The use of Machine learning to energy research: a bibliometric and geopolitical focus

Alfredo Morris
Área de Gestión Eficiente de la Energía
E2 Energía Eficiente, Barranquilla, Atlántico, Colombia
Email- amorris@e2energiaeficiente.com

Sofia Orjuela
Department of Administrative Sciences
Universidad Francisco de Paula Santander, Cúcuta, Norte de Santander, Colombia
Email- sofiaorjuela@ufps.edu.co

Carlos Acevedo
Engineering Faculty
Universidad Francisco de Paula Santander, Cúcuta, Norte de Santander, Colombia
Email- carloshumbertoap@ufps.edu.co

Abstract- The objective of this article is to make a general bibliometric analysis of scientific publications (articles, procedures, books, etc.) related to the implementation of machine learning in processes or research associated in one way or another with energy. A simple and general review of the Web of Science scientific database from which the bibliographic information for this analysis was extracted and was carried out. Data processing and breakdown was carried out with the help of two bibliographic software packages: Histcite® and VOSViewer®, both supported by MS Excel®. It was found that the number of scientific publications associated with this topic is on the rise, being led by world economic powers such as the United States of America and China, who set the procedural guidelines for research on this subject. Therefore, this article is presented as a starting point for those new researchers who are interested in knowing the current state of research associated with machine learning in the area of energy.

Keywords – Machine Learning; Systemic Review; Energy; Geopolitics; Bibliometric

I. INTRODUCTION

Autonomous learning, or machine learning, is one of the disciplines related to artificial intelligence that has received the most attention from the scientific world in recent years. This is reflected in a large number of publications and in a large number of developments in this area [1]. This growth is consistent with the massive amount of data currently being manipulated [2], making it necessary to use analysis tools that help with decision making for a given problem. The versatility of this technology by offering a wide range of algorithms in handling data from different sources [3], has extended its use in different areas of human knowledge, allowing us to find the bulk of publications associated with the use of machine learning in topics related to logistics in the manufacturing industry [4], the future implementation of "smart cities" [5] and, the most common use, in data processing for the public health area [6], with bibliometric analysis articles that serve as a reference for the scientific community seeking to have a starting point that will allow them to begin exploring this topic [1][7][8].

Among the publications related to the use of machine learning, it can be observed that the topics related to "energy," although widely developed [9], do not have an updated bibliometric reference to verify the current status of scientific articles, which makes the process of making research visible and obtaining information from reliable sources more complicated for new researchers entering the field [8]. This represents an alarming situation, knowing that the management and use of energy in a country represents a strong indicator of its level of development [10], which tends to promote the degree of investment in energy-related research processes, from a basic technological level to a more advanced level of development [11].

The present study seeks to make an exploratory bibliometric analysis of the publications related to machine learning and energy worldwide, through the usual study tools such as scientific maps, and analysis of trends in publications [1]. Similarly, and according to other bibliometric analyses that emphasize machine learning [5], seek to review trends in the use of the algorithms normally used in these scientific productions and the programs on which they perform these analyses. In addition, a geographical analysis of scientific publications will be carried out to find trends among machine learning related publications and the degree of energy consumption and development (measured by GDP) of a country.

The document is distributed in 3 sections: Methodology, results and discussion, and conclusions. The methodology will mention the process of obtaining the data, as well as its proper handling and processing. In the results and discussion section, the graphs and information obtained from the respective trend analysis carried out will be presented. Finally, in the conclusions, the main appreciations obtained from the analyses that will be seen throughout the document will be found.

II. METHODOLOGY

Although no specific methodology is used for this type of analysis, as commented by two saints in their bibliometric study of the application of machine learning and data mining in public health [6], it is important to have a reference point when performing this type of analysis. For that reason, for this exploratory study, we will take as reference some points mentioned by Binyousef in his manual for the elaboration of systematic studies of state of the art articles [8]. For this purpose, the objective will be to answer two main questions: What is the current status of information regarding the application in fields related to machine learning applied to energy? And what is the geopolitical landscape regarding this type of publication? It is important to mention that these questions will be sought to be answered on a general level of analysis.

To carry out this study, we proceeded to perform the carving process in Figure 1. First, to search one of the most important scientific databases, such as the Web of Science, initially looking for the term "Machine Learning." For this first case and since no initial filtering was done (the results were shown chronologically from 2000 to 2020), a total of 148,882 publications were obtained. Later, within the category, a filter was made limiting the search to publications that, in the same way, had the word "energy." This search yielded 7,584 publications. In performing this filtering, all publications related to energy in some way were considered, not only those associated with an energy management or energy handling, this with the intention of keeping this research at a general research level.

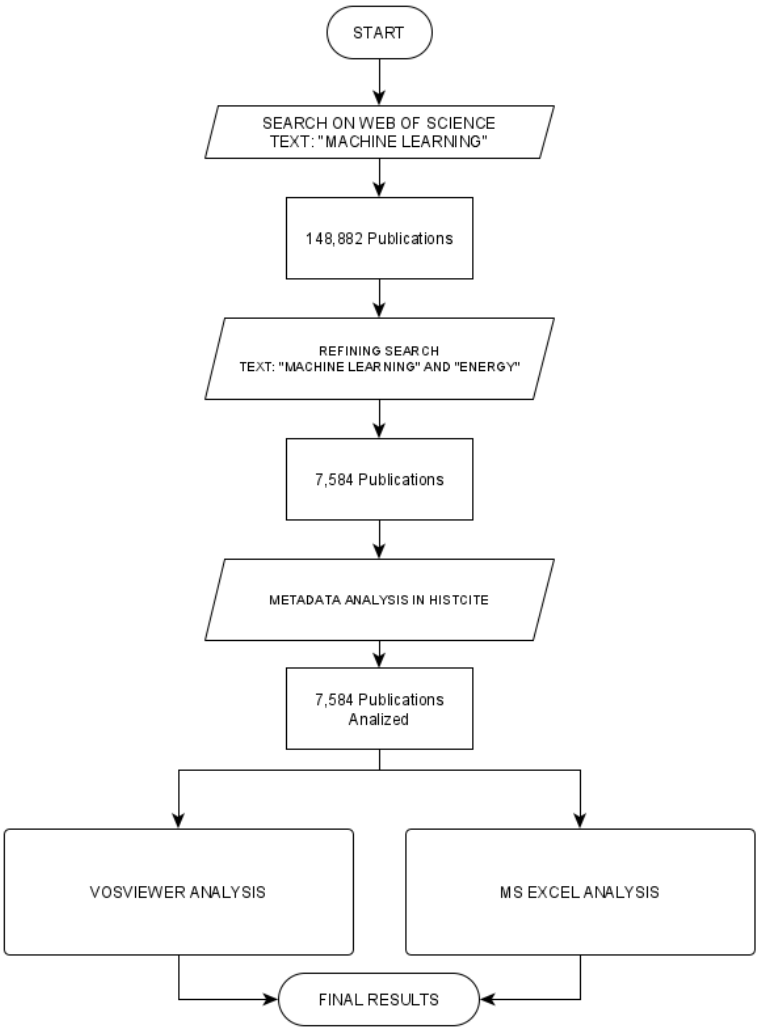


Figure 1 Search Flowchart

Later, having the information, the Histcite® bibliographic information analysis software was used, through which the exploratory analysis of the data was carried out. VOSViewer® was also used to perform graphical analysis through scientific maps, and finally, MS Excel® was used to complement the exploratory analysis performed with the use of Histcite®.

III. RESULTS AND DISCUSSIONS

After the analysis of the 7,584, a sequential growth over the years is observed, is possible to model this sequential behavior through the S curves, which model, with a good degree of confidence, the consumption and publication behavior of new technologies or products [12]. There are several ways to model this behavior. However, the equation with the formula set out in equation (1) will be used.

$$S(x) = min + (max - min) * \left(\frac{1}{1 + \exp(-kx)}\right)^{a} \quad \textbf{(1)}$$

For the present case, min refers to the minimum number of articles reported over the entire time frame analyzed. Max. refers to the maximum of expected publications related to machine learning and energy. Meanwhile, k and a are estimated parameters that have to be greater than zero [12].

After iterations to find the parameters that will result in a better R^2 estimate, with a value of 99.72%, the relationship represented in Figure 2 was found. It can be seen that up to March 2020 the number of scientific articles published already exceeds those published for 2015, and is projected to reach, according to the model, 2666 scientific productions, showing a great interest of the scientific community in this area of analysis, aligned with the political interest of the world community by supporting more initiatives that promote energy management. This increase in scientific production in machine learning applied to energy is in line with the objectives of sustainable energy development [12].

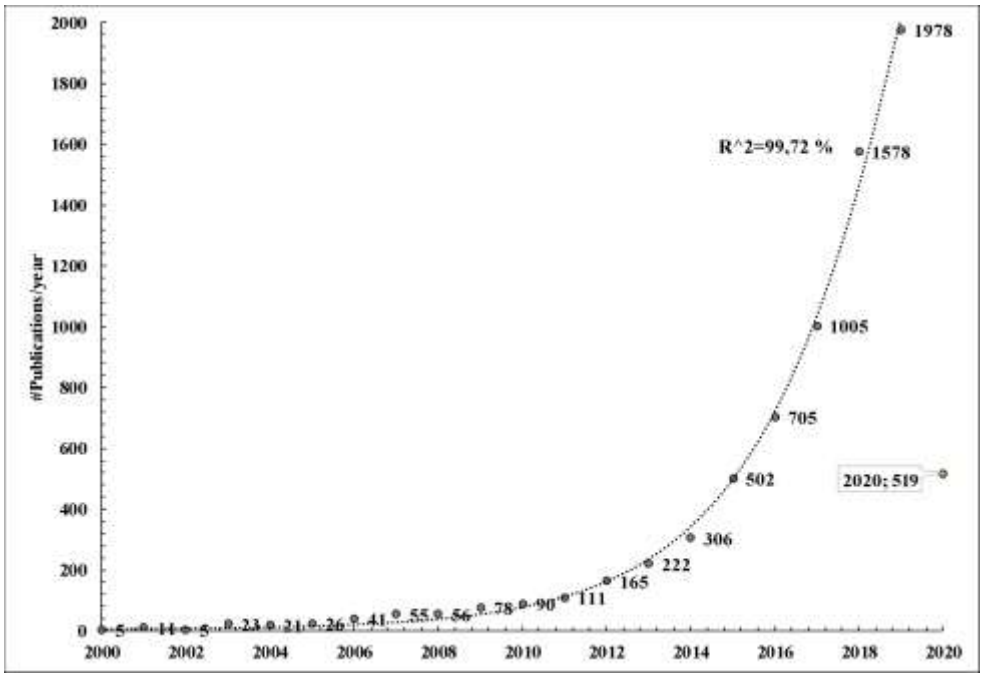


Figure 2 Trend of publications along the years

Table 1 Frequency of documents for each type of publication.

Type of Document	Number of Publications
Proceedings Paper	**2619**
Article	**4808**
Meeting Abstract	**17**
Review	**290**
Editorial Material	**22**
Book Chapter	**66**
Early Access	**81**
Data Paper	**15**
Retracted Publication	**6**

With respect to the type of documentation produced, it is possible to find that most of these are concentrated in articles published in journals and those procedures extracted from conferences, finding few productions associated with other types of documentation, as shown in Figure 3 it is possible to observe that although the articles and procedures cover 95% of scientific productions, there are still other types of documentation, such as literature review articles and meeting summaries associated with machine learning and application in energy.

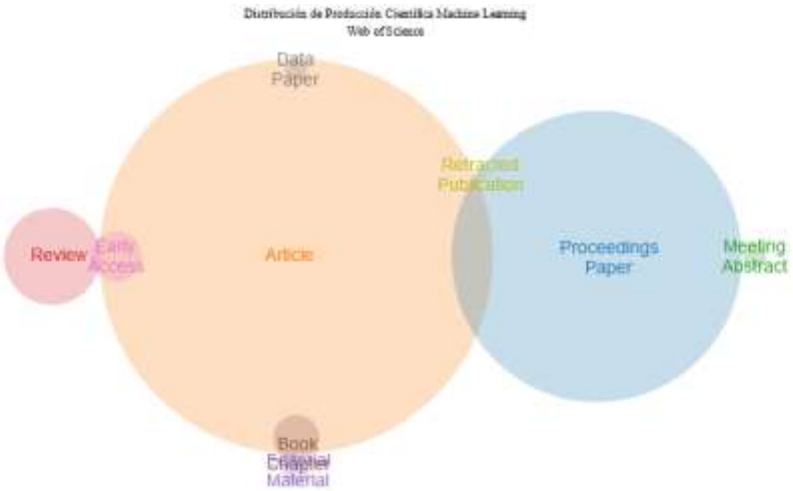


Figure 3 Venn diagram representing the diversity of documentation associated with machine learning and energy.

A co-author analysis of the bibliographic data obtained found something unusual. According to Figure 4, in spite of a large number of scientific productions, the collaboration between these is not so fruitful, with 121 groups of collaboration between authors, but few with a considerable density.

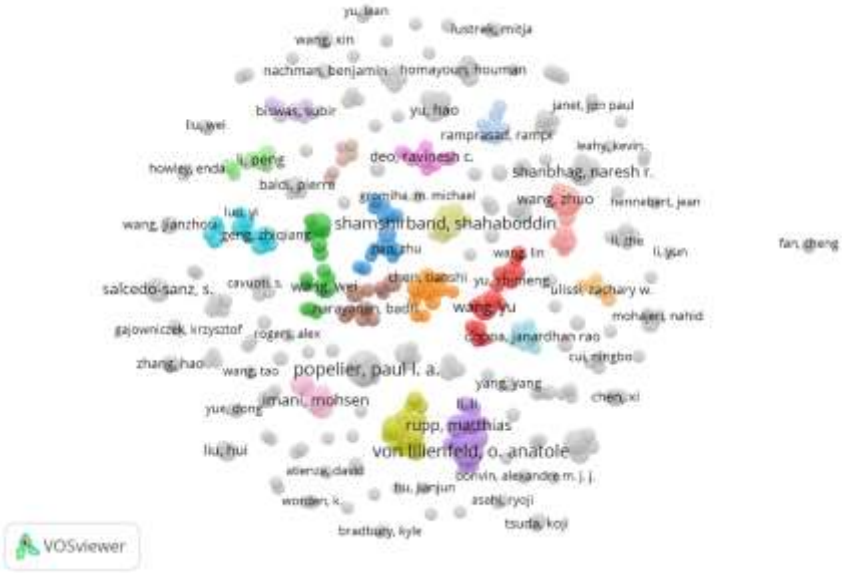


Figure 4 Scientific map of collaborations between authors

When filtering, taking into account the densest scientific collaborations, we find, as a result, the one found in Figure 5, where the number of groupings found is reduced to 11, and a collaborative pattern can be found. The scientific productions that are in the center of the map are directed more to developments related to the energy-saving associated with the architecture of the processors of calculation, as well as the development of more efficient algorithms that save time and energy in the processing. Meanwhile, the groupings of authors found outside the map

are related mostly to applications in different industries of these developments. It is stressed that these developments are not only related to the energy sector.

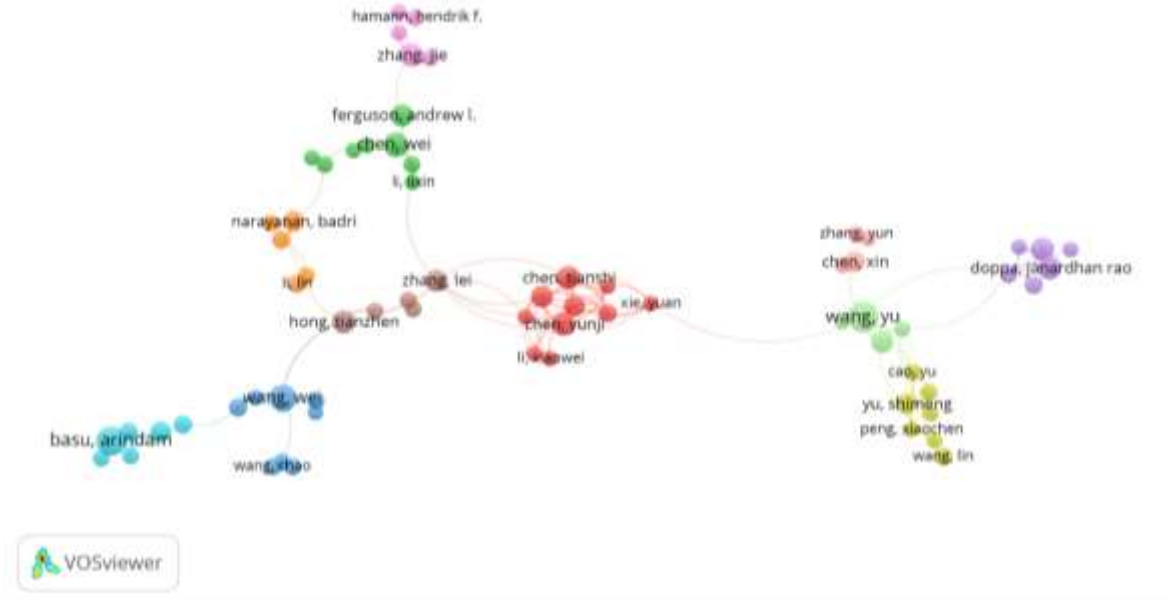


Figure 5 Scientific map of collaborations between authors with more publications

A similar analysis was conducted to evaluate the collaborative relationship between countries, showing strong collaborative relationships between 106 countries that have participated in the development of some scientific documents associated with Machine Learning and Energy, as shown in Figure 6.

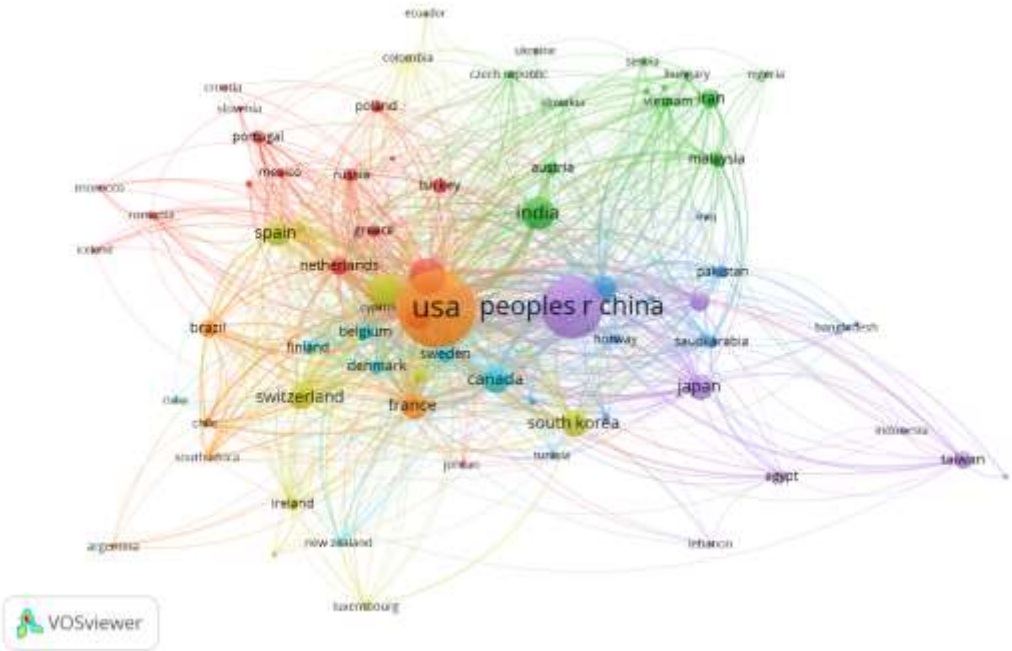


Figure 6 Scientific map between the collaborative environment of countries

From this map, it is possible to observe a great contribution of the United States of America and China in the generation of knowledge associated with machine learning and energy. This great relationship with development can be associated with their great purchasing power and the large investments they make in the area of research. This is

shown in more detail in Fig. 7, where the quantity of items produced is associated with GDP [13]. Similarly, it is possible to show, through the heat map, that the quality of the productions, measured with the TGCS (Total Global Citation Score) calculated with the help of Histcite®, coincides with this trend.

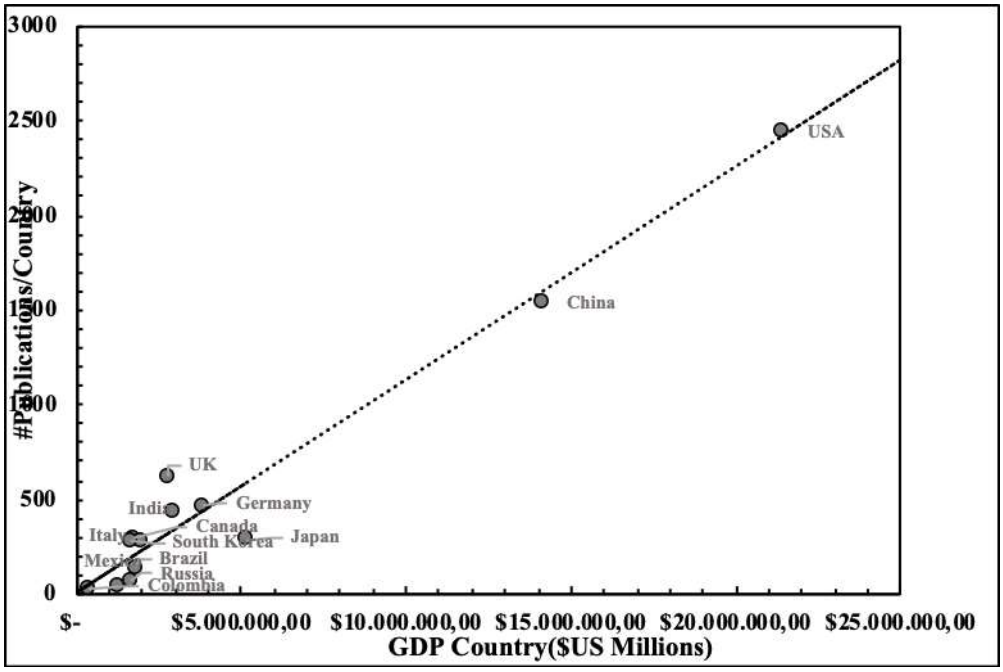


Figure 7 Relation between the number of publications and GDP of countries

This represents a strong indicator for the future study of this area of knowledge, supporting the information provided in Figure 2. Similarly, over the years, there has been a significant increase in the number of countries collaborating with each other, as shown in Figure 8 showing a greater tendency for more authors from various countries to collaborate as the years go by.

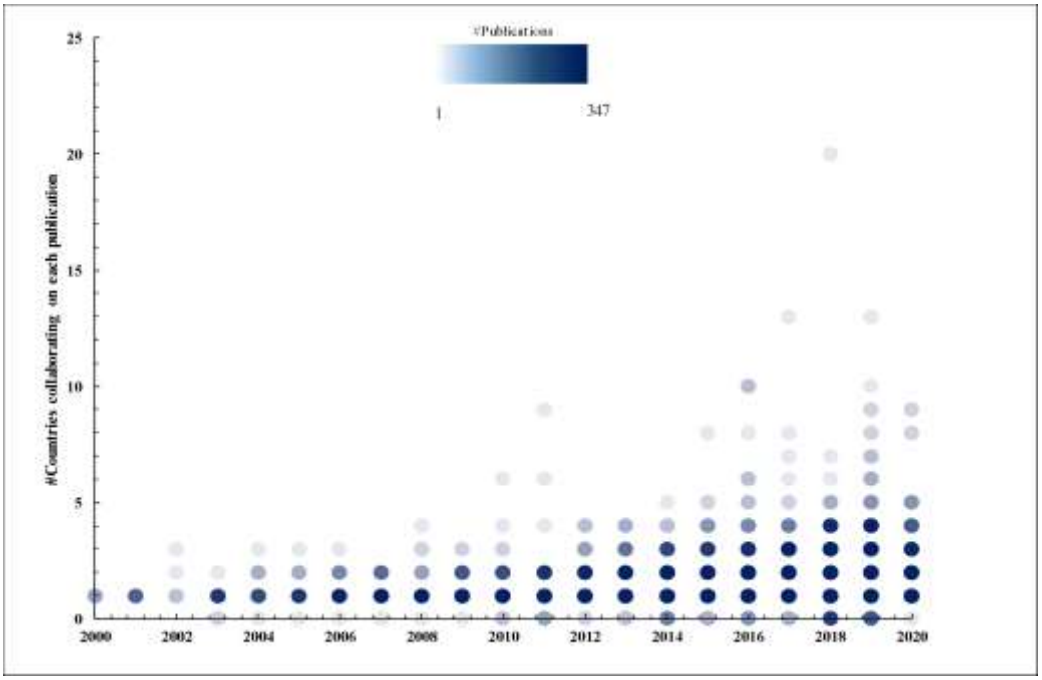
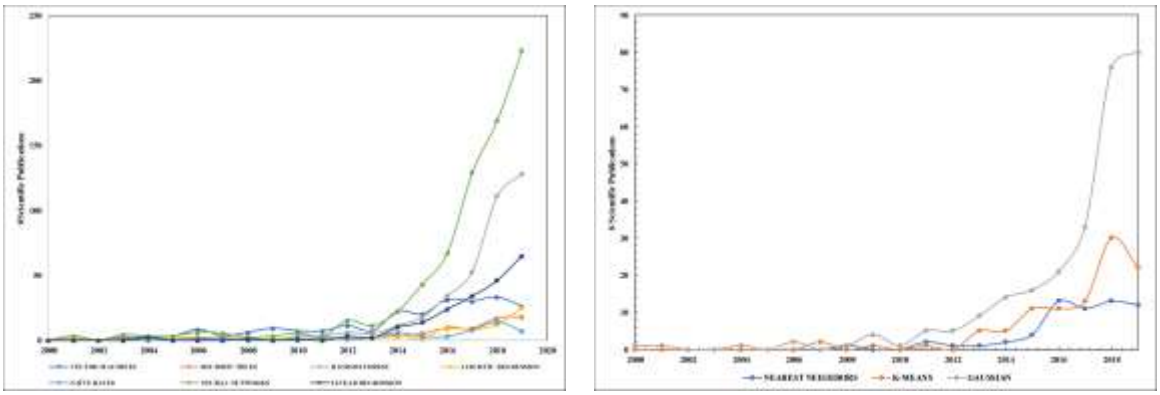


Figure 8 Analysis of collaboration between countries over the years.

Considering the greater relevance that the development of machine learning applied to energy will take in the coming years, it is important to know what the trends in research are [6]. For this purpose, an analysis of the trend in the use of algorithms over the years was carried out, as well as the most used programs in this type of research that were reported in the summary of publications in order to increase their visibility. Two different groups of algorithms were considered: Supervised and unsupervised algorithms. The trend analysis of the use of supervised and unsupervised algorithms are shown in Figure 9 (a) and Figure 9 (b).



(a) (b)

Figure 9 Trend of use of algorithms along the years (a) supervised (b) unsupervised

As illustrated in Figure 9, the use of neural networks is the most frequently mentioned algorithm in the summary of scientific outputs, among both supervised and unsupervised algorithms, with an increasing trend corresponding to the parallel increase in the number of publications. On the other hand, the second position is occupied by random forest algorithms, which by 2018 presented an increase in their frequency of use in scientific publications. Similarly, it is possible to observe that unsupervised algorithms are equally used but with a lower frequency, the greatest exponent being the Gaussian mixture algorithms. In addition, Figure 10 shows that among the countries that have consolidated the trend of using neural network algorithms for studies associated with machine learning and energy, correspond to the same countries that top the lists of productions, associating them with 60% of the publications corresponding to the use of neural networks.

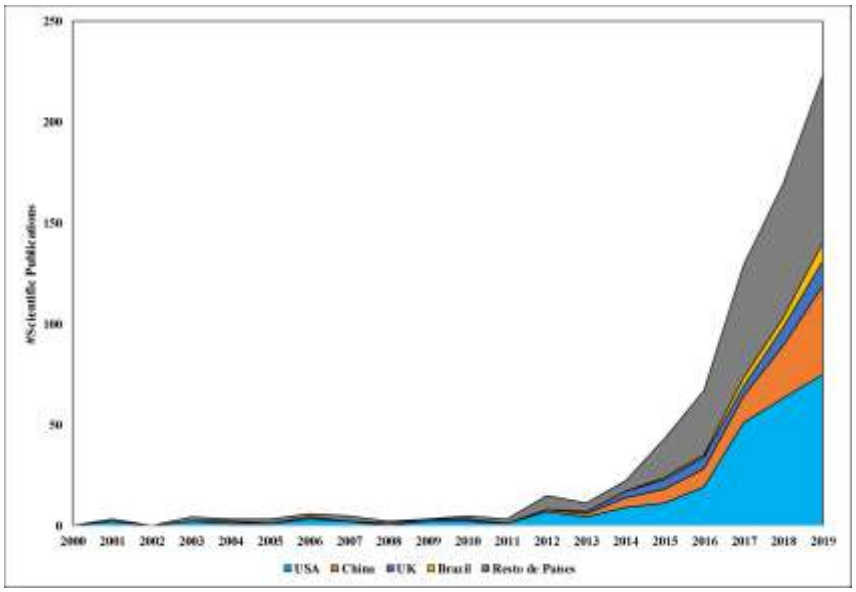


Figure 10 Associated publications of neural networks represented by countries

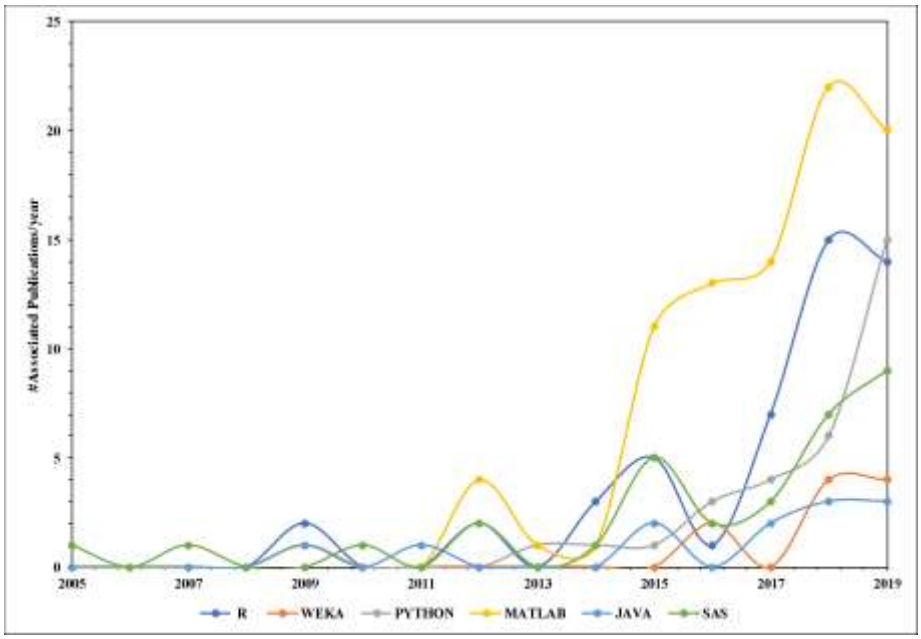


Figure 8 Number of publications associated with machine learning software per year.

IV. CONCLUSIONS

As illustrated in Figure 10, the use of neural networks is the most frequently mentioned algorithm in the summary of scientific outputs, among both supervised and unsupervised algorithms, with an increasing trend corresponding to the parallel increase in the number of publications. On the other hand, the second position is occupied by random forest algorithms. The present study analyzed 7,584 scientific publications associated with the use of machine learning in energy-related applications over a 20-year period (2,000 to 2,020), taking as a reference the methodology proposed by Kitchenman in its proposed methodology for the preparation of review articles. As a result of this analysis, the following significant findings were found:

- The trend in scientific publications in this area is increasing exponentially, with a greater number of scientific publications by March 2020 than for all of 2015.
- The greatest number of scientific publications are associated with articles published in indexed scientific journals, as well as procedures generated from conferences, representing 95% of total publications.
- The strong scientific collaborations associated with this topic are limited compared to the number of scientific productions found. Listing 68 authors out of 22,394 who have collaborated in this area of knowledge.
- The collaborative environment in this field, in geopolitical terms, is open. Finding an appropriate collaborative environment between countries led by the United States and China, which are at the top of the world economy today, according to the GDP of each country.
- Collaboration between countries has been increasing in parallel with the number of publications, with a maximum of 5 authors from different countries producing the same scientific article by 2019.
- As far as the algorithms used are concerned, the development of neural networks for problem-solving has been the most used resource in this area, followed by random forest algorithms.
- Regarding the software/programming languages used for the developments reported through the publications, most of them have opted for the use of Matlab®, led by China. However, there is a significant increase, over the programming language R, to the use of Python in this type of analysis, being led, with a greater contribution of the number of publications, by the USA (Fig. 11).

Based on the analyses made and reported in this article, the following points are suggested for further studies:

- Conduct a geographical study of the most researched sub-themes by country. This is because the topic covered in this article with respect to the use of machine learning applied to energy has been addressed in a general way. This would give a greater understanding of research trends in all countries, allowing the socio-economic status of countries to be related to related research fronts in each country.

- To carry out a scientific map analysis associated with the programs used for the development of these publications, this in order to establish the adaptability of the scientific scenario, in view of the use of new programming languages/software for the analyses associated with this type of technologies.
- Perform analysis to establish trends in scientific journal publication. Seeking to establish whether geopolitical preferences exist in the acceptance of articles related to this topic.

REFERENCES

[1] J. Rincon-Patino, G. Ramirez-Gonzalez, and J. C. Corrales, "Exploring machine learning: A bibliometric general approach using SciMAT," *F1000Research*, vol. 7, no. May, p. 1210, 2018, doi: 10.12688/f1000research.15620.1.

[2] A. Nielsen, *Practical Time Series Analysis_ Prediction with Statistics and Machine Learning*, O'Reilly M. 2019.

[3] N. J. Nilsson, "INTRODUCTION TO MACHINE LEARNING AN EARLY DRAFT OF A PROPOSED TEXTBOOK Department of Computer Science," *Mach. Learn.*, vol. 56, no. 2, pp. 387–99, 2005, doi: 10.1016/j.neuroimage.2010.11.004.

[4] D. Cica, B. Sredanovic, S. Tesic, and D. Kramar, "Predictive modeling of turning operations under different cooling/lubricating conditions for sustainable manufacturing with machine learning techniques," *Appl. Comput. Informatics*, 2020, doi: 10.1016/j.aci.2020.02.001.

[5] M. Zekić-Sušac, S. Mitrović, and A. Has, "Machine learning based system for managing energy efficiency of public sector as an approach towards smart cities," *Int. J. Inf. Manage.*, no. March 2019, p. 102074, 2020, doi: 10.1016/j.ijinfomgt.2020.102074.

[6] B. S. dos Santos, M. T. A. Steiner, A. T. Fenerich, and R. H. P. Lima, "Data mining and machine learning techniques applied to public health problems: A bibliometric analysis from 2009 to 2018," *Comput. Ind. Eng.*, vol. 138, no. April, p. 106120, 2019, doi: 10.1016/j.cie.2019.106120.

[7] F. Y. Y. Ling and W. T. Ng, "Boosting performance of road infrastructure," *Built Environ. Proj. Asset Manag.*, vol. 1, no. 2, pp. 211–225, 2011, doi: http://dx.doi.org/10.1108/20441241111180442.

[8] R. F. Binyousef, A. M. Al-Gahmi, Z. R. Khan, and E. Rawah, "Procedures for Performing Systematic Reviews Barbara Kitchenham," *Ann. Saudi Med.*, vol. 37, no. 1, pp. 79–83, 2004, doi: 10.5144/0256-4947.2017.79.

[9] D. A. C. Narciso and F. G. Martins, "Application of machine learning tools for energy efficiency in industry: A review," *Energy Reports*, vol. 6, pp. 1181–1199, 2020, doi: 10.1016/j.egyr.2020.04.035.

[10] N. Apergis and J. E. Payne, "Renewable energy consumption and economic growth: Evidence from a panel of OECD countries," *Energy Policy*, vol. 38, no. 1, pp. 656–660, 2010, doi: 10.1016/j.enpol.2009.09.002.

[11] El Congreso de la República de Colombia, *LEY 1715 DE 2014 Por medio de la cual se regula la integración de las energías renovables no convencionales al Sistema Energético Nacional.* 2014.

[12] A. Sood and G. J. Tellis, "The S-Curve of Technological Evolution: Marketing Law or Self-Fulfilling Prophecy?," *SSRN Electron. J.*, no. April, 2011, doi: 10.2139/ssrn.981532.

[13] "GDP of Countries." Accessed: May 14, 2020. [Online]. Available: https://www.imf.org/external/pubs/ft/weo/2019/01/weodata/weorept.aspx?pr.x=80&pr.y=9&sy=2019&ey=2019&scsm=1&ssd=1&sort=country&ds=.&br=1&c=512.