

A Bibliometric and Network Analysis of Statistical Distributions in Machine Learning

Manju Jose
Department of Statistics
Alphonsa College, India

Christophe Chesneau
Department of Mathematics
University of Caen, France

Lishamol Tomy
Department of Statistics
Deva Matha College, India

Anjana S
Department of Economics
Alphonsa College, India

ABSTRACT

Statistical distributions constitute a fundamental component of machine learning by providing the probabilistic framework required for modelling uncertainty, variability, and data-generating processes. They underpin a wide range of methodologies, including Bayesian inference, probabilistic graphical models, generative learning and uncertainty-aware artificial intelligence. Despite their widespread application, the evolution and intellectual structure of research on statistical distributions in machine learning have not been systematically examined. This study presents a bibliometric and network analysis of 840 Scopus-indexed publications published between 1997 and 2025. Bibliographic data were analysed using Biblioshiny and VOS viewer to evaluate publication trends, influential sources, author productivity, collaboration patterns and thematic developments. Descriptive indicators were complemented by co-authorship, co-citation, keyword co-occurrence and thematic analyses to explore the social and intellectual structure of the field. The findings reveal sustained growth in research output, increasing international collaboration and the dominance of a core set of journals and authors. Network analyses identify probabilistic modelling, Bayesian learning, uncertainty quantification and generative approaches as the principal intellectual foundations of the field. The results further indicate a gradual shift from classical statistical learning frameworks toward uncertainty-aware and explainable machine learning paradigms. By mapping the evolution, knowledge structure and emerging directions of this research domain, the study provides a comprehensive reference for future investigations at the intersection of statistical distribution theory and machine learning.

Keywords: Statistical distributions, Machine learning, Bibliometric analysis, Research trends

1. Introduction

Statistical distributions play a pivotal role in machine learning by providing mathematical models for representing uncertainty, variability and probabilistic relationships within data. Modern machine learning systems increasingly operate in complex and uncertain environments where reliable prediction requires not only high accuracy but also robust quantification of uncertainty. Probability distributions such as Gaussian, Poisson, Gamma, Dirichlet and mixture distributions form the theoretical basis for parameter estimation, probabilistic inference, latent variable modelling and uncertainty quantification. Consequently, statistical distributions have become indispensable components of contemporary machine learning frameworks (Bishop, 2006; Jordan et al., 1999).

The integration of statistical distribution theory with machine learning has accelerated substantially over the past three decades. Distribution-based approaches underpin numerous methodologies, including Bayesian learning, probabilistic graphical models, support vector machines, ensemble learning methods, deep generative models and uncertainty-aware neural networks. Advances in computational power, data availability and artificial intelligence have

- Received June 20, 2026, in final form August 2026.
- Christophe Chesneau (corresponding author) is affiliated with Department of Mathematics, LMNO, University of Caen, Caen, France
Email: christophe.chesneau@unicaen.fr

further expanded the application of probabilistic methods across domains such as healthcare, finance, environmental science, computer vision and natural language processing. As a result, the literature concerning statistical distributions in machine learning has grown rapidly and become increasingly interdisciplinary (Goodfellow et al., 2016; Kingma & Welling, 2014). Classic work with random forests (Breiman, 2001), support vector machines (Cortes & Vapnik, 1995) and latent variable models (Blei et al., 2003) have shown that using probability improves a model's flexibility and interpretability. Recent advances, like optimal transport theory and generative adversarial networks, demonstrate that having good distribution methods is becoming more vital in machine learning (Arjovsky et al., 2017; Peyré & Cuturi, 2019). Because the study of distributions in machine learning has grown quickly, there's a lot of scattered info available. Research is spread across different fields and journals, which makes it hard to grasp the area's structure and main ideas. Traditional reviews can be helpful, but they can't easily track large trends, citation activity and connections between researchers. Bibliometric analysis is a useful way to measure science, how important it is, and how it has changed over time (Aria & Cuccurullo, 2017).

Bibliometric techniques allow researchers to map how an area grows by looking at things such as publication rates, citations, collaborations and keyword relationships. These reviews give information on how knowledge is created, who the key players are and what new topics are appearing. Although statistical distributions are central to machine learning, there is little comprehensive bibliometric research in this area. Accordingly, this study conducts a comprehensive bibliometric and network analysis of research on statistical distributions used in machine learning. Specifically, the study aims to: (i) examine publication growth and citation patterns, (ii) identify influential journals, authors, and countries, (iii) investigate collaboration and knowledge structures through network analysis and (iv) explore the thematic evolution of the field. By integrating descriptive and network-based bibliometric techniques, the study provides a systematic overview of the intellectual development of research on statistical distributions in machine learning.

2. Methodology

This study employs a bibliometric and science-mapping approach to analyse the evolution and intellectual structure of research on statistical distributions used in machine learning. Bibliometric methods provide quantitative tools for evaluating scientific production, citation impact, collaboration patterns, and thematic development within a research domain (Donthu et al., 2021).

Bibliographic records were retrieved from the Scopus database, selected for its extensive coverage of peer-reviewed journals, conference proceedings, and scholarly publications in computer science, statistics and artificial intelligence. The search was conducted using the query “Statistical distributions used in machine learning”, applied to titles, abstracts, and author keywords. The analysis was restricted to publications published between 1997 and 2025. Journal articles, conference papers and book chapters were included, while duplicate records, incomplete entries, editorials, notes and documents lacking bibliographic information were excluded. Following data cleaning and screening, a final dataset comprising 840 publications was retained for analysis.

The bibliographic data were analysed using Biblioshiny, the web interface of the Bibliometrix package in R and VOS viewer, a specialised software tool for network visualisation (Aria & Cuccurullo, 2017). Descriptive indicators such as annual scientific production, citation impact, author productivity, and source productivity were computed to evaluate research performance. Network analyses included co-authorship, co-citation, country collaboration and keyword co-occurrence mapping to reveal the social, intellectual and

conceptual structures of the field. Thematic mapping and historiographic analyses were further employed to identify major research themes and trace their evolution over time. The combined use of Biblioshiny and VOS viewer provided a comprehensive framework for examining research trends, influential contributions, and emerging directions in studies related to statistical distributions in machine learning.

3. Bibliometric Analysis

Bibliometric analysis gives a way to study the structure, output and changes in research using numbers and systems. By looking at publications, citations, authors and keywords, it shows how knowledge is made, spread and grows in a field. For machine learning's statistical distributions, these methods help find key people, important journals, new topics and teamwork. This gives a full view of the intellectual and social makeup of the literature.

This work uses these techniques to check publication growth, citation habits, author work, source influence and topic changes. Both simple measures and network measures are used to find patterns not easily seen in normal literature reviews. This leads to a better grasp of how statistical distribution research has changed in machine learning and how ideas, authors and groups are linked over time.

3.1 Descriptive Analysis of Bibliometric Data

This descriptive analysis offers a quantitative look at how research on statistical distributions in machine learning is structured and how it has changed. It checks publication rates, citation patterns, key journals, top authors and keyword use over time. These metrics aid in understanding field growth, research output distribution and the main people who have influenced its progress. Examining these measures makes it possible to spot major publication trends, important sources and the research area's general state.

3.1.1 Overview of the Data

Table 1: Overview of the Data

Main information about data	Result
Time span	1997–2025
Sources (Journals, Books, etc.)	378
Documents	840
Annual Growth Rate %	7.87
Document Average Age	7.63
Average citations per document	39.16
References	7292
Authors	2505
Authors of single-authored documents	51
Single-authored documents	57
Co-authors per document	3.62
International co-authorships %	27.02
Article	542
Book chapter	8
Conference paper	290

Table 1 represent the main information about data in analysis included 840 documents on statistical distributions in machine learning from 378 sources, published between 1997 and 2026. These publications include 7,292 cited references, showing a solid scholarly foundation. The dataset's annual growth rate is 7.87 percent, which reflects growing research interest in

probabilistic machine learning methods. In total, 2,505 authors contributed to these publications. Most papers have multiple authors, which suggests a collaborative research culture. The many sources and contributors' points to the interdisciplinary nature of the field, covering statistics, computer science, AI and applied sciences.

3.1.2 Annual Scientific Production

Figure 1 illustrates the temporal evolution of publication output. Research activity remained relatively limited during the late 1990s and early 2000s, reflecting the dominance of deterministic machine learning approaches during that period. A gradual increase is observed after 2005, coinciding with growing interest in Bayesian learning, probabilistic graphical models and statistical inference techniques. The sharp growth after 2017 corresponds with the rapid development of artificial intelligence, deep learning and uncertainty-aware learning frameworks. This pattern suggests that statistical distributions have become increasingly important as machine learning systems have evolved toward more complex and data-intensive applications.

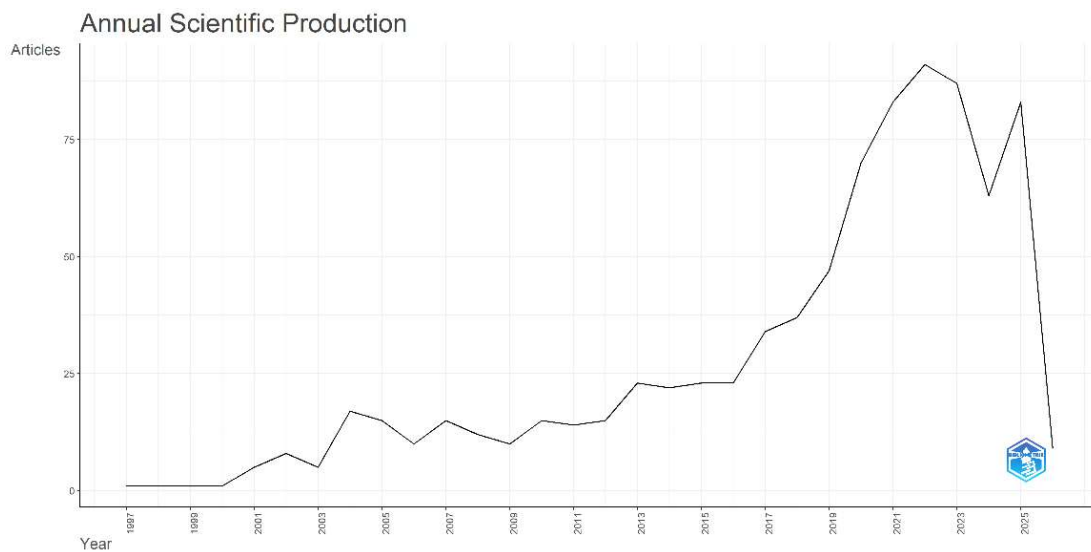


Figure 1: Annual Scientific Production

3.1.3 Average Citations per Year

Figure 2 shows fluctuations in average citations per year, with notable peaks around 1999, 2008, and 2016. These peaks coincide with periods during which influential methodological developments gained prominence. The early peak reflects growing interest in statistical learning theory and probabilistic graphical models, while the 2008 increase corresponds to the wider adoption of Bayesian methods and latent variable models. The peak observed around 2016 aligns with advances in deep generative modelling, variational inference, and uncertainty quantification. These citation patterns indicate that methodological innovations have served as major drivers of intellectual development within the field.

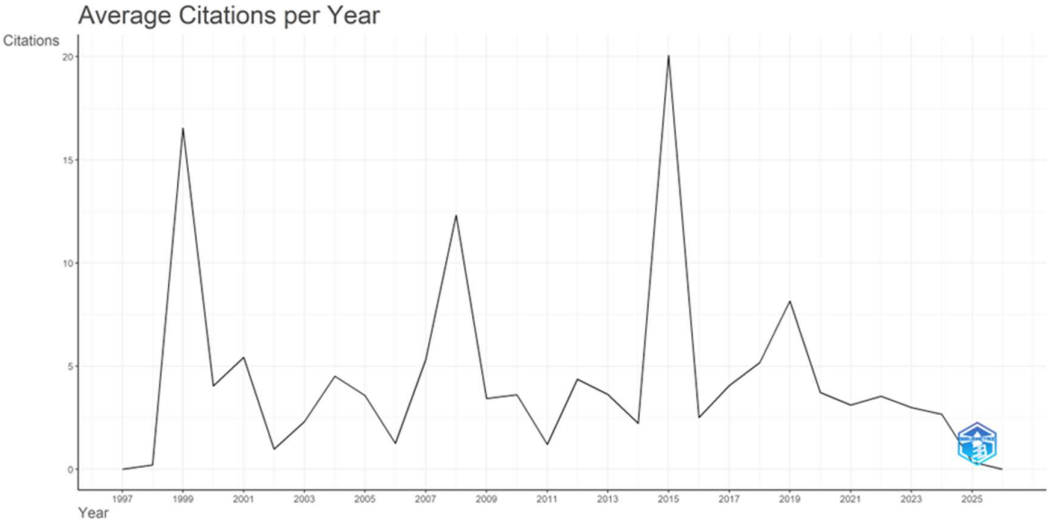


Figure 2: Average Citations per Year

3.1.4 Life Cycle of Publications

The cumulative growth curve and annual publication life-cycle indicate that research on statistical distributions in machine learning has progressed from an emerging research area to a mature and rapidly expanding field. The growth trajectory demonstrates a transition from foundational theoretical investigations toward broader methodological and application-oriented studies. The observed acceleration after 2017 further confirms the increasing relevance of probabilistic approaches in addressing uncertainty, interpretability, and reliability challenges in artificial intelligence.

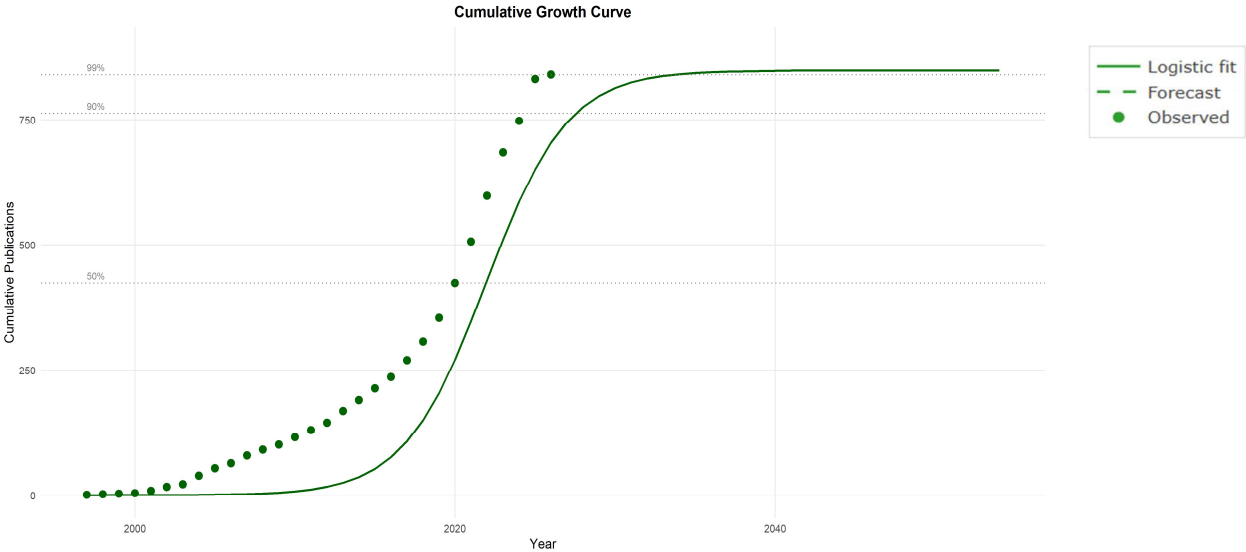


Figure 3: Cumulative Growth Curve

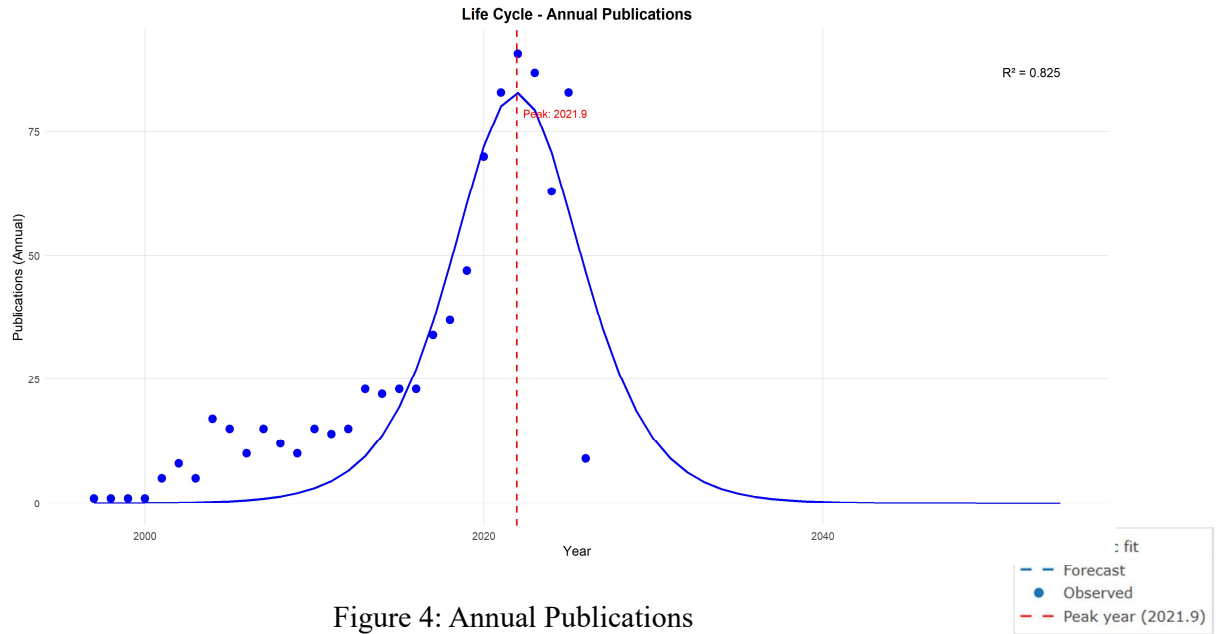


Figure 4: Annual Publications

3.1.5 Three Field Plot

The three-field plot (Figure 5) reveals the knowledge flow connecting influential references, productive authors, and dominant research themes. Highly cited foundational works in probabilistic modelling and machine learning are strongly linked to productive researchers such as Wang J., Wang Y., and Zhang J., who in turn are associated with themes including probability distributions, Bayesian networks and machine learning. This pattern suggests that contemporary research is strongly anchored in classical statistical learning theory while simultaneously expanding toward modern probabilistic and uncertainty-aware machine learning frameworks. The figure highlights how foundational theoretical contributions continue to shape current research priorities and emerging methodologies.

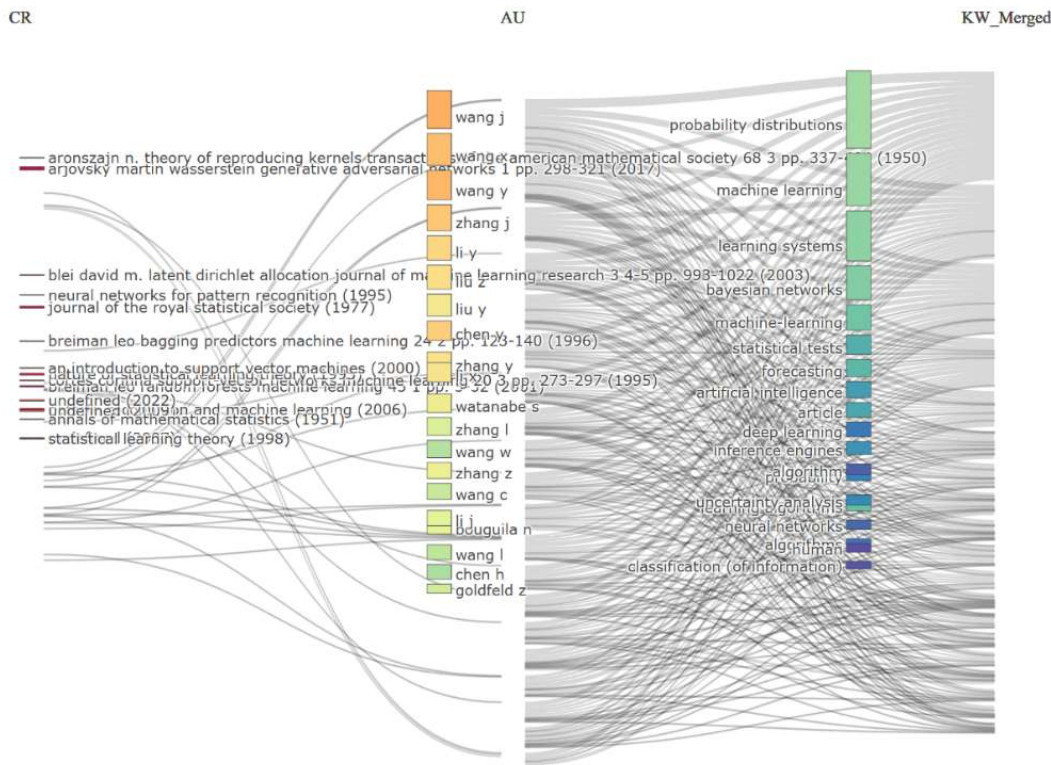


Figure 5: Three Field Plot

3.1.6 Most Relevant Source

Table 2 shows the journals and publications with the most papers on statistical distributions in machine learning. Lecture Notes in Computer Science is the source with the most publications, which suggests it's important for sharing new research in this area from conferences. Proceedings of Machine Learning Research and Neurocomputing follow, with many studies on probabilistic modelling, learning algorithms and distribution-based methods. Journals like IEEE Access, Neural Networks and the Journal of Machine Learning Research also suggest that the topic has interest in both specialized and general publications.

Table 2: Most Relevant Sources

Source	Articles
Lecture Notes in Computer Science	45
Proceedings of Machine Learning Research	28
Neurocomputing	18
IEEE Access	17
Neural Networks	15
Journal of Machine Learning Research	11
IEEE Transactions on Information Theory	10
Geophysical Journal International	9
Advances in Neural Information Processing Systems	8

The variety of publications from these sources shows that research on statistical distributions in machine learning is interdisciplinary. The results appear in different journals and conferences for different groups. This shows the field has become accepted in many areas of science and engineering.

The curves in Figure 6, which track publication trends from key sources, show how various journals have aided this field's expansion. Lecture Notes in Computer Science consistently increased its publication rate during the period studied, confirming its sustained importance as a main platform for novel studies. Neurocomputing grew noticeably, especially after 2015, suggesting rising interest in machine learning using distribution in neural and computational modelling research.

Proceedings of Machine Learning Research has increased sharply in recent years, which mirrors the rising importance of conference proceedings in fast-moving areas like machine learning and AI. At the same time, IEEE Access and Neural Networks have grown steadily but moderately, adding regularly to the existing literature. These patterns suggest a balanced publishing situation where both established journals and newer platforms actively contribute to shaping the field.

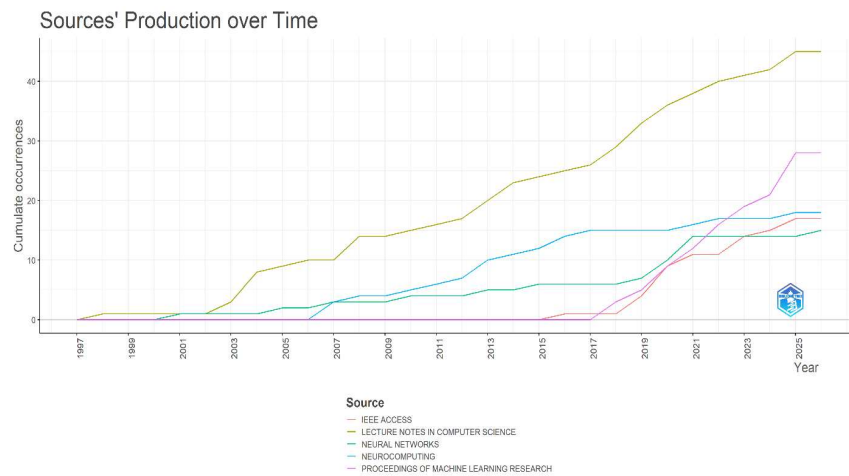


Figure 6: Sources Production Over Time

Table 3 shows that source local impact indicators gauge the influence of journals in this area of study. Journals like Neurocomputing have the highest impact values, because many of their articles are often cited by other papers in this dataset. This hints that Neurocomputing is important for publishing influential research that is commonly used by scholars working on statistical distributions in machine learning.

Other sources, like Neural Networks, IEEE Access, and Lecture Notes in Computer Science, also have strong citation numbers, which reflect their importance in spreading relevant research. Some journals publish many papers, but their influence changes based on how often those papers are cited. This change shows the difference between productivity and scholarly impact and proves that both things are important when measuring a source's contribution to the field.

Table 3: Sources Local Impact

Source	h_index	g-index	m-index	Total Citations (TC)	NP	Publication year
Neurocomputing	19	86	0.679	7567	126	1999
Neural Networks	11	18	0.550	1542	18	2007
IEEE Access	10	15	0.385	318	15	2001
Lecture Notes in Computer Science	9	41	0.310	1697	45	1998
Proceedings of Machine Learning Research	8	11	0.889	152	28	2018
IEEE Transactions on Information Theory	7	10	0.280	285	10	2002
IEEE Transactions on Neural Networks	7	7	0.292	343	7	2003
Monthly Notices of the Royal Astronomical Society	7	8	0.333	175	8	2006
Geophysical Journal International	6	9	0.353	266	9	2010

Bradford's Law of Scattering, as shown in Figure 7, says that a few core journals put out many papers, while many other journals put out only a few. The Bradford's Law plot in this work has a clear core zone where a few sources account for a large part of the total research.

Journals like Lecture Notes in Computer Science, Proceedings of Machine Learning Research, Neurocomputing, IEEE Access, Neural Networks, Journal of Machine Learning Research and IEEE Transactions on Information Theory are the main sources for research on statistical distributions in machine learning. These journals are the main places where new ideas and results are shared. Outside of this core, many other journals add smaller numbers of papers, which shows how interdisciplinary the field is. This result backs up the idea that while research is spread out, a few key sources control academic communication in this area.

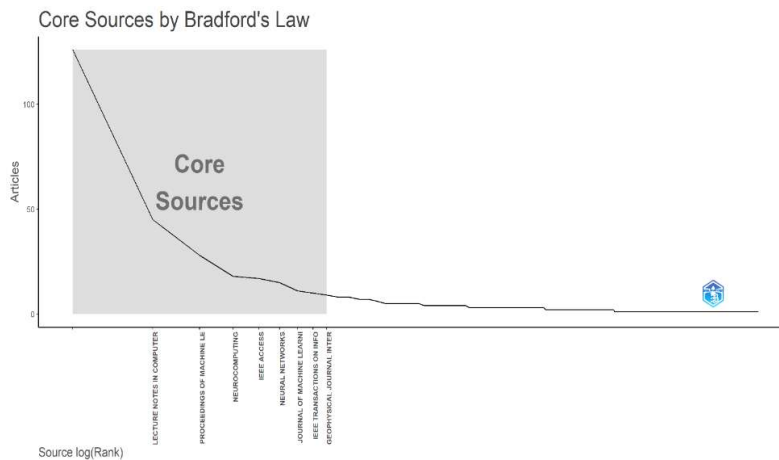


Figure 7: Core Sources by Bradfords 's Law

3.1.7 Top Productive and Highly Cited Authors

To understand how individual researchers help the study of statistical distributions in machine learning, we can look at the number of papers they publish and how often their work is cited. Authors who publish a lot are considered productive, and those whose work is cited often have a big impact on the field. Table 4 indicates that Wang J, Wang Y, Wang X, and Zhang J are highly productive, with each publishing several papers. Their ongoing work shows their strong interest and involvement in creating machine learning methods based on distributions.

Table 4: Top Productive Authors

Author	Articles	Articles Fractionalized
Wang J	12	2.50
Wang Y	11	2.94
Wang X	10	2.01
Zhang J	10	3.38
Chen Y	9	2.20
Li Y	9	1.55
Liu Z	9	2.26
Zhang Y	9	2.79
Li X	8	2.29
Liu Y	8	2.18

Local citation analysis pinpoints the authors whose work is cited most often within the dataset. This shows their influence on statistical distributions in machine learning. Instead of global citations, which assess overall academic influence, local citations reveal how much an author's research shapes studies in this area. Authors with many local citations are key to the development of ideas and methods here.

Figure 8 shows that Signoretto M and Suykens J.A.K. have the most local citations, meaning their work is commonly used as a basis by researchers in distribution-based machine learning. Cuturi M and Peyré G also have high local citation values, showing their important role in advancing probabilistic modelling and computational methods. Other authors, like Bárány V, Blei D.M., Griffiths T.L. and Jordan M.I., are notable in the citation network, which reflects their basic contributions to probabilistic learning and Bayesian modelling. These trends suggest that a small group of researchers has greatly shaped the intellectual direction of this area.

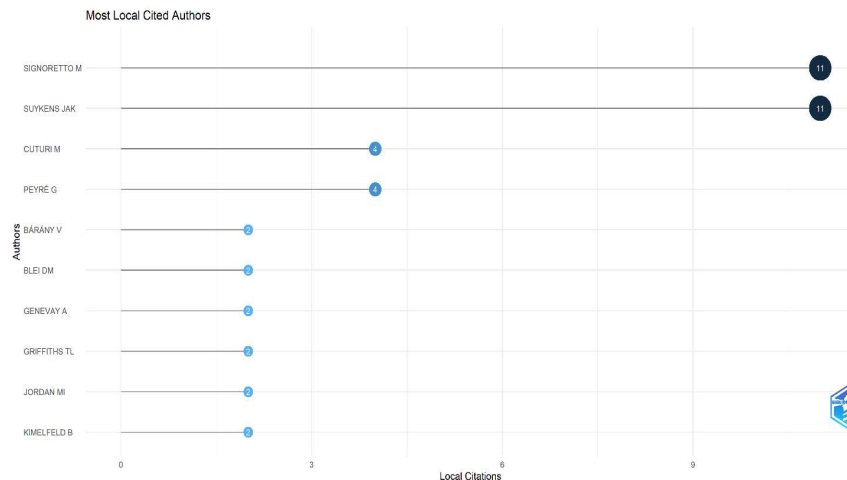


Figure 8: Most Locally Cited authors

3.1.8 Word Cloud

Figure 9 shows the most common keywords in the data; bigger words mean they appear more often. The key words probability distributions, machine learning, learning systems, and Bayesian networks suggest that probabilistic modelling is very important to this field. These words show how statistical theory and modern learning methods are closely linked. Other keywords like artificial intelligence, neural networks, uncertainty analysis, Gaussian distribution, and regression analysis tell us that distribution-based methods are popular in both theoretical and applied machine learning. These keywords mean that this research focuses on building models and using probabilistic thinking to solve actual problems.



Figure 9: Word Cloud

3.2 Network Analysis

Network analysis helps people get a better sense of how machine learning research on statistical distributions is structured via relationships among writers, countries, papers, and topics. Unlike basic stats, which focus on how many papers are published and how often they are cited, network methods check how different things link together and how info moves in the research world. Showing these relationships as networks helps spot teamwork structures, important studies, and topic links that shape how the field grows.

This work built various types of bibliometric networks with Biblioshiny, like co-authorship, country teamwork, co-citation and keyword networks. These networks show groups of researchers who often work together, countries with strong research links, groups of very important papers and closely related research subjects. Looking at these networks helps us understand the social and intellectual structure of machine learning research on statistical distributions and to spot the key areas that push the field forward.

3.2.1 Countries Collaboration

Figure 10 reveals a highly internationalised research landscape. The dominance of the United States and China reflects the concentration of global investments in artificial intelligence, advanced computing infrastructure, and large-scale machine learning research programmes. Their central position in the collaboration network suggests that future methodological developments are likely to be strongly influenced by research priorities emerging from these countries. The participation of countries across North America, Europe and Asia demonstrates that research on statistical distributions has become a globally distributed endeavour. Such collaboration facilitates methodological innovation by integrating expertise from statistics, computer science, and application-specific domains.

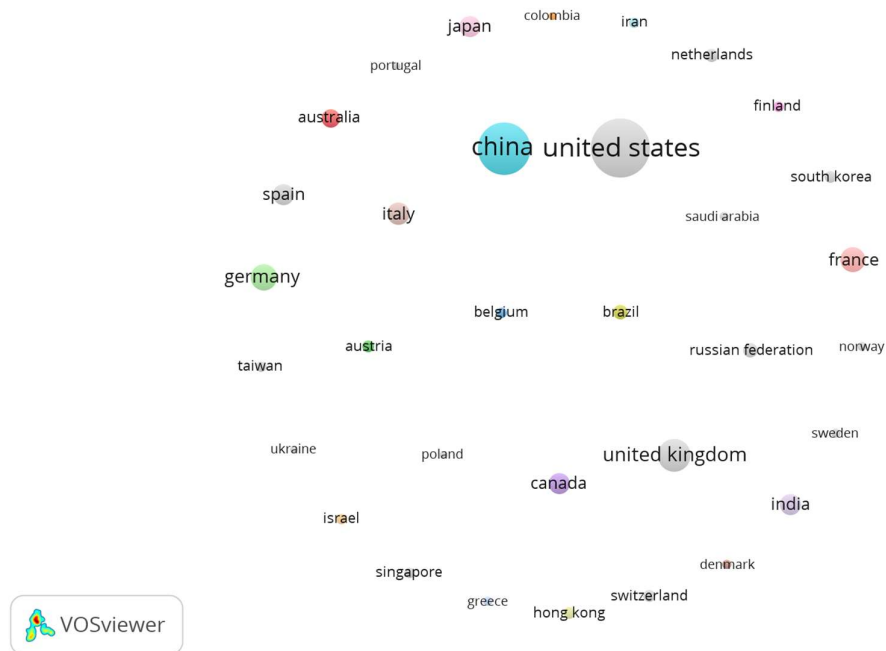


Figure 10: Country wise collaborations

3.2.2 Collaboration network

The co-authorship network (Figure 11) reveals several well-defined collaborative clusters, indicating that knowledge production in this field is strongly team-based. Authors such as Wang J., Wang X., Liu Z., and Zhang J. occupy central positions within the network and act as important connectors among research groups. The presence of multiple clusters rather than a single dominant community suggests that the field is characterised by diverse research directions, including theoretical developments, methodological innovation and application-oriented studies. The network structure further highlights the importance of interdisciplinary collaboration in advancing machine learning methodologies based on statistical distributions.

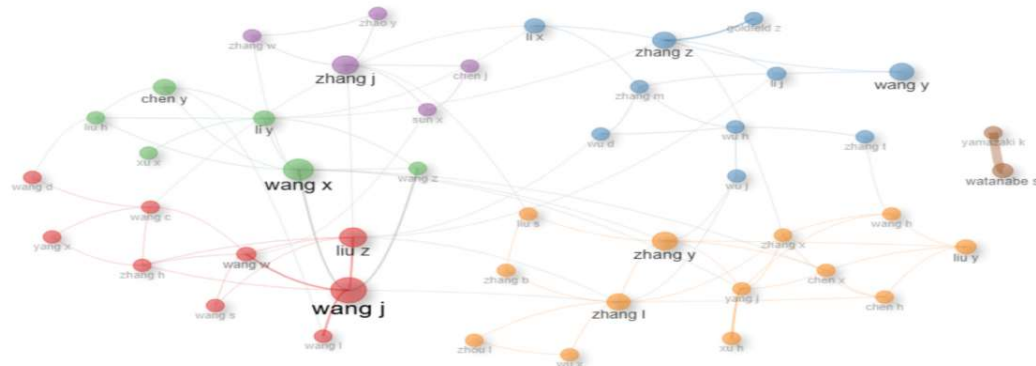


Figure 11: Collaboration Network

3.2.3 Co-citation Network

The co-citation network (Figure 12) reveals two major intellectual clusters. The first cluster is centred on specialised machine learning and statistical journals, including Journal of Machine Learning Research, IEEE Transactions on Information Theory, Advances in Neural Information Processing Systems and Journal of the American Statistical Association. This cluster represents the theoretical and methodological foundation of the field. The second cluster comprises multidisciplinary journals such as Science and Nature, reflecting the growing influence of distribution-based machine learning research beyond traditional computational disciplines. The interaction between these clusters demonstrates how foundational statistical theory has evolved into widely applicable machine learning methodologies with broad scientific impact.

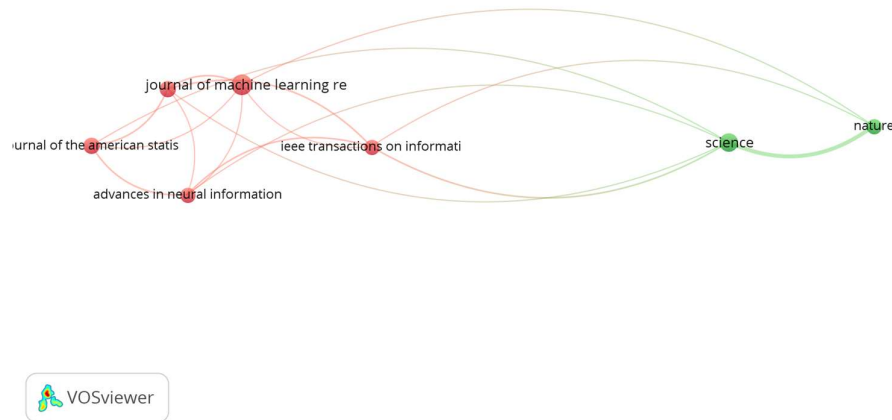


Figure 12: Co-citation Network

The separation of the co-citation network into specialised machine learning journals and multidisciplinary scientific journals reflects an important stage in the maturation of the field. Early research primarily focused on theoretical developments in statistical learning and probabilistic inference, whereas more recent studies increasingly address real-world applications in healthcare, environmental science and autonomous systems. The coexistence of these clusters indicates that statistical distributions have evolved from a methodological concern to a broadly applicable scientific tool.

3.2.4 Historiograph

The historiograph (Figure 13) illustrates the chronological evolution of the field through direct citation relationships. Early influential studies established the theoretical foundations of statistical learning, Bayesian inference, and probabilistic modelling. Subsequent research expanded these concepts into areas such as latent variable modelling, generative learning, and uncertainty-aware artificial intelligence. The historiographic structure confirms that the field has evolved cumulatively, with contemporary advances building directly upon a stable body of foundational theoretical work.

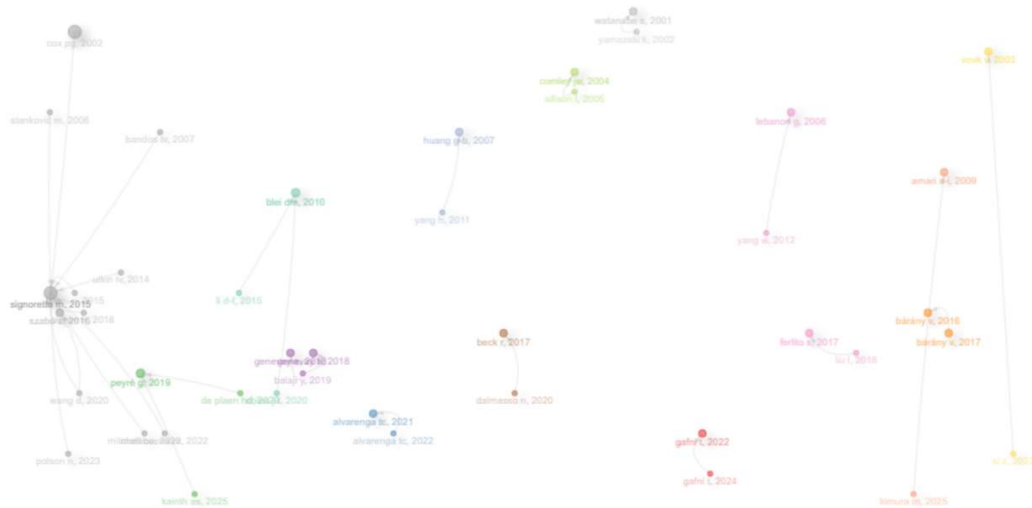


Figure 13: Historiograph

3.2.5 Co-occurrence Network

The keyword co-occurrence network (Figure 14) demonstrates that probability distributions, machine learning, learning systems, and Bayesian networks constitute the conceptual core of the field. The close association among these keywords highlights the central role of probabilistic reasoning in machine learning development. Additional clusters associated with neural networks, uncertainty analysis and artificial intelligence indicate the increasing integration of statistical distributions into advanced computational learning frameworks. These findings suggest a gradual shift from traditional statistical modelling toward uncertainty-aware and explainable machine learning systems.

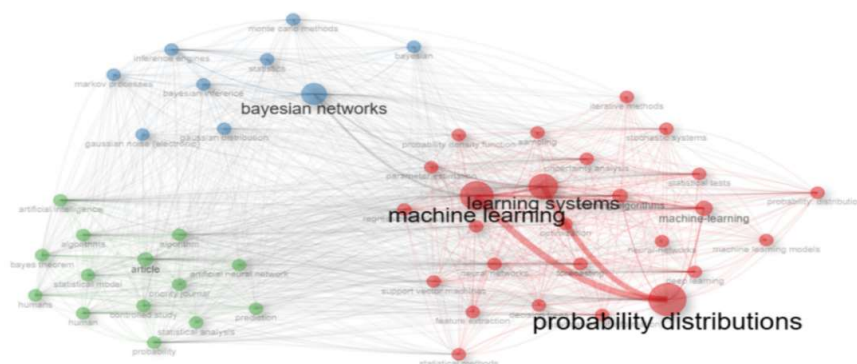


Figure 14: Co-occurrence Network

3.2.6 Thematic Map

The thematic map (Figure 15) provides insight into the maturity and relevance of research themes. Probability distributions, machine learning and learning systems occupy the motor-theme quadrant, indicating that they are both highly developed and central to the field's intellectual structure. Support vector machines and feature extraction appear as specialised niche themes, reflecting their continued relevance within specific methodological contexts. Themes such as Bayesian networks and Gaussian distributions are positioned within the emerging or declining quadrant. This classification does not imply reduced importance; rather, it suggests a transition in research focus. Many concepts associated with Bayesian networks and Gaussian modelling have become integrated into broader frameworks such as deep probabilistic learning, generative modelling, and uncertainty-aware artificial intelligence. Consequently, the thematic map reflects the dynamic evolution of research priorities rather than the disappearance of these concepts.

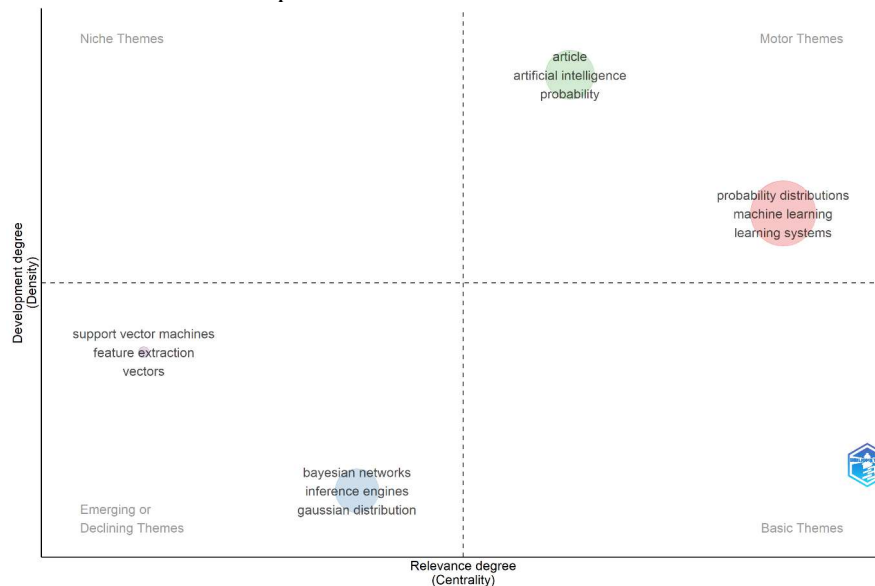


Figure 15: Thematic Map

3.3 Discussion

Collectively, the findings indicate that research on statistical distributions in machine learning has evolved from a predominantly theoretical area into a mature and interdisciplinary research domain. The strong growth in publications, extensive international collaboration and emergence of sophisticated probabilistic methodologies demonstrate the increasing importance of uncertainty modelling in artificial intelligence. The integration of machine learning journals with multidisciplinary scientific sources further highlights the broad applicability of distribution-based methods across diverse research fields.

The thematic and network analyses reveal a clear intellectual transition from classical statistical learning approaches toward probabilistic deep learning, generative modelling and trustworthy artificial intelligence. These developments suggest that future research is likely to focus on uncertainty quantification, explainability, robustness and the development of flexible statistical distributions capable of modelling increasingly complex data structures. Consequently, statistical distribution theory is expected to remain a fundamental component of next-generation machine learning research.

The prominence of authors from East Asia, particularly China, reflects the rapid expansion of artificial intelligence research across the region during the last decade. However,

this pattern may also be influenced by database coverage and publication practices within Scopus. Consequently, productivity rankings should be interpreted alongside potential database and language-related biases.

3.4 Applications and Potential Research Areas of Statistical Distributions in Machine Learning

Statistical distributions are very important in machine learning. They allow models to show uncertainty, variability, and probability that exist in real-world data. In tasks like classification, regression, clustering, and forecasting, distributions are used to model likelihood functions, error structures, and predictive uncertainty. These functions are important in areas like healthcare, finance, environmental studies, and self-driving systems, where choices must be made with incomplete info (Signoretto, M., et.al, (2012)). By adding distributional assumptions, machine learning models give not only predictions but also measures of confidence, improving how reliable they are (Suykens, J. A. K., et.al., (2002)).

In today's machine learning, statistical distributions support methods like Bayesian learning, probabilistic graphical models, ensemble ways, and deep generative models. Things like Bayesian neural networks, variational autoencoders, and stochastic simulation depend on probability distributions to learn latent representations, create synthetic data, and conduct inference that accounts for uncertainty. These have been used in image and signal processing, anomaly detection, recommendation systems, and time-series analysis, showing how useful distribution-based modelling is (Yamazaki, K., & Tanaka, H. (2002)).

Outside of current applications, statistical distributions could be used in machine learning research. New areas include uncertainty-aware and trustworthy AI, where probabilistic modelling can make things more reliable and easier to understand. There are also chances to create adaptable distribution families, like mixture, heavy-tailed, and nonparametric distributions, made for complex and big data (Watanabe, S. (2001)). Combining statistical distributions with explainable AI and risk-sensitive learning is a good direction for improving both theory and use of machine learning systems. These research areas show how important statistical distributions are in forming future machine learning methods.

3.5 Limitations and Future scope

Despite providing a comprehensive overview of research on statistical distributions in machine learning, this study has several limitations. First, the analysis relies exclusively on the Scopus database, which may not fully capture relevant publications indexed in Web of Science, IEEE Xplore or other specialised repositories. Consequently, some influential conference proceedings and emerging research contributions may not be represented. Second, the findings are dependent on the selected search strategy and bibliographic metadata, which may exclude studies employing alternative terminology. Third, bibliometric indicators primarily measure publication and citation characteristics and do not directly evaluate methodological quality or practical effectiveness.

Future research could address these limitations by integrating multiple databases and employing advanced text-mining, topic-modelling and machine-learning techniques to enhance thematic identification. Comparative investigations of specific distribution families, such as Bayesian, mixture, heavy-tailed and nonparametric distributions, would provide deeper insights into their relative influence across machine learning applications. Additionally, the growing emphasis on explainable, trustworthy and uncertainty-aware artificial intelligence presents significant opportunities for further exploration of probabilistic and distribution-based methodologies.

4 Conclusion

This study provides a comprehensive bibliometric and network-based assessment of research on statistical distributions used in machine learning between 1997 and 2025. The findings demonstrate sustained growth in scientific output and increasing international collaboration, reflecting the expanding importance of probabilistic approaches within machine learning research. Bibliometric indicators reveal a concentration of influential contributions within a relatively small group of journals and authors, while network analyses highlight the central roles of probabilistic modelling, Bayesian learning, uncertainty quantification and generative methods in shaping the field.

The thematic and co-citation structures indicate that research has evolved from foundational statistical learning frameworks toward more sophisticated approaches that address uncertainty, interpretability, and reliability in artificial intelligence systems. The presence of strong connections between specialised machine learning journals and high-impact multidisciplinary sources further demonstrates the broad scientific relevance of statistical distribution theory. Overall, the study contributes a structured understanding of the intellectual, social and thematic evolution of research on statistical distributions in machine learning. The findings provide a useful foundation for future investigations and suggest that probabilistic and distribution-based methodologies will remain central to the development of next-generation machine learning systems.

References

- [1]. Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein generative adversarial networks. *Proceedings of the 34th International Conference on Machine Learning*, 214–223.
- [2]. Aria, M., & Cuccurullo, C. (2017). Bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, 11(4), 959–975.
- [3]. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022.
- [4]. Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2), 123–140.
- [5]. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- [6]. Bishop, C. M. (2006). Pattern recognition and machine learning. *Springer*.
- [7]. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297.
- [8]. Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems*, 26, 2292–2300.
- [9]. Donthu, N., Kumar, S., Mukherjee, D., Pandey, N. & Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research*, 133, 285–296.
- [10]. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.
- [11]. Griffiths, T. L., & Ghahramani, Z. (2005). Infinite latent feature models and the Indian buffet process. *Advances in Neural Information Processing Systems*, 18, 475–482.
- [12]. Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., & Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine Learning*, 37(2), 183–233.
- [13]. Kingma, D. P., & Welling, M. (2014). Auto-encoding variational Bayes. *Proceedings of the International Conference on Learning Representations*, 1–14.

- [14]. Peyré, G., & Cuturi, M. (2019). Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5–6), 355–607.
- [15]. Signoretto, M., Van de Plas, R., De Moor, B., & Suykens, J. A. K. (2013). Nuclear norms for tensors and their use for convex multilinear estimation. *Linear Algebra and Its Applications*, 438(8), 3855–3873.
- [16]. Suykens, J. A. K., Van Gestel, T., De Brabanter, J., De Moor, B., & Vandewalle, J. (2002). Least squares support vector machines. *World Scientific*.
- [17]. Watanabe, S. (2001). Algebraic geometry and statistical learning theory. *Neural Networks*, 14(9), 1259–1273.
- [18]. Yamazaki, K., & Tanaka, H. (2002). Probabilistic modelling in machine learning. *Neural Computation*, 14(8), 1809–1834.