Application analysis of machine learning in fault diagnosis: A bibliometric analysis

Wenyan Tu¹, Ji Fang¹, Yihua He¹ and Jin Huang¹
¹Huazhong Univ Sci & Technol, Lib, Wuhan, Hubei, Peoples R China

Corresponding author and e-mail: Jin Huang, hjhust@hust.edu.cn

Abstract. This article to analyze the application of machine learning in fault diagnosis by using bibliometrics, co-citation network analysis and cluster analysis methods. The analysis found that in the application research of machine learning in fault diagnosis, China has the largest number of published papers and cited papers, among which IEEE access is the most published journal, MECHANICAL SYSTEMS AND SIGNAL PROCESSING is the most cited journal, followed by the United States and India. Using document co-citation relationship analysis, a set of key documents in this field was identified. Currently, the important algorithms fault diagnosis include support vector data description method, transfer learning algorithm, convolutional neural network algorithm, natural inspired optimization algorithm, bayesian network, wavelet packet decomposition algorithm, fuzzy logic algorithm and so on. All of them are clustered according to keywords, and their application fields include acoustic emission and fault diagnosis of doubly-fed induction generators.

1. Introduction
The earliest definition of fault is the deviation of observed variables or calculated parameters from the acceptable range in a process [1-2]. Fault diagnosis refers to the process of finding the fault of equipment or system. Fault diagnosis is developed due to the need of establishing monitoring system [3], which has gone through three stages [3-4]. The first stage is fault diagnosis based on the experience of experts and simple instruments; the second stage is the diagnosis technology based on signal analysis and modeling by means of sensing and detection technology; the second stage is fault diagnosis based on signal analysis and modeling; The third stage: with the development of computer, artificial intelligence and information processing technology, fault diagnosis technology has entered a new development stage—intelligent fault diagnosis stage based on machine learning [4-8].

This article uses bibliometrics, co-citation network analysis, and cluster analysis methods to analyze the overall status of the application of machine learning in fault diagnosis, combined with visual analysis methods to try to find the most concerned hot issues and the research field with the most development potential, which will guide further in-depth research in the future.

2. Data sources and methods

2.1. Data sources
Based on the core collection data of Web of Science and the Incites database, this article uses the subject search method to construct a search strategy of TS=("machine learning" and "fault diagnosis"). The
publication time of the literature is as of 2019, regardless of language, and the type of literature for article, review and proceedings paper, 752 academic papers were finally retrieved.

2.2. Analysis methods and tools
This article uses bibliometric analysis methods to analyze and explore the 752 papers retrieved from multiple perspectives such as the country where the paper was published, the influence of the journal, the co-citation of the literature, and the distribution of subject areas.

When analyzing the co-citation network of documents, use CiteSpace visualization software for visual analysis and display, revealing important documents and key documents in this research field. Through keyword clustering, understand and analyze specific research content and main research fields.

3. Bibliometric analysis

3.1. Basic overview of national or regional papers
Through the use of Incites data analysis platform to analyze the 752 papers that have been retrieved, it can be seen that in the application analysis of machine learning in fault diagnosis, China has published the most papers, followed by the United States and India. China is also cited the most frequently. In terms of institutions, China has the largest number of institutions in the top 100 cited by frequency, with 31 institutions, but the number of institutions in the top 10 and 50 cited by frequency is less than that of the United States. In addition to China and the United States, India is the country with the most cited institutions in the top 100, but there is still a big gap between China and the United States.

Table 1. Ranking of the top ten countries and institutions in the number of published papers.

| Name    | Total number of papers | Total citations | Top 10 institutions by citation frequency | Top 50 institutions by citation frequency | Top 100 institutions by citation frequency |
|---------|------------------------|-----------------|------------------------------------------|-------------------------------------------|--------------------------------------------|
| China   | 300                    | 2580            | 3                                        | 14                                        | 31                                         |
| America | 115                    | 2343            | 5                                        | 17                                        | 28                                         |
| India   | 70                     | 980             | 2                                        | 6                                         | 8                                          |
| Canada  | 40                     | 236             | 0                                        | 0                                         | 2                                          |
| England | 37                     | 537             | 0                                        | 4                                         | 6                                          |
| Korea   | 29                     | 156             | 0                                        | 0                                         | 2                                          |
| France  | 26                     | 299             | 0                                        | 2                                         | 4                                          |
| Spain   | 21                     | 385             | 0                                        | 3                                         | 4                                          |
| Ecuador | 15                     | 204             | 0                                        | 0                                         | 0                                          |
| Australia | 15                   | 96              | 0                                        | 0                                         | 2                                          |

3.2. Analysis of high level journals
According to the analysis of the number of publications and citation frequency of the source journals of 752 papers, there are five journals that rank top ten in both publication volume and citation frequency, and all the partitions in the JCR database are in region Q1. Among them, the journal with the largest publication volume is IEEE ACCESS, and the journal with the highest citation frequency is MECHANICAL SYSTEMS AND SIGNAL PROCESSING. IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS has the highest impact factor.

Table 2. Top ten journals with published papers and cited frequency.

| Journal                  | Papers | Freq | Periodical Division | Journal impact factors |
|--------------------------|--------|------|---------------------|------------------------|
| IEEE ACCESS              | 26     | 206  | Q1                  | 4.098                  |
| Expert Systems With Applications | 15     | 560  | Q1                  | 4.292                  |
3.3. Co-citation analysis of literature

Co-citation analysis means that two articles appear together in the bibliography of the third cited article, and the two articles form a co-citation relationship. Usually, two or more papers with co-citation relationship are similar in research methods or research topics, and a group of key papers in a certain topic field can be identified through co-citation analysis of the literature. By using CiteSpace to analyze the co-citation of 752 literatures, figure 2-2 is obtained. From the figure, the larger the node is, the higher the citation frequency is, and the connection between nodes indicates the co-citation relationship between the two literatures. The nodes in red represent emergence citations, which are nodes guiding the sudden rise or fall of dosage, and represent possible changes in the research direction of this field. From the perspective of analysis, the three literatures ranking top in citation frequency are shown in Table 2-3, and the citation frequency of these three literatures is increasing year by year, which may become important and key achievements in this field.

Table 3. Top three papers cited by frequency.

| Author            | Title                                                                 | Source        | Freq | Year |
|-------------------|-----------------------------------------------------------------------|---------------|------|------|
| Jia F             | Deep neural networks: A promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data | Mech Syst Signal Pr | 53   | 2016 |
| LeCun Y           | Deep learning                                                        | Nature        | 42   | 2015 |
| Janssens, Olivier | Convolutional Neural Network Based Fault Detection for Rotating Machinery | J Sound Vib   | 27   | 2016 |

Figure 1. Chronological distribution of the cited frequency of the top three papers.

Through burst analysis of literature, it can be concluded that the articles written by Sugumaran V, Widodo A and Witten I H have been cited with relatively high burst strength. Further analysis of the change of citation frequency shows that these articles attracted a high level of attention in the early stage,
but with the change of time, the attention declined, which also indicates that the research focus or direction has changed.

**Table 4.** The top three papers in mutant intensity.

| Author        | Title                                                                 | Source        | Freq | Burst | Year |
|---------------|-----------------------------------------------------------------------|---------------|------|-------|------|
| Sugumaran V   | Feature selection using Decision Tree and classification through Proximal Support Vector Machine for fault diagnostics of roller bearing | MECH SYST SIGNAL PR | 15   | 7.67  | 2007 |
| Widodo A      | Support vector machine in machine condition monitoring and fault diagnosis | MECH SYST SIGNAL PR | 13   | 6.81  | 2007 |
| Witten I H    | Data mining practica                                                   | DATA MINING PRACTICA | 7    | 4.56  | 2005 |

**Figure 2.** Document Co-Citation Network.

### 3.4. Analysis of discipline field

According to the discipline classification of Web of Science, an analysis of 752 papers related to the application of machine learning in fault diagnosis showed that the disciplines were mainly distributed in the fields of engineering and computer science. Specifically, it had been cited more frequently in refined fields such as artificial intelligence, automation and control systems, information systems, mechanical engineering, interdisciplinary applications, instruments & instrumentation. Among them, the engineering field had the highest frequency of appearance, and the artificial intelligence field had the highest betweenness centrality.
Table 5. Distribution of disciplinary.

| Subject Categories                              | Frequency | Betweenness Centrality |
|------------------------------------------------|-----------|------------------------|
| Engineering                                    | 449       | 0.69                   |
| Computer Science                               | 290       | 0                      |
| Engineering, Electrical & Electronic           | 242       | 0.41                   |
| Computer Science, Artificial Intelligence      | 149       | 1.03                   |
| Automation & Control Systems                   | 87        | 0.35                   |
| Engineering, Multidisciplinary                 | 71        | 0                      |
| Computer Science, Information Systems          | 70        | 0.18                   |
| Engineering, Mechanical                        | 62        | 0.69                   |
| Computer Science, Interdisciplinary Applications| 58        | 0                      |
| Instruments & Instrumentation                  | 54        | 0.06                   |

3.5. Cluster analysis

Keywords are an important manifestation of the research content of an article. By analyzing keywords, the research direction and research hotspots in this field can be roughly understood. Further clustering analysis was carried out on the keywords, and the clustering results were evaluated by two indexes: the clustering module value (Q value) and the clustering average contour value (S value). It is generally believed that Q>0.3 means that the cluster structure is significant, S>0.5 clustering is reasonable, and S>0.7 means that the clustering is convincing. Through the cluster analysis of the keywords of the 752 articles, 9 clusters were obtained. Further analysis of clustering can divide the main research content into two categories, one was the algorithm research of machine learning; the other was the specific application of machine learning in different types of faults.

3.5.1. Algorithm research in the field of fault diagnosis. From the keyword clustering table, there were many researches on the algorithm in fault diagnosis. Among them, clustering 0 (support vector data description method), clustering 1 (transfer learning algorithm), clustering 3 (convolution neural network algorithm), clustering 4 (natural inspired optimization algorithm), clustering 6 (Bayesian network), clustering 7 (wavelet packet decomposition algorithm) and clustering 8 (fuzzy logic algorithm) were related to the algorithm.

3.5.2. Application research in the field of fault diagnosis. In the field of practical application, the research on fault diagnosis mainly involved acoustic emission and doubly-fed induction generators.

Cluster 2: The application of artificial neural network in acoustic emission. Acoustic emission is a common physical phenomenon. Acoustic emission technology is a dynamic non-destructive testing method that judges the degree of internal damage of the structure based on the stress wave emitted from the material or structure. Some progress has been made in the use of artificial neural networks for acoustic emission signal processing. Because the artificial neural network has the functions of self-organization, self-adaptive and self-learning, and the network has high robustness, it can solve the problem of noise interference in acoustic emission detection, and can accurately judge the activity of acoustic emission source [9].

Cluster 5: Fault detection of doubly fed induction generator. Doubly fed induction motor is mainly used in wind power generation system. Doubly fed induction generator (DFIG) is the main type of wind power generation. With the development of wind power generation, the research on condition monitoring and fault diagnosis of wind turbine is becoming more and more important. At present, early fault intelligent diagnosis method based on multi-sensor information fusion and wavelet analysis method are important analysis methods for fault detection of doubly fed induction generator [10].
### Table 6. Keyword cluster.

| Cluster id | Size | Silhouette | Mean year | Keywords |
|------------|------|------------|-----------|----------|
| 0          | 30   | 0.695      | 2011      | functional failure (9.42, 0.005); machine learning (6.95, 0.01); svdd (6.26, 0.05); turbofan engine (6.77, 0.01); transfer learning (6.77, 0.01); statistical features (4.26, 0.05); pattern recognition (7.76, 0.01); acoustic emission (7.76, 0.01); artificial neural network (4.25, 0.05); cnn (14.48, 0.001); deep learning (12.13, 0.001); k-means (9.6, 0.005); big data (10.2, 0.005); fault injection (5.07, 0.05); nature-inspired optimization algorithms (5.07, 0.05); doubly-fed induction generator (dfig) (9.08, 0.005); deep classifier (9.08, 0.005); rotor current (9.08, 0.005); fault detection (9.08, 0.005); logical analysis of data (6.92, 0.01); bayesian networks (6.92, 0.01); tennessee eastman process (6.92, 0.01); machine learning (5.05, 0.05); wavelet packet decomposition (4.52, 0.05); asymmetrical fault (4.52, 0.05); fuzzy logic (6.59, 0.05); model-based reasoning (6.49, 0.05); artificial intelligence (6.49, 0.05); |

4. Conclusions

Machine learning is one of the most intelligent and cutting-edge research fields of artificial intelligence. The core of intelligent fault diagnosis is to effectively acquire, transmit, process, regenerate and utilize diagnostic information, so that it has the ability to accurately identify and predict the status of the diagnostic object in a given environment [11]. At present, in the research of fault diagnosis, China is not only the country with the most published papers, but also the country with the most cited frequency. Based on the analysis of document co-citation relationship, a group of key documents in this field are identified. It is concluded that the papers published by Jia F, LeCun Y, Janssens, Olivier have received more citations. From the keyword clustering analysis, the important fault diagnosis algorithms include support vector data description method, transfer learning algorithm, convolution neural network algorithm, natural inspired optimization algorithm, bayesian network, wavelet packet decomposition algorithm, fuzzy logic algorithm and so on. Their research fields consist of acoustic emission and fault diagnosis of doubly fed induction generator.

With the development of artificial intelligence technology and the improvement of mechanical equipment automation and intelligent level in China, machine learning algorithm will play a greater role in the improvement of power system, mechanical equipment and performance.

References

[1] HIMMELBLAU D M. Fault Detection and Diagnosis in Chemical and Petrochemical Processes [M]. New York: Elsevier Science Ltd, 1978:45–70.
[2] Zhou Donghua, Xi Yugeng. Fault detection and diagnosis technology [J]. Journal of Control
theory and Applications, 1991, 8 (1):1-7.

[3] Li Wei. Advance of Intelligent Fault Diagnosis for Complex System and Its Present Situation [J]. Computer Simulation, 2004, 21(10):4-7.

[4] Zhou Donghua, Liu Yang, He Xiao. Review on Fault Diagnosis Techniques for Closed-loop Systems [J]. Acta Automatica Sinica, 2013, 39(11):1933-1943.

[5] Li Chunlin, Xiong Jianbin, Su Naiquan, Zhang Qinghua, Liang Qiong, Ye Baoyu. Application Review of Deep Learning in Fault Diagnosis [J]. Machine Tool & Hydraulics, 2020, 48(13):174-184.

[6] Wang Xiaofeng, Mao Deqiang, Feng Shangcong. Review on modern fault diagnosis technologies [J]. China Measurement & Test, 2013, 39(06):93-98.

[7] Zhou Donghua, Shi Jiantao, He Xiao. Review of Intermittent Fault Diagnosis Techniques for Dynamic Systems [J]. Acta Automatica Sinica, 2014, 40(02):161-171.

[8] Wen Chenlin, Lv Feiya. Review on Deep Learning Based Fault Diagnosis [J]. Journal of Electronics & Information Technology, 2020, 42(01):234-248.

[9] Li Jialin, Dong Yunchao, Ma Yukuan. Pattern Recognition of the Characteristics of Acoustic Emission Source Using Neural Network [J]. Nondestructive Testing, 2001(06):231-233+272

[10] Xiao Yali, Fang Ruiming, Li Wenpeng. Overview of the Condition Monitoring Method for Doubly-Fed Induction Wind Power Generation Systems [J]. Power System and Clean Energy, 2012, 28(03):67-73+78.

[11] Xiao Xiao. Machine learning in intelligent fault diagnosis [D]. Chongqing University, 2002.