Vol. 3, No. 1, 1-13. DOI: 10.47909/ijsmc.36

Pro-Metrics

ORIGINAL RESEARCH

# Comparing research trends through author-provided keywords with machine extracted terms: A ML algorithm approach using publications data on neurological disorders

**Priya Tiwari[1], Saloni Chaudhary[1], Debasis Majhi[1], Bhaskar Mukherjee[2]**

[1] Department of Library & Information Science, Banaras Hindu University, Varanasi-221005, India.
[2] Department of Library & Information Science, Banaras Hindu University, Varanasi-221005, India.
**Email:** mukherjee.bhaskar@gmail.com
Corresponding author.

**ABSTRACT**

**Objective.** This study aimed to identify the primary research areas, countries, and organizational involvement in publications on neurological disorders through an analysis of human-assigned keywords. These results were then compared with unsupervised and machine-algorithm-based extracted terms from the title and abstract of the publications to gain knowledge about deficiencies of both techniques. This has enabled us to understand how far machine-derived terms through titles and abstracts can be a substitute for human-assigned keywords of scientific research articles.

**Design/Methodology/Approach.** While significant research areas on neurological disorders were identified from the author-provided keywords of downloaded publications of *Web of Science* and *PubMed,* these results were compared by the terms extracted from titles and abstracts through unsupervised based models like VOSviewer and machine-algorithm-based techniques like YAKE and CounterVectorizer.

**Results/Discussion.** We observed that the post-covid-19 era witnessed more research on various neurological disorders, but authors still chose more generic terms in the keyword list than specific ones. The unsupervised extraction tool, like VOSviewer, identified many other extraneous and insignificant terms along with significant ones. However, our self-developed machine learning algorithm using CountVectorizer and YAKE provided precise results subject to adding more stop-words in the dictionary of the stop-word list of the NLTK tool kit.

**Conclusions.** We observed that although author provided keywords play a vital role as they are assigned in a broader sense by the author to increase readability, these concept terms lacked specificity for in-depth analysis. We suggested that the ML algorithm being more compatible with unstructured data was a valid alternative to the author-generated keywords for more accurate results.

**Received:** 21-01-2023. **Accepted:** 08-05-2023

**Editor:** Carlos Luis González-Valiente

**How to cite:** Tiwari, P., Chaudhary, S., Majhi, D., & Mukherjee, B. (2023). Comparing research trends through author-provided keywords with machine extracted terms: A ML algorithm approach using publications data on neurological disorders. *Iberoamerican Journal of Science Measurement and Communication; 3*(1), 1-13. DOI: 10.47909/ijsmc.36

**Originality/Value.** To our knowledge, this is the first-ever study that compared the results of author-provided keywords with machine-extracted terms with real datasets, which may be an essential lead in the machine learning domain. Replicating these techniques with large datasets from different fields may be a valuable knowledge resource for experts and stakeholders.

**Keywords:** machine learning algorithm; Covid-19-neurological disorders; neurological disorders; automatic extraction; title extraction-ml algorithm.

## 1. INTRODUCTION

Detecting research trends has become more challenging due to the continuous increase of scientific literature in each field. Trend analysis helps policymakers and researchers to map the intellectual structure of the discipline, understand the dynamics in the discipline (Duvvuru *et al.,* 2013), and also assist in discovering emerging topics in science (Small, Boyack, & Klavans, 2014). The title, abstract and author-provided keywords play a significant role in determining research trends. While the title conveys the primary meaning, the abstract summarizes the whole thought. On the other hand, the author-provided keywords are the important terms of the research that the author considers to be the most relevant to their research (Lu *et al.,* 2019), reflecting the personal view on the subject presented in a document (Kevork & Vrechopoulos, 2009). In the present digital-dominant era, careful elaboration of titles, abstracts, and keywords is fundamental for the text to be searched by search engines. When author-assigned keywords are chosen methodically, they may improve the effectiveness of keyword searches (Maurer, McCutcheon, & Schwing, 2011), Although online databases like *Web of Science,* and *Scopus,* along with these three fields, also added indexer keywords in their dataset for further recovery of publications quickly, mostly these words are broadly descriptive (Zhang *et al.,* 2016), therefore may not helpful in understanding the dynamic of discipline and discover emerging topics.

To comprehend the dynamics of research trends, reading the title and abstract individually one by one and extracting its trend is a grim task. With the continuous growth in scientific literature and the evolution of new technologies, trend analysis techniques have transformed over time. Keyword-based trend analysis through network-based methods, such as keyword co-occurrence, has been proven effective in identifying research trends and hotspots (Cheng *et al.,* 2020). However, it is generally restricted to the simple description of the network (Huang & Zhao, 2019). In contrast, machine learning approaches like text mining and natural language processing can be employed in trend analysis (Sarker, 2021). The data avalanche and challenges associated with analyzing those complex datasets within the allotted time frame are likely factors in why researchers are increasingly resorting to machine learning rather than conventional trend analysis techniques.

Machine learning includes the use of machine learning models and algorithms. Algorithms for machine learning are programmes that analyze datasets for patterns and laws. There are various machine learning algorithms: supervised, unsupervised, semi-supervised and reinforcement learning. Although machine learning algorithms have some benefits, supervised learning approaches need a lengthy training procedure and massive collections of human-annotated documents to fully understand a language (Uddin *et al.,* 2019). In contrast, their plug-and-play capabilities allow broad unsupervised approaches to be readily applied to a document in various languages or domains with little effort (Zamri *et al.,* 2022).

The use of unsupervised approaches has grown in favour as an alternative to the tiresome process of manually labelling big collections of documents(Quan, Wang & Ren, 2014). The standard procedure in unsupervised techniques is called TF.IDF, which compares the frequency of a term in a document to that of a term in a bigger collection. But TF.IDF needs a sizable corpus, which might not always be accessible (Papagiannopoulou & Tsoumakas, 2020). To identify significant terms from the dataset, the current study uses more recently developed, unsupervised automatic keyword extraction algorithms Yet Another Keyword Extractor (YAKE) and CountVectorizer.

Besides, for clustering unlabelled datasets, the VOSviewer tool can be utilized to visualize term relations between clusters accompanied by a term map that displays the relation between the most important terms in publications, their corresponding countries, organization etc. and co-occurrence relations between these terms (Van Eck & Waltman, 2019)

YAKE is a simple, unsupervised automatic keyword extraction technique that finds the most relevant keywords in the text using statistical text features retrieved from specific documents. This system is independent of dictionaries, text size, domain, language, or training in a particular set of documents(Campos *et al.,* 2020). It specifies a collection of five features that capture keyword characteristics. These features are heuristically merged to give each term a single score. The keyword's importance increases with a lower score.

Additionally, a great utility offered by the Python scikit-learn module is CountVectorizer. It converts a given text into a vector based on the number of times (count) that each word appears across the full text. This is useful when we have several texts and want to turn each word into a vector for further use in text analysis. Thus, this paper identifies limitations in comparing machine-learning-based algorithms utilizing YAKE and CountVectorizer and human-assigned keyword methods utilizing VOSviewer.

The COVID-19 pandemic, a global health crisis that emerged in late 2019, has profoundly impacted people's lives worldwide, causing significant health, economic, and social disruptions. The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is the cause of Covid-19, a unique illness with a long history. Covid-19 has been found to have serious mental health repercussions despite being a respiratory illness because social isolation triggers psychiatric symptoms linked to neuropsychiatric and neurological problems (Roy *et al.,* 2021). Neurological dysfunction harms the brain's functioning regardless of age, and they have the potential to seriously impair the nervous system and strike human society with fatalities (Wu *et al.,* 2020). This cognitive blockage further disrupts the quality of life and prolificity(Graham *et al.,* 2021; Rothstein, 2023). Since Coronavirus has had a long history, have any new sub-field of research emerged during covid-19 or can any significant volume of changes in research-related neurological disorders in the post-Covid-19 be observed? Considering the intricacy of Covid-19 pathogenesis, it becomes essential to study the disorders caused due to neurological imbalance. Secondly, as it has been noted that author keywords may be subject to personal biases, failing to show the interdisciplinarity of their publications, can we develop any machine learning algorithm as an alternative to human-assigned keywords of scientific research articles?

## 2. OBJECTIVES

- To identify predominant research areas, countries, and organizational involvement in neurological disorders by analyzing human-assigned tags in the publications on neurological disorders.
- To identify the limitations or deficiencies derived from the comparison between machine-learning algorithm-based methods and the human-assigned methods using title and abstract keywords of publications.

## 3. METHODOLOGY

### 3.1. Data source and search strategy

Before understanding the viability of title extraction using a machine learning algorithm, tracking the significant areas of a research domain was essential. Also, to know whether these terms are consistent over time, it was necessary to check the consistency of terms by comparing results of at least two-time windows. For that, initially, we executed a Boolean search in the advanced search interface of the Web of Science (WoS) (search field-topic field) and PubMed (search field-all fields) databases during the first quarter of 2023 to know the research of the domain. A search string consisting of terms *neurological** AND *covid-19, neurological** OR *SARS-Cov-2* has been searched in the WoS database under the topic field. The search was also done under the author keyword for *disorder* AND *neurological** in the topic field in the PubMed database. The results from both databases were downloaded, merged in a single CSV file for author's convenience, and removed duplicate records using the MS-EXCEL command-line from the corpus. In total, 13281

unique articles were collected initially from the two databases for study. These articles have a field tag 'DE', representing the author keyword in WoS.

Similarly, the field author keyword of the PubMed database also contains author-chosen keywords of the title. Both databases' datasets have been imported into a Python-based program to understand the frequently occurring disorders. Each disorder has been fetched individually from the prominently identified 25 disorders into WoS and PubMed again for noting publications that appeared before 2020.

### 3.2. Data processing

In the next step, the title and abstract fields of identified publications have been imported to VOSviewer to extract the significant terms. A threshold of 10 terms has been fixed to identify the top words that occur at least 10 times. Further, the same dataset was incorporated into our own-developed algorithm in Python. Pandas library and Natural Language Toolkit (NLTK) library were run in the algorithms. The uploaded records were first tokenized in which each word, letter, and punctuation mark is considered a token. All tokens were converted to lowercase, and then NLTK enabled stop-words to remove insignificant words like "the", "if" "but", "a", or "an", and so on, drawing no information and taking a long time to process. After that, a new stop-word library was developed to eliminate other non-significant words with little or unusual significance in the text corpus. For removing double spaces, special characters, numbers and punctuations, regular expressions (re) were used. Then, we run the *correct()* function from the TextBlob library to perform spelling corrections. Finally, WordNetLemmatizer was run to lemmatize words with singular, plurals etc.

### 3.3. Exploring research dimensions of the downloaded dataset

To conclude whether Coronavirus has significantly impacted different neurological disorders, the results were compared in two-time windows, i.e. Covid-19 period (i.e., 2020-2022) and pre Covid-19 period (until 2019). Simple descriptive statistics have been employed here. For identifying the country and organizational involvement in neurological research, VOSviewer was employed.

### 3.4. Comparison of different unsupervised technique

During this phase, the same file was uploaded to a self-developed machine-learning algorithm under Jupyter Notebook for applying CountVectorizer and YAKE. We used the scikit-learn machine learning library for the keywords extraction, and from the sklearn feature_extraction module, we used CountVectorizer for machine-extracted keywords using the ngram_range 2,3. As a result, we retrieved machine-extracted keywords from the title and abstract fields of the identified neurological disorders. Subsequently, yake KeywordExtractor was executed from the YAKE library using the max_ngram_size=3, deduplication_threshold=0.9 and extracting required keywords from the dataset. YAKE extracts keywords without a training set or external corpus, which is advantageous in cases where access to the training set is restricted. Lastly, the results obtained from both unsupervised techniques were compared to understand the trend of the subject of this study and to identify the limitations or deficiencies in both the methods, machine learning algorithm-based and human-assigned.

## 4. RESULTS

### 4.1. Publication trend

The results of 13281 unique publications show a major portion, i.e., 9701 publications appeared as 'articles', followed by 3351 as 'reviews'. All these publications appeared in 2107 journals, with more articles appearing in the *Journal of the Neurological Sciences* (365) followed by *Neurological Sciences* (335). So far, all these publications have received 162273 citations, with an average of 12.22 citations per article. There are almost 7 articles from which almost 12% of the total citations came, each of which received more than 1000 citations per article. Considering the year of publications, it was seen that almost 40% of publications appeared in 2021, which was double that of 2020 but a fall of 19% of publications observed between

2021 and 2022. On analyzing the abstracts of the published articles, it was understood that there were at least 479 publications where authors reported that patients have comorbidity with various other diseases along with Covid-19. The major age group of patients was adults, followed by children of the average age group 12. Adults were affected by neurological manifestations like stroke, Alzheimer's disease, and Parkinson's disease. Major neurological manifestations among children were ascending weakness with areflexia, diminished visual acuity, encephalopathy or weakness with plasma creatinine kinase (CK) elevation. Older patients exhibit seizures, stroke, flaccid paraparesis, corticospinal weakness, and even coma.

### 4.2. Predominant research areas

Table 1 shows the number of publications produced by medical professionals during the last three decades (as indexed in WoS and PubMed) on the top 25 neurological disorders. As shown in Table 1, stroke in the previous three decades consisting of ischemic, acute and haemorrhagic stroke, is the predominant disorder which sparked interest among medical scientists leading to the highest number of publications in the particular disorder. Following stroke, epilepsy, Alzheimer's disease, multiple sclerosis, and neuroinflammation were major disorders on which research was conducted significantly during 2020-22. Col 1 indicates that publications on various neurological disorders increased significantly during the Covid-19 period. To compare the covid-19 and pre covid period results, i.e., 2020-22 with pre-2019, respectively, we adjusted the total publications by average publications per year. It indicates that the average number of publications per year was comparatively less before 2019 than those from 2020-22.

| Disorders | NP until 2019 | APY until 2019 | NP during 2020-22 | APY during 2020-22 | Average Increase (per Year) |
|---|---|---|---|---|---|
| Stroke (Ischemic/Acute/haemoregic) | 1919 | 63.97 | 1866 | 622.00 | 558.03 |
| Epilepsy | 4019 | 133.97 | 1464 | 488.00 | 354.03 |
| Alzheimer's Disease | 2158 | 71.93 | 1087 | 362.33 | 290.40 |
| Multiple Sclerosis | 2715 | 108.60 | 973 | 324.33 | 215.73 |
| Neuroinflammation | 1043 | 35.97 | 931 | 310.33 | 274.36 |
| Blood-brain barrier | 754 | 30.16 | 709 | 236.33 | 206.17 |
| Encephalitis | 1522 | 66.17 | 686 | 228.67 | 162.50 |
| Encephalopathy | 2011 | 67.03 | 662 | 220.67 | 153.64 |
| Thrombosis | 676 | 29.39 | 611 | 203.67 | 174.28 |
| Parkinson's Disease | 2266 | 78.14 | 580 | 193.33 | 115.19 |
| Central Nervous System | 1409 | 54.19 | 526 | 175.33 | 121.14 |
| Depression | 1576 | 58.37 | 407 | 135.67 | 77.30 |
| Dementia | 1955 | 85.00 | 376 | 125.33 | 40.33 |
| Fatigue | 472 | 15.73 | 361 | 120.33 | 104.60 |
| Cognitive Impairment | 1116 | 42.92 | 355 | 118.33 | 75.41 |
| Neurological Anxiety | 602 | 20.76 | 306 | 102.00 | 81.24 |
| Headache/migrane | 1912 | 63.73 | 298 | 99.33 | 35.60 |
| Seizures | 1039 | 34.63 | 267 | 89.00 | 54.37 |
| Guillain-Barre Syndrome | 431 | 14.37 | 171 | 57.00 | 42.63 |
| Mental Health | 930 | 31.00 | 156 | 52.00 | 21.00 |
| Movement Disorders | 324 | 11.17 | 151 | 50.33 | 39.16 |
| Transverse Myelitis | 127 | 4.54 | 124 | 41.33 | 36.79 |
| Delirium | 157 | 7.14 | 119 | 39.67 | 32.53 |
| Cerebral Venous Sinus Thrombosis | 78 | 3.55 | 48 | 16.00 | 12.45 |
| Cytokine storm | 316 | 11.29 | 47 | 15.67 | 4.38 |

**Table 1.** Top twenty-five neurological disorder before and after Covid-19 based on human-assigned keywords. Note: NP=Number of Publication, APY=Average publication/year.

### 4.3. Productive countries and research active organizations

To understand which country and organization are dominantly associated with research on neurological disorders during the last 3 years, 2020-22, we have exported the WoS and PubMed condensed data into VOSviewer. Figures 1 a and b represent the country and organizational clusters. While importing the data into VOSviewer, through co-occurrence analysis, we observed that a total of 129 countries are involved in this research domain. Setting a threshold of five publications per country, 86 countries meet the threshold point. Among these 86 countries, the USA leads the table with 4227 publications, followed by Italy (1818), England (1357) and fourth rank held by India with 909 publications. Below India, Germany, and People's Republic of China is there. The USA, among the countries, followed by Italy, and England are the predominant countries which produce more research on these domains. It is interesting to note that nodes of the most advanced countries are nearer to each other, meaning the quantity and occurrence of collaboration between these countries is relatively high. India, among different nodes, also played a significant role by conducting research in cooperation with USA, England and Netherlands.

Regarding organizational clusters, it is clear that Harvard Medical School, Tehran University of Medical Sciences, and the University of Milan are some major organizations whose researchers actively research neurological disorders during 2020-22. The number of clusters in the organizational visualization shows that 7 clusters exist, indicated by 7 different colours. These 7 clusters are indicative of seven major organizations which are producing a significant portion of research being conducted in this field.

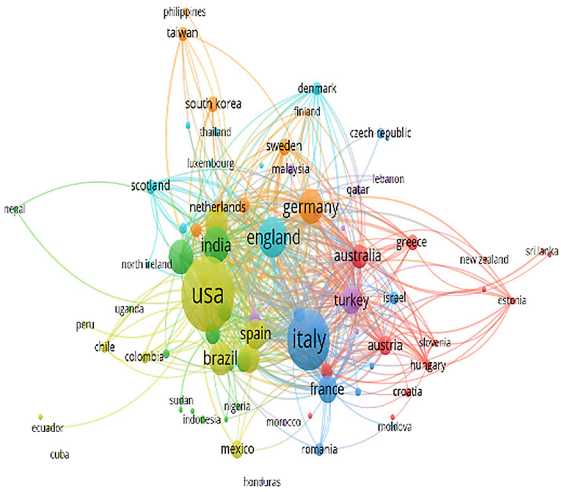


**Figure 1a.** Country Cluster.

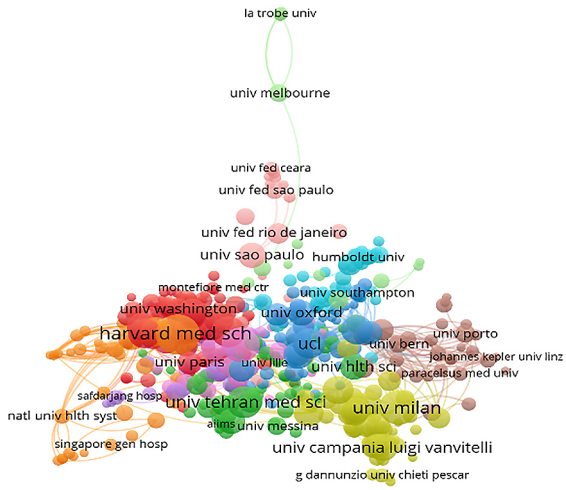


**Figure 1b.** Organizational Cluster.

### 4.4. Comparison of cluster-based terms with ML algorithm-based terms from the title of publications

Table 1 shows the result of predominant neurological disorders based on author-produced keywords (human-assigned). But it was observed that many downloaded results were without any keywords. This is one of the biggest challenges while analyzing research trends. To overcome such challenges, we have approached two machine-enabled processes to understand how far the automatic extraction process can overcome the shortcomings of the human-assigned process. We first utilized an unsupervised approach using VOSviewer and then NLP-based Python modules like CountVectorizer and YAKE. VOSviewer has excellent functionality to extract significant terms from the title and abstract. These techniques have been exploited to extract significant terms present in the form of neurological disorders that appeared in the downloaded articles' titles and abstracts. In Table 2, V_Terms is the term VOSviewer identified and extracted from the title. It is clear that among them, Encephalitis

is the most important term with the highest occurrence in titles of the dataset. But it was also seen that some irrelevant terms, such as *study, review, investigation,* and *infection,* which have little or less significance in the context of important keywords, are also extracted by VOSviewer while extracting terms from the title. This is one of the drawbacks we faced for such program-based extraction through VOSviewer.

| V_Terms | Freq. | RS | C_Terms | Freq. | RS |
|---|---|---|---|---|---|
| encephalitis | 223 | 0.7091 | ischemic stroke | 1078 | 0.002063 |
| thrombosis | 220 | 0.5464 | blood brain barrier | 687 | 0.002078 |
| traumatic brain injury | 126 | 0.4287 | cognitive impairment | 280 | 0.001052 |
| blood brain barrier | 115 | 0.5666 | neurological disorders | 274 | 0.002306 |
| guillain barre syndrome | 106 | 0.7504 | traumatic brain injury | 177 | 0.000277 |
| intracerebral hemorrhage | 88 | 0.5428 | cerebral venous sinus | 176 | 0.001479 |
| cerebral venous thrombosis | 83 | 0.7532 | cerebrospinal fluid | 152 | 0.002352 |
| parkinson | 64 | 0.4596 | venous sinus thrombosis | 143 | 0.000277 |
| autoimmune encephalitis | 63 | 0.6819 | guillain barre syndrome | 125 | 0.002478 |
| epileptic seizure | 62 | 0.857 | intracerebral hemorrhage | 108 | 0.000536 |
| posterior reversible encephalopathy syndrome | 60 | 0.8469 | epileptic seizure | 98 | 0.008874 |
| subarachnoid hemorrhage | 60 | 0.619 | mental health | 83 | 0.006321 |
| cerebral ischemia | 52 | 0.6547 | temporal lobe epilespy | 78 | 0.002504 |
| intravenous thrombolysis | 51 | 1.1955 | subarachnoid hemorrhage | 77 | 0.002214 |
| delirium | 45 | 0.8368 | mild cognitive impairment | 76 | 0.003711 |
| severe acute respiratory syndrome coronavirus | 41 | 0.4206 | encephalopathy syndrome | 75 | 0.010874 |
| mental health | 40 | 1.1776 | ischemia reperfusion injury | 75 | 0.002461 |
| acute ischemic stroke patient | 39 | 1.204 | nf kappa | 74 | 0.008163 |
| transverse myelitis | 38 | 0.9448 | ischemic encephalopathy | 71 | 0.009444 |
| chronic fatigue syndrome | 36 | 3.405 | posterior reversible encephalopathy | 67 | 0.000261 |
| Palsy | 36 | 1.2827 | transverse myelitis | 65 | 0.012042 |
| middle cerebral artery occlusion | 35 | 0.6071 | artery occlusion | 63 | 0.003451 |
| stroke patient | 35 | 1.0383 | spinal cord injury | 57 | 0.001367 |
| epilepsy surgery | 32 | 0.9745 | relapsing remitting sclerosis | 55 | 0.021474 |
| alzheimer | 32 | 0.4513 | parkinson | 46 | 0.006245 |

**Table 2.** Comparative frequency & relevance of Title terms extracted through VOSviewer and machine learning approach. Note: RS=Relevancy Score, Freq.=Frequency., V_Terms=VOSviewer extracted keywords, C_Terms=ML algorithm extracted keywords.

To overcome this problem, we have developed a Python-based program for machine algorithm-based extraction from title and abstract. In Table 2, C_Terms indicates the terms our program extracted from the title proper. Comparing the results of Table 2 with VOSviewer extracted terms and terms extracted through our program under Python, one can easily infer that the diseases identified in our process are more accurate and exhaustive than the VOSviewer removed terms. Additionally, in Table 2 we have also explained the relevancy score of these terms in the context of the title. Recently, VOSviewer has begun providing a feature to indicate the relevancy score of extracted terms. The relevancy score denotes co-occurrence with a limited connection of nodes, meaning if the relevancy score is high, the term has occurred in a more specific sense and vice-versa. For our machine extraction of significant terms, we exploit another module of NLP named YAKE to explore the relevancy score of significant terms. In the YAKE module, the lesser the score, the higher the term's relevancy. Therefore, in this context, it can be said that

some of the terms have fewer occurrences in the machine extraction process, but in all of the titles, they are assumed to be more relevant. Likewise, we represent the clusters of significant terms through VOSviewer which is shown in Figure 2. So many terms have been identified with at least 17 clusters. Still, *telemedicine, brain, Parkinson,* and *severe acute respiratory syndrome* have been identified as significant terms among all other clusters.

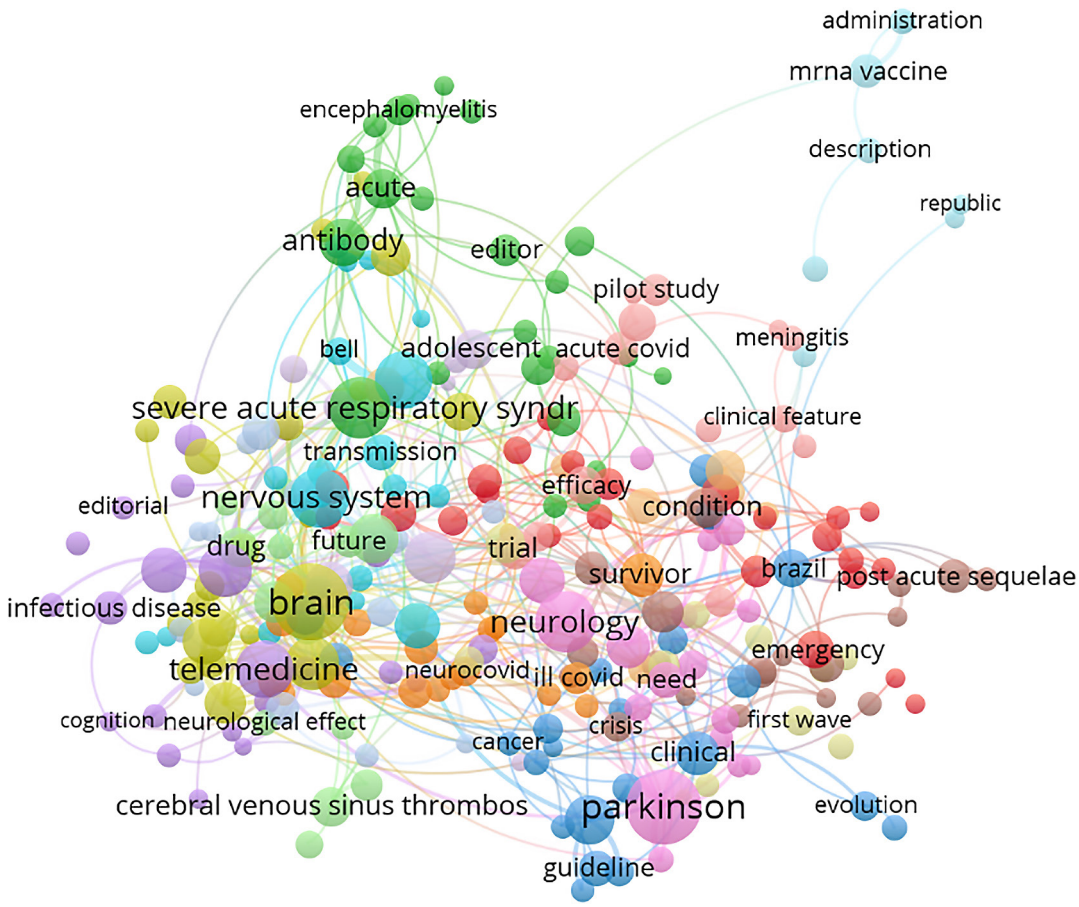


**Figure 2.** Clusters of Title extracted terms through VOSviewer.

### 4.5. Comparison of cluster-based terms with ML algorithm-based terms from abstracts of publications

Further, we have compared the significant terms that have been extracted from the abstracts of the titles. The results are shown in Table 3. The terms extracted through VOSviewer are broader than those extracted through our self-developed Python algorithm. Some of the lesser significant diseases have been identified in VOSviewer as substantial. However, the program we developed using Python has also produced more significant terms extracted from the abstracts. One of the possible reasons for differences in the extraction of substantial terms from abstracts using the two approaches in this study is that in our program, we have used stop words.

In the stop word list, we incorporated several insignificant words in the context of clusters of total results. This facility of NLP-based programs for machine extraction enables us to prepare more significant terms to display more relevant results to investigate the growth of a subject. Likewise, we present the clusters of significant terms through VOSviewer shown in Figure 3. As indicated in the graph, many terms have been identified with at least 7 clusters, but *telemedicine, brain, vaccine,* and *storm* have been identified as some of the significant terms among all other clusters.

| V_Terms | Freq. | RS | C_Terms | Freq. | RS |
|---|---|---|---|---|---|
| covid | 3028 | 0.8223 | ischemic stroke | 3600 | 0.000105 |
| neuroinflammation | 2023 | 0.917 | neurological disorders | 2299 | 0.00017 |
| encephalitis | 1083 | 0.6845 | central nervous cns | 1746 | 0.000366 |
| blood brain barrier | 925 | 0.7031 | neurological diseases | 1519 | 0.00017 |
| neurodegenerative disease | 892 | 0.4293 | cognitive impairment | 1279 | 0.000822 |
| acute ischemic stroke | 885 | 1.0089 | blood brain barrier | 1273 | 0.000939 |
| thrombosis | 853 | 0.9597 | Acute ischaemic stroke | 1042 | 0.001104 |
| brain injury | 609 | 0.5027 | scleriosis ms | 935 | 0.00105 |
| thrombolysis | 516 | 1.5862 | cerebrospinal fluid | 826 | 0.001091 |
| middle cerebral artery occlusion | 381 | 0.9999 | alzheimer | 742 | 0.001049 |
| nf kappa b | 377 | 1.2563 | neurodegenerative diseases | 709 | 0.001104 |
| stroke patient | 372 | 0.6633 | parkinson | 639 | 0.001078 |
| hypertension | 367 | 0.6923 | spinal cord injury | 594 | 0.002457 |
| hemorrhage | 361 | 0.424 | cerebral ischemia | 584 | 0.001426 |
| headach | 359 | 1.2704 | cerebral artery occlusion | 572 | 0.0025478 |
| movement disorder | 339 | 0.7464 | nf kappa | 553 | 0.000497 |
| cerebral ischemia | 330 | 0.6659 | guillain barre syndrome | 548 | 0.000247 |
| cytokine | 315 | 0.3943 | mental health | 523 | 0.000745 |
| brain edema | 299 | 0.7661 | epileptic seizure | 425 | 0.001124 |
| weakness | 294 | 1.0082 | brain imaging | 413 | 0.001073 |
| meningitis | 278 | 0.5551 | traumatic brain injury | 380 | 0.0024157 |
| depressive symptom | 270 | 1.4229 | anxiety/depression | 347 | 0.0011473 |
| guillain barre syndrome | 258 | 1.3513 | venous thrombosis | 321 | 0.0021478 |
| emotional dysregulation | 224 | 0.3723 | sinus thrombosis | 316 | 0.0023147 |
| mild cognitive impairment | 222 | 0.964 | brain edema | 315 | 0.0031472 |

**Table 3.** Comparative frequency of Abstract terms extracted through VOSviewer and machine learning approach. Note: RS=Relevancy Score, Freq.=Frequency, V_Terms=VOSviewer extracted keywords, C_Terms=ML algorithm extracted keywords.

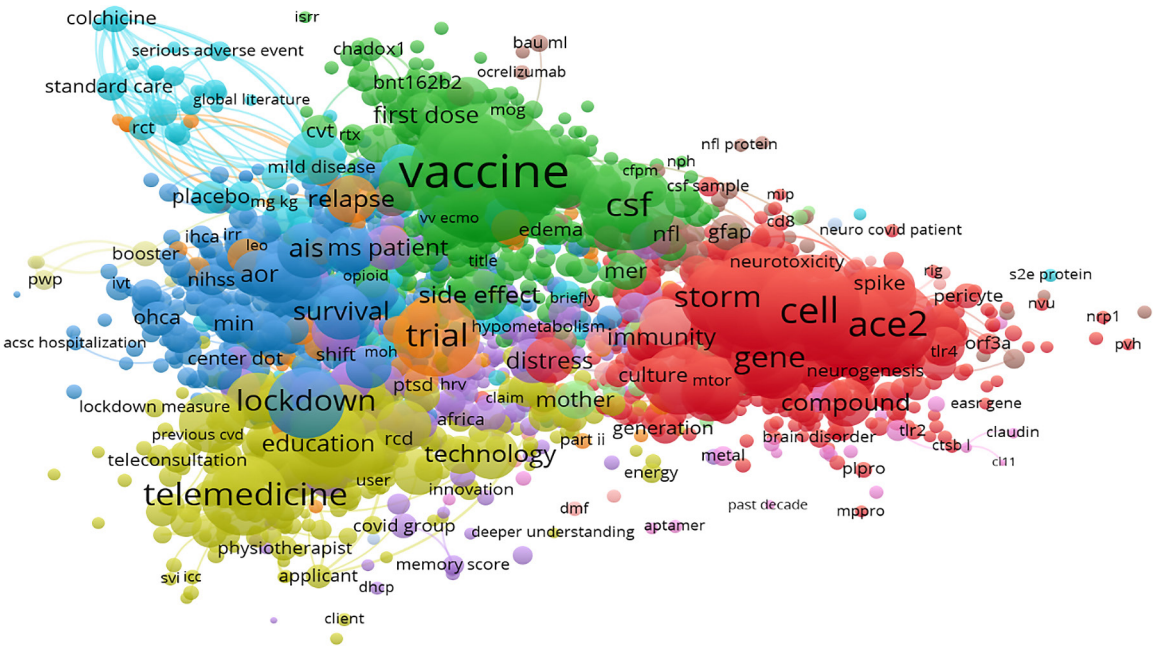


**Figure 3.** Abstract extracted terms through VOSviewer.

## 5. DISCUSSION

The primary purpose of this paper is to analyze and compare the results obtained through different keyword extraction techniques, i.e., author-assigned and Machine learning algorithm-based extraction. We looked at a field that has recently gained a lot of attention to show the results these procedures may provide: neurological diseases due to Covid-19. Before starting the extraction process immediately, we determined the publication trend in the specific domain and other dimensions like country and institutional involvement. As mentioned in Table 1, we identified the top 25 key terms relating to neurological disorders, which indicates the consistent appearance of these potential terms throughout the given timeframe, i.e., 2020-2022. During the analysis, we found that *covid-19* was the impelling cause for these identified diseases as we witnessed a hike in publication records post covid. As evident from the table above. *Stroke* is the top identified key term, including other types of strokes like *Ischaemic, Acute* and *Haemorrhagic stroke*. The feasible explanation for the highest occurrence of this broader concept (stroke) in terms of neurological disorder could be attributed to the choice made by the authors in assigning the keyword to the particular record. Authors voluntarily present their views regarding the subject matter by assigning keywords to the paper. These keywords play a vital role as the author assigns them in a broader sense to increase readability. However, this generalized representation of the concept clouds the identification of a specific concept during analysis. For example, thrombosis occurred frequently compared to cerebrothrombosis, which is more of a particular disorder. Another drawback of using these author-produced vital terms is the inflected nature of these critical terms. These terms differ from title to title and are even indexed under different names. For example- Guillain Barre Syndrome is mentioned and identified differently under Guillain Barrè syndrome, Guillaine - Barre Syndrome, GBS, and Guillain-Barre syndrome; another instance noticed was for the key term Seizure and Seizures, which were identified differently during analysis. In contrast, we can always refer to author-assigned keywords to get a birds-eye view of the publication. To understand the specific nature of the record, we need to investigate different variables like the abstract and title of the record for better representation.

As for the countries playing a pivotal role in neurological research worldwide, the USA is the central node in the visualized network, followed by other advanced countries like Italy and England. Similarly, India and other developing nations are also conducting specific research in the domain, despite being hit severely by the pandemic. To channel these research findings to a broader audience, a global response accompanied by improved coordination among scientists and international health professionals is indispensable to better prepare for the next unforeseen epidemic. As one of the instrumental institutions in uniting neuroscience research efforts globally, Harvard Medical School is the most central node, followed by Tehran University of Medical Sciences and the University of Milan. These institutions are working consistently towards excellence and inclusion in this research domain.

Further, we moved towards fulfilling the main objective, where we attempted to analyze the anomalies in author-assigned keywords as they failed to cater to the specific needs of the researchers. To overcome these issues, we used ML algorithms to extract key terms with a more detailed representation of the records. Once the extraction was done, we tried to evaluate and compare the results to test their compatibility. The superiority of machine-extracted keywords over author-produced keywords in the context of the specificity of key terms can be inferred from Table 2. In Column 1, VOSviewer extracted significant terms from the titles of the records that have been displayed. While the encephalitis term has the highest occurrence, many other extraneous and insignificant terms were found in the results extracted from the titles. Column 4 lists the significant terms extracted through our self-developed ML algorithm. The most important term extracted through our program is Ischemic Stroke, accompanied by other concrete terms such as blood-brain barrier and cognitive impairment. It is intriguing to observe that the absence of stop word functionality in VOSviewer is the root cause behind the inclusion of insignificant terms in the derived clusters. Moreover, handling and visualization of large datasets is a

tedious task. Therefore, our self-developed ML algorithm yielded a relatively higher occurrence of compound terms while VOSviewer extracted significant terms that included most root words.

Following a similar approach, we analyzed the comparative frequency of abstract terms in Table 3. Deploying both VOSviewer and machine learning techniques, we extracted the key terms from the abstract separately. We found that VOSviewer identified key terms from the abstract, like covid and neuroinflammation, alongside significant terms like encephalitis, blood brain barrier, acute ischemic stroke etc. In addition to that, it also extracted insignificant connections within the text, like antibodies and telemedicine. Whereas machine extracted, key terms were more specific, indicating that we used a stop word list to discard the unnecessary and redundant words from the text to get more explicit results. While present-day researchers utilize readily available NLTK Toolkit provided 'English' stopwords list for fulfilling their motives, the technical jargon of biomedical research cannot rely on such generic stopwords lists. Therefore, creating a new list of stopwords based on subject terminology is essential. Thus, identifying insignificant, generic and uninformative stopwords in this domain was undertaken using alternative statistical metrics, including entropy, inverse document frequency, and word frequency (Sarica & Luo, 2021). Without having to perform the manual and ad hoc finding and removal of uninformative words, researchers and analysts working on textual data and technical language analysis can immediately employ it to denoise and filter their technical textual data.

## 6. CONCLUSION

In this study, we mined a corpus of 13281 unique publications on neurological disorders caused due to Covid-19 from two significant databases, WoS and PubMed. *Stroke* consisting of both ischemic and acute stroke is the predominant disorder which sparked interest among medical scientists leading to the highest number of publications in the particular disorder. Following *stroke, Epilepsy, Alzheimer's Disease, Multiple Sclerosis,* and *encephalitis* were few major disorders on which research was conducted significantly during 2020-22. It was found that the USA, among the countries, followed by Italy and England are countries predominantly associated with research on these domains. India, among other nodes, also played a significant role by conducting research in collaboration with USA, England and Netherlands. In terms of predominant organizations, it was discovered that Harvard Medical School, Tehran University of Medical Sciences, and the University of Milan are some of the major organizations whose researchers actively performed research relating to neurological disorders during 2020-22.

Further, we compared title and abstract keywords extracted through clustering and topic modelling using VOSviewer, CountVectorizer and YAKE. While the encephalitis term emerged as the most significant title-extracted term by VOSviewer, many other extraneous and insignificant terms, such as 'study', 'review', 'investigation' etc. found a place in the result set. On the other hand, the most significant term that our machine learning algorithm extracted from the title proper was ischemic stroke, followed by blood-brain barrier, cognitive impairment etc., which were more accurate and relevant than VOSviewer extracted terms. Similarly, comparing the significant terms that have been extracted from the abstracts of the titles, we observed that covid and ischemic stroke were the most important terms extracted through VOSviewer and machine learning methods, respectively. Discussing and evaluating the techniques based on the results obtained, we discovered that our algorithm-extracted keywords show higher relevancy. Summarizing all the findings, we can suggest that the ML algorithm being more compatible with unstructured data, is a valid alternative to the author-generated keywords for more accurate results.

### Contribution statement

Writing-original draft, data curation, formal analysis: Priya Tiwari.

Conceptualization, visualization: Saloni Chaudhary.

Software, writing-review and editing: Debasis Majhi.

Project administration, investigation, methodology, supervision, writing-review and editing: Dr. Bhaskar Mukherjee.

## Conflict of interest

The authors do not have any conflict of interest.

## Statement of data consent

The numeric data generated through Python-based programs during the development of the study has been included in the manuscript. ●

## REFERENCES


Campos, R., Mangaravite, V., Pasquali, A., Jorge, A., Nunes, C., & Jatowt, A. (2020). YAKE! Keyword extraction from single documents using multiple local features. *Information Sciences, 509,* 257-289. doi: 10.1016/j.ins.2019.09.013

Cheng, Q., Wang, J., Lu, W., Huang, Y., & Bu, Y. (2020). Keyword-citation-keyword network: A new perspective of discipline knowledge structure analysis. *Scientometrics, 124*(3), 1923-1943. doi: 10.1007/s11192-020-03576-5

Duvvuru, A., Radhakrishnan, S., More, D., Kamarthi, S., & Sultornsanee, S. (2013). Analyzing Structural & Temporal Characteristics of Keyword System in Academic Research Articles. *Procedia Computer Science, 20,* 439-445. doi: 10.1016/j.procs.2013.09.300

Graham, E. L., Clark, J. R., Orban, Z. S., Lim, P. H., Szymanski, A. L., Taylor, C., ... Koralnik, I. J. (2021). Persistent neurologic symptoms and cognitive dysfunction in non-hospitalized Covid-19 "long haulers." *Annals of Clinical and Translational Neurology, 8*(5), 1073-1085. doi: 10.1002/acn3.51350

Huang, T.-Y., & Zhao, B. (2019). Measuring popularity of ecological topics in a temporal dynamical knowledge network. *PLOS ONE, 14*(1), e0208370. doi: 10.1371/journal.pone.0208370

Kevork, E. K., & Vrechopoulos, A. P. (2009). CRM literature: Conceptual and functional insights by keyword analysis. *Marketing Intelligence & Planning, 27*(1), 48-85. doi: 10.1108/02634500910928362

Lu, W., Li, X., Liu, Z., & Cheng, Q. (2019). *How do Author-Selected Keywords Function Semantically in Scientific Manuscripts?*

Maurer, M. B., McCutcheon, S., & Schwing, T. (2011). Who's Doing What? Findability and Author-Supplied ETD Metadata in the Library Catalog. *Cataloging & Classification Quarterly, 49*(4), 277-310. doi: 10.1080/01639374.2011.573440

Papagiannopoulou, E., & Tsoumakas, G. (2020). A review of keyphrase extraction. *WIREs Data Mining and Knowledge Discovery, 10*(2), e1339. doi: 10.1002/widm.1339

Quan, C., Wang, M., & Ren, F. (2014). An Unsupervised Text Mining Method for Relation Extraction from Biomedical Literature. *PLOS ONE, 9*(7), e102039. doi: 10.1371/journal.pone.0102039

Rothstein, T. L. (2023). Cortical Grey matter volume depletion links to neurological sequelae in post COVID-19 "long haulers." *BMC Neurology, 23*(1), 22. doi: 10.1186/s12883-023-03049-1

Roy, D., Ghosh, R., Dubey, S., Dubey, M. J., Benito-León, J., & Kanti Ray, B. (2021). Neurological and Neuropsychiatric Impacts of COVID-19 Pandemic. *The Canadian Journal of Neurological Sciences. Le Journal Canadien Des Sciences Neurologiques, 48*(1), 9-24. doi: 10.1017/cjn.2020.173

Sarica, S., & Luo, J. (2021). Stopwords in technical language processing. *PLOS ONE, 16*(8), e0254937. doi: 10.1371/journal.pone.0254937

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science, 2*(3), 160. doi: 10.1007/s42979-021-00592-x

Small, H., Boyack, K. W., & Klavans, R. (2014). Identifying emerging topics in science and technology. *Research Policy, 43*(8), 1450-1467. doi: 10.1016/j.respol.2014.02.005

Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. *BMC Medical Informatics and Decision Making, 19*(1), 281. doi: 10.1186/s12911-019-1004-8

Wu, Y., Xu, X., Chen, Z., Duan, J., Hashimoto, K., Yang, L., ... Yang, C. (2020). Nervous system involvement after infection with COVID-19 and other coronaviruses. *Brain, Behavior, and Immunity, 87,* 18-22. doi: 10.1016/j.bbi.2020.03.031

Zamri, N., Pairan, M. A., Azman, W. N. A. W., Abas, S. S., Abdullah, L., Naim, S., ... Gao, M. (2022). A comparison of unsupervised and supervised machine learning algorithms to predict water pollutions. *Procedia Computer Science, 204,* 172-179. doi: 10.1016/j.procs.2022.08.021

Zhang, J., Yu, Q., Zheng, F., Long, C., Lu, Z., & Duan, Z. (2016). Comparing keywords plus of WOS and author keywords: A case study of patient adherence research. *Journal of the Association for Information Science and Technology, 67*(4), 967-972. doi: 10.1002/asi.23437

Van Eck, N. J., & Waltman, L. (2018). *VOSviewer Manual*. https://www.vosviewer.com/ documentation/Manual_VOSviewer_1.6.9.pdf