granular classification system could offer deeper insights into the specific roles and responsibilities undertaken by authors.

Second, cross-disciplinary comparison. Investigating author contributions across different academic disciplines could provide valuable insights into disciplinary variations in collaboration practices. Comparing contribution patterns in fields with distinct research cultures and norms could shed light on factors influencing authorship dynamics, such as disciplinary conventions, team size preferences, and credit attribution practices.

Third, validation and generalization. Validating the findings of this study using independent datasets and diverse research contexts is essential to ensure the generalizability of our results. Future research could replicate our analysis using datasets from different sources or specific research domains to assess the robustness of observed patterns and trends.

6. Acknowledgment

The authors gratefully acknowledge the assistance provided by ChatGPT in generating and refining the content presented in this paper. For a more comprehensive analysis of the code underlying this study, please refer to this GitHub repository: https://github.com/itaiassraf/Big-Data-Mining-Project.

7. References

[1] D. Tkaczyk, P. Szostek, M. Fedoryszak, P. Dendek and L. Bolikowski, "CERMINE: automatic extraction of structured metadata from scientific literature," International Journal on Document Analysis and Recognition, vol. 18, no. 4, pp. 317-335, 2015.

[2] P. Lopez, "GROBID: combining automatic bibliographic data recognition and term extraction for scholarship publications," Research and Advanced Technology for Digital Libraries, 2009

[3] A. Constantin, S. Pettifer and A. Voronkov, "PDFX: fully-automated pdf-to-xml conversion of scientific literature," DocEng, pp. 177-180, 2013

[4] I. Councill, C. Giles and M.-Y. Kan, "ParsCit: an open-source CRF reference string parsing package," in International Conference on Language Resources and Evaluation, 2008.

[5] Tkaczyk, Dominika, Andrew Collins, and Joeran Beel. “NaïveRole: Author-Contribution Extraction from Biomedical Publications.” In 27th AIAI Irish Conference on Artificial Intelligence and Cognitive Science, 2019.

[6] Rieko Ueda PhD, Yuji Nishizaki MD, Yasuhiro Homma MD, Patrick Devos MS, Shoji Sanada M. The relationship between contributions of authors and author order, 2021