Vol 1 | Issue 1 | Pages 20-41

Archives of Information Science and Technology

Research Article

Applying Deep Learning Techniques for Big Data Analytics: A Systematic Literature Review

Memudu Tunde Muhammed[1], Obidallah Waeal J[1,2] and Raahemi Bijan[1]*

[1]Knowledge Discovery and Data Mining Lab, Telfer School of Management University of Ottawa, Ontario, Canada

[2]College of Computer and Information Sciences, Imam Mohammed ibn Saud University, Saudi Arabia

Abstract

Context: Data are being generated from numerous sources and applications and are thereby becoming increasingly complex. The increase in the use of technologies-such as phones, machines, vehicles, sports activities, and academic activities to carry out social and economic functions has also led to various forms of data being generated. The complexity, velocity, versatility, and volume of these data have introduced "big data", which is also called "large data". Because big data analysis is becoming a challenge to the exponential growth of data, deep learning, which is an aspect of machine learning, is considered a method of analyzing big data due to its use of excellent and advanced classification techniques and the hierarchical layer techniques. In this paper, we analyze how deep learning techniques and algorithms have been applied to big data, the types of datasets, the algorithms used, and the trend toward this area of study.

Objective: To identify, summarize, and systematically compare the current deep learning techniques and algorithms for big data; the various datasets to which deep learning algorithms are mostly applied; and the areas and fields in which the selected studies are being conducted, while providing answers to a specific set of research questions that goes thus: What are the relevant techniques, methods, and algorithms of deep learning in big data analysis? What are the most common datasets used for validation? And What are the trends and future research directions? These research questions helps us focus on the algorithms used in big data, which also helps with the results section of our paper, type of dataset that is being used for deep learning algorithms, and lastly to identify the most common area i.e. the field of study, the country, and the year in which this has trended.

Method: We conducted a systematic literature review (SLR) using predefined procedures that involved automatically searching five public digital libraries: IEEE, SCOPUS, ScienceDirect, Web of Science, and the ACM Digital Library. Of the original 863 papers retrieved from these search engines and during our two-stage scanning, 74 primary studies were identified, and we eventually selected 33 final papers, which we used in the synthesis of this study.

Results: This SLR includes definitions of big data and deep learning, the areas in which deep learning algorithms have been applied to big data, and the various types of deep learning techniques and algorithms used for big data. A synthesis that resulted in knowledge of the current state of the art in deep learning algorithms applied to big data. This SLR also identifies the trends in deep learning research that have been applied to big data over the past 10 years.

Conclusion: The application of deep learning has gained considerable attention since 2015. However, research needs to be done to improve the ways in which deep learning algorithms will be applied to varieties of data. Based on our review and analysis of the final selected studies, the big datasets to which these deep learning algorithms are applied are mostly image datasets. This demonstrates how well these algorithms can perform on image datasets. Furthermore, our review identifies the trends in the research on deep learning algorithms for big data, which will help researchers to understand the current state of the art regarding the use of deep learning algorithm on big data.

Keywords

Big data, Big data analytics, Systematic literature review, Deep learning

Introduction

As big data is now gaining recognition and is a fast-moving target area of focus in our current technology state and societies, it is now exceeding the amount of data we used to have. However, we still need to store, access, manage, and handle it [1]. The following three aspects characterize big data: (a) Large amounts of data, (b) The data cannot be classified as regular relational databases, and (c) The rapid generation, capture, and processing of data [2]. In addition, large data is changing health care, science, engineering, finance, business, and, ultimately, society [2]. The impact of big data on our society is very broad and will always gain attention from both technical and nontechnical experts [3]. Apparently, we live in a data-flood era, as evidenced by the amount of data from various sources, as well as their rate of growth [3]. Over the past two decades, the increasing capacity of computing has produced overwhelming number of data streams [4]. Data are generally growing at a rapid rate, thereby making it difficult to handle large amounts [5]. According to [5] one major problem with big data is the growth of data beyond the available computing resources.

However, extracting and analyzing relevant information from diverse and rapidly growing data in large quantities is a challenging task [6]. Data mining, knowledge discovery, or the extraction of useful and interesting information might be complex, especially with regard to big data. The analysis of a large data repository, for example, can be conducted using a complex program because its main goal is to extract useful information from the data [6]. These techniques are related to rule learning, classification, data mining, cluster analysis, machine learning (deep learning), and text analysis [6].

Machine learning, which has recently become a trend, can be used to solve the problems related to big data [7]. This study was driven by the potential of deep learning algorithms to classify data and provide hierarchical layer abstractions, as well as by the commercialization of machine learning frameworks, which makes it easier for researchers to quickly implement and deploy machine learning solutions regarding big data [7]. The relevant techniques and algorithms of deep learning, which is an aspect of machine learning, have made it useful for big data analytics [7]. Deep learning is currently an active area of research within machine learning and pattern recognition. It has enjoyed considerable success in a wide range of areas, such as speech recognition, computer vision, and natural language processing [6]. Because big data presents tremendous opportunities and potential for change within all organizations, applying deep learning to it will facilitate the achievement of impressive and meaningful results [8].

The aim of this SLR is to identify, explore, summarize, and synthesize how deep learning algorithms have been applied to big data, the types of data used, how big these data are, how they are generated, where they are being generated from, and the results. The final objective is to provide the reader with the current state of big data and deep learning by identifying the algorithms, datasets, and trends in this direction (over the years, in different countries and fields of study, and their various applications).

Compared to the narrative literature review, this SLR follows a defined and rigorous sequence of methodological research steps. SLRs depend on well-defined and evaluated study protocols to extract, synthesize, and report the results. We adopted the guidelines provided by [9], which outline the step-by-step procedures for conducting an SLR. According to [9] an SLR is defined as “a means of identifying, evaluating and interpreting all available research relevant to a particular research question, or topic area, or phenomenon of interest”.

We follow the general guidelines and procedures outlined by [9] for the three phases of the SLR: the planning, conducting, and reporting phases. The planning phase involves four main activities: (a) Identifying the need for the SLR, (b) Defining the research questions, (c) Developing the review protocol, and (d) Evaluating the review protocol. To frame the study, we identified three research questions. Based on the search questions and search strings identified in the planning phase, we searched five well-known digital databases and retrieved the results of 863 papers, which used several formats, including journal articles, peer-reviewed articles, conference proceedings, full conference proceeding books, and book chapters.

We made use of tools such as Parsifal and Mendeley. Parsifal [10] is an online tool designed to support researchers performing SLRs within the context of software engineering. Geographically distributed researchers can work together within a shared workspace, designing the protocol and conducting the research. Parsifal is an online tool designed to support researchers performing SLRs within the context of software engineering. Geographically distributed researchers can work together within a shared workspace, designing the protocol and conducting the research. Mendeley [11] is a free reference manager and academic social network that can facilitate the organization of research, collaboration with others online, and the discovery of the latest research. The 74 primary selected papers were chosen based on study selection processes during the planning phase. We used quality assessment to weight the importance of studies when the results were being synthesized and to support the validity of the selected papers in this review. The final activity led to the identification of 33 papers as the final selected papers, from which we obtained the information required to answer the research questions and to analyze and summarize the results.

In summary, we found one SLR related to these areas, and it focused solely on the features of deep learning in big data analysis, as elaborated by [6]. In our SLR, we focus on the deep learning algorithms applied on big data, the common dataset used for validation, and the trend toward this area of research.

The structure of this paper is organized as follows: Section 2 provides background about big data, its characteristics, problems of big data, and big data analytics In Section 3, we describe our research method: the search, selection, data extraction and quality assessment. In Sections 4 and 5, we report the steps for executing the SLR and the results of this study. Section 6 discusses possible threats to the validity of this SLR. Finally, we conclude by summarizing some of the outcomes of this study in Section 7.

***Corresponding author:** Obidallah Waeal J, Knowledge Discovery and Data Mining Lab, Telfer School of Management, University of Ottawa, 55 Laurier Ave., E., Ontario, Canada; College of Computer and Information Sciences, Imam Mohammed ibn Saud University, Saudi Arabia

Accepted: December 29, 2018

Published online: December 31, 2018

Citation: Muhammed MT, Obidallah WJ, Bijan R (2018) Applying Deep Learning Techniques for Big Data Analytics: A Systematic Literature Review. Arch Inf Sci Tech 1(1):20-41

Table 1: The difference between traditional data and big data [3].

Characteristics	**Big Data**	**Traditional Data**
Volume	Terabytes or Petabytes	Gigabytes
Data Source	Fully distributed	Centralized
Generated Rate	More rapid, seconds	Hourly, daily
Storage	HDFS, NOSQL	RDBMS
Data Integration	Difficult	Easy
Access	Near real-time, batch processing	Interactive
Structure	Unstructured, Semi-structured	Structured

Table 2: Big data issues discussed in different literature reviewed in this paper.

	Big Data Issues				
Reference Articles	[15]	[4]	[2]	[3]	[16]
Security	Yes	Yes	Yes	Yes	Yes
Privacy	Yes	No	Yes	No	No
Heterogeneity	Yes	No	Yes	No	No
Scalability	Yes	No	Yes	No	Yes
Data Integrity	No	No	Yes	Yes	Yes
Data Quality	No	No	Yes	Yes	Yes
Data Governance	Yes	No	Yes	No	No
Storage and Management	No	Yes	No	No	Yes
Computation and Analysis	No	Yes	No	No	Yes
Data Recovery	Yes	No	No	No	No
Data Transformation	No	No	Yes	Yes	Yes

Background

Many authors of various studies have provided different definitions of big data and deep learning; in most cases, they have been clear, similar, and easy to understand. According to Hu, et al. big data is classified into big data science and frameworks. Big data science is “the study of techniques covering the acquisition, conditioning, and evaluation of big data”, whereas big data frameworks are “software libraries along with their associated algorithms that enable distributed processing and analysis of big data problems across clusters of computer units” [3].

Big data

Big data generally refers to a volume of data that cannot be processed effectively using ordinary database methods [1,2] proposed a useful definition based on the literature and journals they consulted for their research, observation, and analysis of the essence of big data. According to their definition, “big data is a set of techniques and technologies that require new forms of integration to uncover large hidden values from large datasets that are diverse, complex, and of a massive scale” [2]. Gartner recently provided the following definition: “Big Data are high-volume, high-velocity, and/or high-variety information assets that require new forms of processing to enable enhanced decision-making, insight discovery and process optimization” [4]. The following attributes can also be used to define big data: volume, velocity, veracity and variety [5].

Big data is commonly attributed to the 4Vs, and it has also been referred to as the problem of big data [12]. It can also be defined as large, heterogeneous, and often unstructured data that are difficult to analyze and extract from using ordinary data management tools and techniques [13]. According to another definition, big data constitutes huge datasets with high-volume data, varieties of information, and great diversity, including structured, semi-structured, and unstructured data, which arrive faster (velocity) than traditional datasets [14].

The characteristics of big data: Big data has been characterized differently by various authors of different studies; however, in general, the most popular ways in which big data has been classified revolve around the Vs. (volume, velocity, variety, veracity, valence, and value). The following 4V characteristics have been generally accepted because they speak clearly to what big data is [2]: volume refers to large amounts of data; velocity to the speed of data generation; variety to structured data, unstructured data, and images; and veracity to trust and integrity. Table 1 adopted form [3] to show a comparison between big data and traditional data.

Big data issues: The following are the big data issues mentioned throughout the additional journals and articles reviewed in this survey: security, privacy, heterogeneity, data governance, disaster recovery, big data storage and management, big data computation and analysis, scalability and availability, data integrity, data transformation, and data quality. Table 2 shows a concept matrix of how big data issues have been discussed in various literatures.

One major problem with regard to big data is security, which is the issue that is addressed most often in the papers in Table 2. Big data can be difficult to secure because of its volume, its variety, the speed of the information, and the openness to a threat. Security is not the only aspect of concern; rather, all the other problems mentioned above have led to the various ways in which big data can be analyzed, as well as the different algorithms that can be used which motivate both researchers and practitioners.

Big data analytics and deep learning: Big data analytics is now being used in most aspects of studies and organizations, such as businesses and the science and technology areas as a way to build and predict models and trends [15]. Big data analytics also shares concerns with other data-related areas of study and disciplines, which have benefited from the previous body of knowledge that was developed in past years [16].

According to [14] big data analytics has a significant impact on organizations by creating competitive advantage and new ideas, as well as helping to generate revenue or increase the chances of revenue generation [14]. Therefore, many

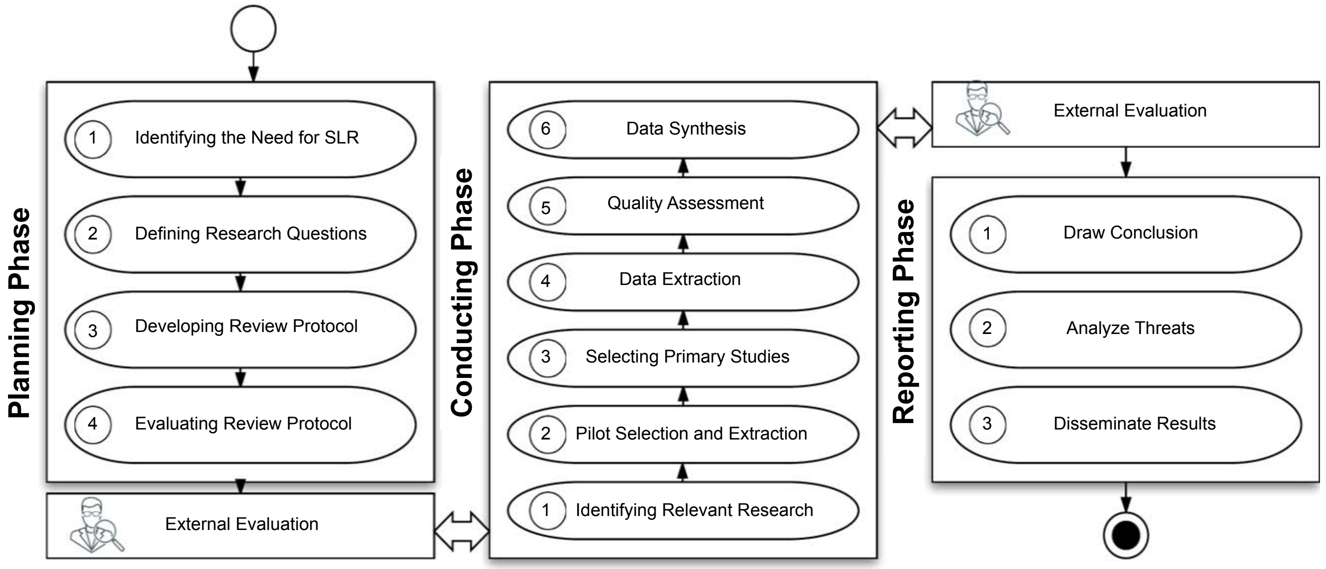

Planning Phase
1 Identifying the Need for SLR
2 Defining Research Questions
3 Developing Review Protocol
4 Evaluating Review Protocol
External Evaluation
Conducting Phase
1 Identifying Relevant Research
2 Pilot Selection and Extraction
3 Selecting Primary Studies
4 Data Extraction
5 Quality Assessment
6 Data Synthesis
External Evaluation
Reporting Phase
1 Draw Conclusion
2 Analyze Threats
3 Disseminate Results

Figure 1: Research method followed in the SLR [22].

analytical techniques, algorithms, and tools facilitate the acquisition of relevant information from raw and unprocessed big data [6]. Some of these techniques are association rule learning, data mining, cluster analysis, machine learning, text analytics, and business intelligence tools. For the purpose of this SLR, we focused primarily on deep learning algorithms for big data.

Deep learning is an area of machine learning research that is now trending in the artificial intelligence arena [17]. As a facet of the machine learning model, deep learning uses the supervised or unsupervised method to "learn hierarchical features for the tasks of classification and pattern recognition" [18]. According to [19], deep learning aims to move machine learning closer to one of its original goals, which is artificial intelligence. Deep learning has been used in many areas of life and research studies and has been applied in a variety of fields, such as medicine, botany, image recognition, food processing, mechanics, Web mining, text mining, and data mining.

Deep learning is a machine learning technique that extracts higher-level representations from datasets by creating and piling different layers like neurons [20]. According to recent studies, deep learning has demonstrated that it performs well in the areas of image processing, speech recognition, Web search, recommendation systems, and more [20]. In addition, [21] applied machine learning to medical data (electronic health records) on a large population of patients who may be associated with radiation oncology. This is one area of the medical field in which machine learning has been applied.

According to [8], deep learning has two well-established deep architectures: deep belief networks (DBNs) and convolutional neural networks (CNNs). DBN uses deep architecture to learn the feature representations from the tagged and unlabeled data presented to it, and it includes unsupervised pre-training and supervised fine-tuning strategies to build the model [8]. CNN consists of many hierarchies, some of which are used for feature representations (or feature maps), and others are traditional neural networks used for classification [8].

Research Method

We followed the guidelines provided by a systematic literature done in an engineering field [22], these steps were primarily recommended by [9] with its three phases as shown in Figure 1. The three phases include planning, conducting and reporting the review. An external evaluation of the outcome added at the end of each phase as a checkpoint to move to the next phase.

The following are the research phases:

- **Planning phase:** Identification of the need for SLR, definition of the research question, development of the review protocol, and evaluation of the review protocol.
- **Conducting phase:** Identification of the relevant research, pilot selection and extraction, selection of primary studies, data extraction, quality assessment, and data synthesis.
- **Reporting phase:** Drawing of conclusion, analysis of threat, dissemination of results.

The planning phase of SLR

Defining the research questions: We considered the research questions from the following five viewpoints: population, intervention, comparison, outcome, and context (PICOC) to define the scope, goals and string of the SLR as outlined in Table 3.

The research questions are as follows:

- What are the relevant techniques, methods, and algorithms of deep learning in big data analysis?
- What are the most common datasets used for validation?
- What are the trends and future research directions?

Table 4 comprises the explanations for the research questions and how they helped to narrow the scope of our review.

Search strategy: We need to identify the right search terms and keywords. Followed by the five criteria by [9], we used the used Population and Intervention in our search terms and keywords. The strategy for our search keyword is:

(P1 OR P2 …OR Pn) AND (I1 OR I2 …OR In)

Pn: population terms, In: intervention terms

From the Table 3, we have the population as "big data", "large data" and the intervention as "deep learning" Our search criteria is very straight and narrowing as we didn't consider different spelling on the keywords, in our case we didn't use asterisks (*) else we only used the Boolean Operators AND and OR. Our search strings are:

("Big data" OR "large data") AND "Deep learning"

We only used the advanced option for the ACM digital library in other for our search criteria to work properly and give relevant studies to our search keywords. We only focused on journals, peer-reviewed articles and conference papers. We run our search within a year span which is 10 years back from when this review was conducted in late 2017. We search for the studies published within 2007-2017. Table 5 illustrates the result of our first run of the search strings.

Table 3: PICOC Criteria.

Criteria	Describe/Reflect
Population	Big data, large data
Intervention	Deep learning
Comparison	A comparison of methods, algorithms, datasets
Outcome	A systematic literature survey report including synthesis of most relevant articles published on deep learning in big data analytics
Context	A systematic investigation to consolidate a peer-reviewed and academic research, Classification and comparison, trends and future research directions

Study selection process: At this stage, we identify relevant studies by defining the inclusion and exclusion criteria; removing duplicates of studies; screening titles, keywords, and abstracts; carrying out full-text screening, and extracting data. These processes were followed one stage after the other to obtain a good set of primary and final selected studies (Figure 2).

Quality assessment of primary selected studies: As part of the review protocol stage and as recommended by [9], it is important to assess the quality of the primary and the final selected studies used in this SLR. We have also used quality assessment as a means of assessing the importance of the selected studies when the results are being synthesized and supporting the validity of the selected papers analyzed in this review. We have based the quality assessment of our primary selected studies on the questions below. We seek to respect the work of all authors, and we attempt to make our quality assessment qualitative rather than subjective; we have, therefore, decided not to give a quality score. We have considered only a qualitative comparison-YES and NO-between our primary selected studies so as not to be biased by the quality of the selected studies. It is clear that the quality of each study is based on our quality assessments questions below.

Quality assessment questions

1. **Were the names of these deep learning techniques or algorithms mentioned?**

- Yes: They mentioned the names of the algorithms, techniques, or frameworks used.
- No: They did not mention the names, but it was stated that they applied these algorithms or techniques.

2. **Were the sizes of these data and the collection methods clearly mentioned?**

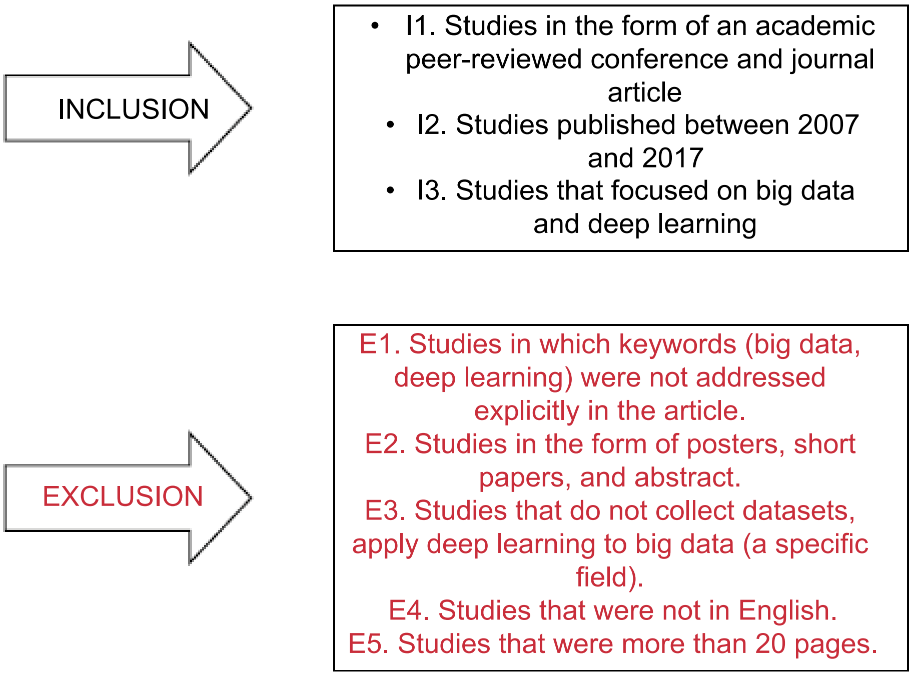
- Yes: They mentioned the sizes and how big the data are.
- No: They did not mention the sizes, but they collected the data.

Table 4: Research Question and Motivation.

No	Research questions	Motivations
RQ1	What are the relevant techniques, methods, and algorithms of deep learning in big data analysis?	This question helps us focus on the algorithms used in big data, which also helps with the results section of our paper.
RQ2	What are the most common datasets used for validation?	This enables us identify the type of dataset that is being used for deep learning algorithms.
RQ3	What are the trends and future research directions?	This helps us identify the most common area, the field of study, the country, and the year in which this has trended.

Table 5: Selected libraries and initially retrieved results.

No	Database Name	Web address	No of Initially Retrieved
1	ACM Digital Library	http://dl.acm.org/	40
2	ScienceDirect	http://www.sciencedirect.com/	76
3	Scopus	https://www.scopus.com/	433
4	IEEE Xplore	http://ieeexplore.ieee.org/	105
5	Web of Science	https://webofknowledge.com/	209
Total Number of papers		863	



- I1. Studies in the form of an academic peer-reviewed conference and journal article
- I2. Studies published between 2007 and 2017
- I3. Studies that focused on big data and deep learning

E1. Studies in which keywords (big data, deep learning) were not addressed explicitly in the article.
E2. Studies in the form of posters, short papers, and abstract.
E3. Studies that do not collect datasets, apply deep learning to big data (a specific field).
E4. Studies that were not in English.
E5. Studies that were more than 20 pages.

Figure 2: Inclusion and Exclusion Criteria.

Table 6: Data extraction form.

No	Data Extraction Category	Description	Purpose
1	Identifier	Identifier number (DOI)	
2	Date	Data extraction date	
Study Description			
1	Title	The title of the study	
2	Authors Name	Name of the study authors	
3	Country	Country of the study publication (1st author)	
4	Publication Year	Publication year	
5	Type	Conference proceedings, journal article	
6	Venue	Name of the Conference or Journal	
7	Author's Affiliations	The affiliation of the authors (1st author)	
Study Content			
1	Objectives	The objectives of the study	
2	Method	The method used to support the objectives	RQ1
3	Algorithm	The name of the algorithm used in the study	RQ1
4	Dataset	The name of the dataset used for validation	RQ2
5	Size of Dataset	How big the data set is	RQ2
6	Future directions and trend	Future directions, trends and application	RQ3

3. Were the findings credible, and did they contribute to our research questions? That is to say, can we extract the data we need from the articles?

- Yes: The findings in the articles or papers were credible and were useful for our research questions.

- No: The findings in the papers could not contribute to our research questions.

Data extraction: The data extraction phase involves collecting data and information, which is relevant to the research questions from the primary selected papers. Table 6 shows the data extraction form.

External evaluation: We evaluated our protocols by consulting the university librarian, doctoral students who have conducted SLRs in similar fields, and a professor from the university to demonstrate that we are on the right track. For instance, we consulted the university librarian, whom we asked to show us how to retrieve relevant resources from different digital libraries and how to assess the validity of our search keywords. Based on the feedback received, we refined the review scope and improved the search strategy and the inclusion and exclusion criteria to better narrow the scope of the study.

Conducting phase

This section provides details on how the SLR was conducted by showing how we identified our studies, the pilot selection, and the overall selection process.

Identifying the relevant research: We searched the five selected digital databases and retrieved the results. The search strings used were modified to meet the restrictions and requirements of the digital libraries. The databases of some digital libraries were more detailed and comprehensive than others. The search was conducted in September 2017 and was limited to studies published between 2007 and 2017. Appendix 1 shows the details search queries for each library.

The total number of articles found was 863 from all the search libraries, according to the defined search keys, queries, and strings above, and, as we stated in Table 5, the number of articles retrieved from each library. We found 177 sets of duplicates in the results. Our selection process was based on the remaining 686 after removing the duplicates.

Pilot selection and extraction: Before we began our selection process, we ran a pilot selection and extraction on 25 studies. These 25 studies were chosen from among the first search results from all the libraries. We choose five articles randomly from each library, and then we split these 25 articles into two baskets: relevant and irrelevant as indicated in Table 7. From the 25 articles chosen, we had a duplicate, which was removed, leaving 24 studies in general. We attempted to classify them into these baskets to see how many relevant and related articles we had for our review and how many were irrelevant to our study. The main purpose of doing this was to determine whether our search criteria and queries were good and led to relevant and related article, as we intended.

Selecting the primary studies: We excluded and included studies based on the predefined inclusion and exclusion criteria as indicated in Figure 3.

- Our initial search consisted of 863 papers. After deleting the 177 duplicate papers, the selection process consisted of three stages for screening the 686 papers that were left from the selected digital databases.
- Abstract and title screening: We screen out papers by filtering keywords, abstracts, titles based on the inclusion and exclusion criteria. Based on this stage of screening, we found that some of the retrieved papers did not address big data and deep learning explicitly. These papers might have been retrieved because they had "deep learning", "big data", or "large data" somewhere between the abstracts or keywords and titles of the articles; we, therefore, excluded 576 articles.
- From the 110 that were left, we excluded five studies because the university was unable to access them, and their full texts were, therefore, not available.

Table 7: Pilot selection table.

	Relevant	
No	**Title**	**Database**
1	A Hierarchical Fused Fuzzy Deep Neural Network for Data Classification	IEEE
2	Changing Mobile Data Analysis through Deep Learning	IEEE
3	Deep Learning Based Approach for Bearing Fault Diagnosis	IEEE
4	Deep Learning for Consumer Devices and Services: Pushing the limits for machine learning, artificial intelligence, and computer vision.	IEEE
5	How meta-heuristic algorithms contribute to deep learning in the hype of big data analytics	SCOPUS
6	Improving deep neural network design with new text data representations	SCOPUS
7	Learning Transportation Modes from Smartphone Sensors Based on Deep Neural Network	Web of Science
8	A GPU deep learning metaheuristic based model for time series forecasting	Web of Science
9	Prototyping a GPGPU Neural Network for Deep-Learning Big Data Analysis	Web of Science
10	Deep Learning at Scale and at Ease	ACM
11	Deep Learning for Mobile Multimedia: A Survey	ACM
12	Machine learning on big data: Opportunities and challenges	ScienceDirect
	Irrelevant	
No	**Title**	**Database**
1	Stacked Extreme Learning Machines	IEEE
2	4th International Conference on Advanced Computing, Networking and Informatics, ICACNI 2016	SCOPUS
3	Constructing fine-granularity functional brain network atlases via deep convolutional autoencoder	SCOPUS
4	InSTechAH: Cost-effectively autoscaling smart computing Hadoop cluster in private cloud	SCOPUS
5	Big Data technologies: A survey	ScienceDirect
6	Big data issues in smart grid - A review	ScienceDirect
7	From machine learning to deep learning: Progress in machine intelligence for rational drug discovery	ScienceDirect
8	Big Data sources and methods for social and economic analyses	ScienceDirect
9	Smart servitization within the context of industrial user-supplier relationships: Contingencies according to a machine tool manufacturer	Web of Science
10	A Small-Footprint Accelerator for Large-Scale Neural Networks	ACM
11	Joint embeddings of shapes and images via CNN image purification	ACM
12	Deep joint demosaicking and denoising	ACM

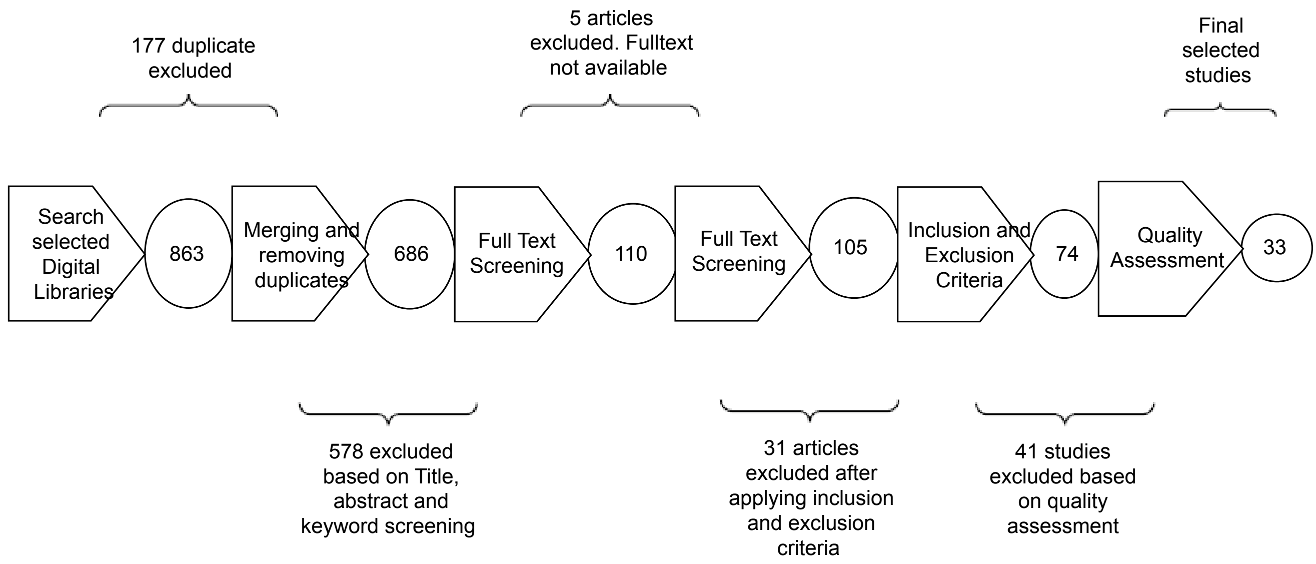

177 duplicate excluded
5 articles excluded. Fulltext not available
Final selected studies
Search selected Digital Libraries
863
Merging and removing duplicates
686
Full Text Screening
110
Full Text Screening
105
Inclusion and Exclusion Criteria
74
Quality Assessment
33
578 excluded based on Title, abstract and keyword screening
31 articles excluded after applying inclusion and exclusion criteria
41 studies excluded based on quality assessment

Figure 3: Overall selection process.

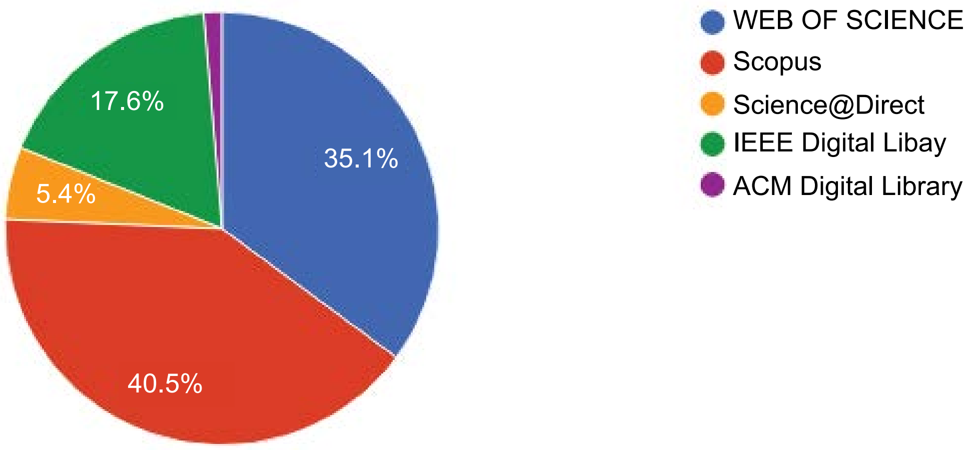


Figure 4: Distribution of primary studies across libraries.

- From our full-text screening, we excluded 31 articles based on our exclusion and inclusion criteria, which led to 74 articles.
- Full-text screening: A total of 74 papers were scanned during this stage, of which 41 were excluded based on our quality assessment, which helped us narrow the scope of our research.

Distribution and acceptance of primary studies across libraries: Figure 4 shows the distribution of our primary studies across the five libraries we used in our study. We retrieved most of the articles with the highest percentage rates from SCOPUS and those with the lowest percentage rates from the ACM Digital Library after inclusion and exclusion criteria, abstract, keyword, and title screening. Figure 5 shows the accepted articles per source-that is, the accepted studies from the five digital libraries we used for our SLR. As previously mentioned, we retrieved the highest number of studies, as well as the most accepted ones, from SCOPUS.

Analyzing the Selected Papers

In this section, we present the fundamental analysis of the results of the primary selected studies.

Quality of selected studies

According to [9], there is no generally accepted method of conducting study quality assessment. The quality of a study varies according to individuals' different perspectives. Thus, [9] highlighted some of the reasons for quality assessment: "To investigate whether quality differences provide an explanation for differences in study results, as a means of weighting the importance of individual studies when results are being synthesized".

In our SLR, the quality of the final selected papers was evaluated based on quality assessment questions we elaborated in Section 3.1.4. Table 8 provides the number of citations in our final selected studies, as well as the venues and the years. Our final selected papers, from which we extracted data to provide the answers to our research

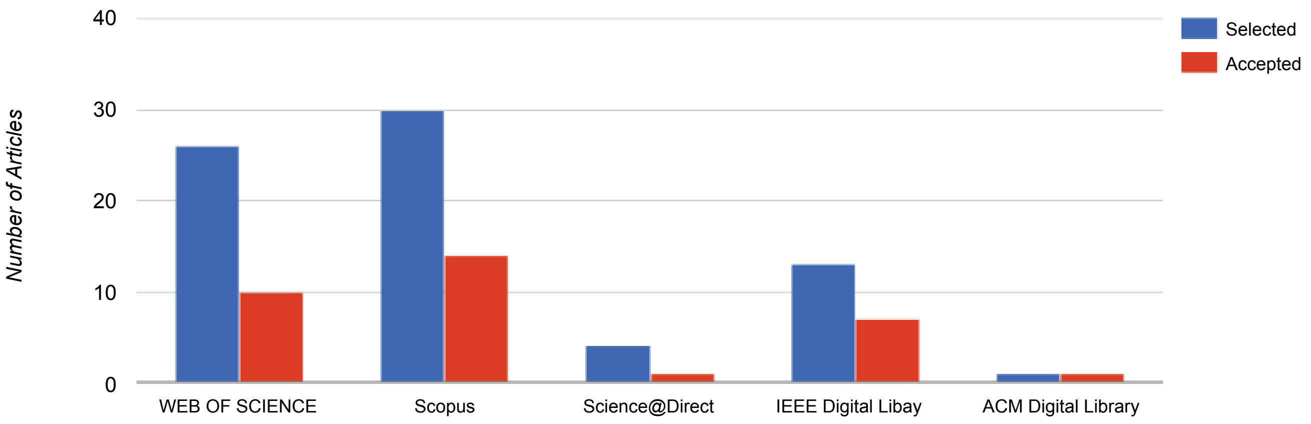


Figure 5: Acceptance of studies across the libraries.

Table 8: List of final selected studies.

ID	Title	Venues	Year	Citation
[FS1]	Shifu: Deep Learning Based Advisor-advisee Relationship Mining in Scholarly Big Data.	International World Wide Web Conference Committee (IW3C2) (C)	2017	1
[FS2]	GPU-Accelerated Parallel Hierarchical Extreme Learning Machine on Flink for Big Data	IEEE Transactions on Systems, Man, and Cybernetics: Systems	2017	3
[FS3]	Learning deep representation with large-scale attributes	Proceedings of the IEEE International Conference on Computer Vision	2015	13
[FS4]	Large-scale deep learning for computer-aided detection of mammographic lesions	MEDICAL IMAGE ANALYSIS (J)	2017	39
[FS5]	Mining Fashion Outfit Composition Using An End-to-End Deep Learning Approach on Set Data	IEEE TRANSACTIONS ON MULTIMEDIA (J)	2016	4
[FS6]	A novel sparse representation classification face recognition based on deep learning	IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (c)	2016	-
[FS7]	T-LRA: Trend-Based Learning Rate Annealing for Deep Neural Networks	IEEE International Conference on Multimedia Big Data BigMM (c)	2017	
[FS8]	Adaptive neuron apoptosis for accelerating deep learning on large scale systems	IEEE International Conference on Big Data (Big Data) (c)	2016	3
[FS9]	Retrieval from and Understanding of Large-Scale Multi – modal Medical Datasets : A Review	IEEE TRANSACTIONS ON MULTIMEDIA (J)	2017	-
[FS10]	Big Data and Deep Analytics Applied to the Common Tactical Air Picture (CTAP) and Combat Identification (CID)	International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2016) (c)	2016	-
[FS11]	Mobile Big Data Analytics Using Deep Learning and Apache Spark	IEEE NETWORK (J)	2016	27
[FS12]	Comparison between Multi-Class Classifiers and Deep Learning with Focus on Industry 4.0	Cybernetics & informatics K$I (C)	2016	1
[FS13]	Deep Neural Networks for Traffic Flow Prediction	International Conference on Big Data and Smart Computing IEEE BigComp (c)	2017	1
[FS14]	Gender Classification by Deep Learning on Millions of Weakly Labelled Images	International Conference on Data Mining Workshops (C)	2017	3
[FS15]	Large Deep Neural Networks for MS Lesion Segmentation	conference-proceedings-of-spie (C)	2017	-
[FS16]	Deep net architectures for visual-based clothing image recognition on large database	SOFT COMPUTING (J)	2017	1
[FS17]	Deep Computation Model for Unsupervised Feature Learning on Big Data	IEEE TRANSACTIONS ON SERVICES COMPUTING (c)	2016	19
[FS18]	Social Network Analysis of TV Drama Characters via Deep Concept Hierarchies	IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (C)	2015	6

[FS19]	Small boxes big data: A deep learning approach to optimize variable sized bin packing	BigDataService (C)	2017	-
[FS20]	Big-Data-Generated Traffic Flow Prediction Using Deep Learning and Dempster-Shafer Theory	International Joint Conference on Neural Networks IJCNN (C)	2016	3
[FS21]	Deep learning in remote sensing scene classification: a data augmentation enhanced convolutional neural network framework	GISCIENCE {\&} REMOTE SENSING (J)	2017	2
[FS22]	DP-miRNA: An improved prediction of precursor microRNA using deep learning model	International Conference on Big Data and Smart Computing IEEE BigComp (c)	2017	2
[FS23]	Predicting the Success of Bank Telemarketing using Deep Convolutional Neural Network	The International Conference on Soft Computing and Pattern Recognition SoCPaR (C)	2015	1
[FS24]	A Deep Learning Approach to Android Malware Feature Learning and Detection	IEEE Trustcom/BigDataSE/ISPA (J)	2016	4
[FS25]	Mass detection in digital breast tomosynthesis: Deep convolutional neural network with transfer learning from mammography	MEDICAL PHYSICS (J)	2016	14
[FS26]	A Novel Multimode Fault Classification Method Based on Deep Learning	Journal of Control Science and Engineering (j)	2017	1
[FS27]	Learning Transportation Modes From Smartphone Sensors Based on Deep Neural Network	IEEE SENSORS JOURNAL (J)	2017	-
[FS28]	Automated IT system failure prediction: A deep learning approach	Big Data (C)	2016	1
[FS29]	Deep Convolutional Computation Model for Feature Learning on Big Data in Internet of Things	IEEE Transactions on Industrial Informatics (J)	2017	1
[FS30]	Fast auto-clean CNN model for online prediction of food materials	Journal of Parallel and Distributed Computing	2017	1
[FS31]	Large-scale restricted Boltzmann machines on single GPU	Big Data (C)	2013	5
[FS32]	Weakly Semi-supervised Deep Learning for Multi-label Image Annotation	IEEE Transactions on Big Data (J)	2015	22
[FS33]	A Novel Left Ventricular Volumes Prediction Method Based on Deep Learning Network in Cardiac MRI	Computing in Cardiology (C)	2016	6

questions, as well as all the answers for all three quality assessment questions. Most of the studies were published in 2016 and 2017. While a few of them do not have citations yet, most have a significant number of citations. Table 8 shows the list of our final selected papers.

Publication venues and ranking

The publication venues included a list of 28 conferences with rankings based on Qualis and Excellence in Research for Australia, as well as 22 journals, outlined in Table 9 and Table 10 with their ranks, impact factors, abbreviations, and influence scores. As illustrated in Table 9 and Table 10, we also used the publication venues, rankings, impact factors, and influence scores to assess the quality of our studies. We divided the tables into two sections: final selected and not selected articles (i.e., articles in the primary selected category that are not part of the final selected papers). Based on the tables, we could find most of the ranks for only the final selected papers (articles), which illustrate the quality of the papers selected for the synthesis.

Distribution of primary selected papers over the year

All the selected papers that met our criteria were published in last ten years i.e. 2007-2017, and it shows a growing interest in the topic in the previous three years which are 2015, 2016, 2017. Figure 6 illustrates the distribution of papers over years, which shows that a significant amount of papers was published in 2016 and 2017. Since our study search finished September 2017, all studies published after September 2017 wasn't included in our study.

Publication of papers by country

The contribution of papers by countries is outlined in Figure 7 below which shows 20 different countries have contributed in this area big data or large data and deep learning. China has the largest number of contributions with 26 selected papers, followed by the USA with 16 selected papers, followed by Japan, Korea, and India, then Canada and Australia, Greece and the rest shown in Figure 7. This distribution based on the first author country, and it does not formulate any theory about the geographical allocation of teams working big data and deep learning at the time of the review. However, it shows a growing interest in the conducted study research area from different countries and teams.

Publication type

From the beginning of our study, we made it clear what kind of articles we want to consider for our study. These criteria was included in our inclusion and exclusion criteria. We have only considered peer review journals, articles and conference proceedings. The figure below shows the distribution of articles and conference proceedings we used for our study. We have the higher number of studies from conference proceedings as we know thus are is basically just

Table 9: List of Journals ranking in selected studies.

journals	**Venues**	**Abbreviation**	**Impact Factor**	**Article Influence Score**
	Final Selected Papers			
	IEEE Transactions on Multimedia		3.509	1.123
	IEEE Transactions on Big Data	TBD	1.203	-
	IEEE Transactions on Systems, Man, and Cybernetics: Systems	-	2.350	0.759
	Journal of Computing in Cardiology	CiC	1.30	-
	Journal of Parallel and Distributed Computing	JPDC	1.930	0.638
	IEEE Transactions on Industrial Informatics	IEEE T IND INFORM	6.764	1.964
	IEEE Sensors Journal	IEEE SJ	2.512	0.556
	Journals of Medical Physics	-	2.617	0.807
	IEEE Trustcom/BigDataSE/ISPA	-		
	GIScience & Remote Sensing	-	3.049	-
	Journals of Medical Image Analysis	MICCAI	4.188	1.948
	IEEE Transactions on Services Computing	TSC	3.520	0.913
	IEEE Network	-	7.230	2.261
	Journals of Soft Computing	IJSC	2.472	0.750
	IEEE computer society		1.755	0.644
journals	Not Selected Papers			
	Procedia Computer Science		1.08	
	Future Generation Computer Systems		3.997	1.151
	Knowledge-Based Systems	KNOWL-BASED SYST	5.02	0.68
	IT Professional	IT Prof Mag	1.35	
	Transportation Research Part C: Emerging Technologies		3.805	1.935
	Applied Intelligence	Appl Intell	2.67	0.30
	Procedia Engineering		0.73	
	International Journal of Molecular Sciences	IJMS	4.01	
	Cancer Letters		6.375	2.288
	Journal of Manufacturing Systems	JMSY	2.770	1.349
	Apsipa Transactions on Signal and Information Processing		1.64	
	IEEE International Conference on Robotics and Automation	ICRA		
	IEEE Transactions on Parallel and Distributed Systems	TPDS	4.181	1.207

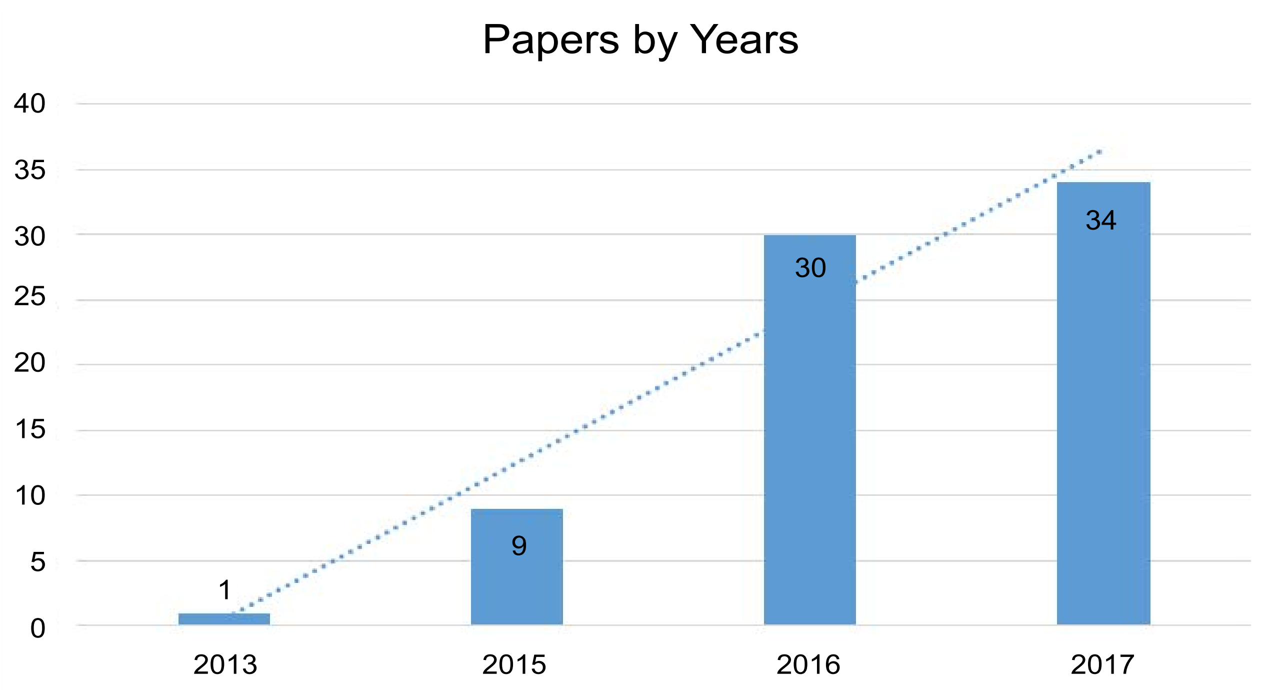


Figure 6: Distribution of primary selected papers over the year.

Table 10: List of conferences proceedings ranking in selected studies.

	Venues	Abbreviation	Rank (Qualis and ERA)
Conferences	**Final selected papers**		
	International World Wide Web Conference Committee	IW3C2	A (ERA)
	IEEE Third International Conference on Multimedia Big Data	BigMM	-
	IEEE International Conference on Big Data	Big Data	
	IEEE International Conference on Computer Vision	ICCV	A(ERA)
	International Conference of Soft Computing and Pattern Recognition	SoCPaR	-
	International Joint Conference on Neural Networks	IJCNN	A(ERA)
	IEEE International Conference on Big Data Computing Service and Applications, BigDataService	BigDataService	-
	IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining	ASONAM	B4(qualis)
	Conference-proceedings-of-spie	SPIE	C(ERA)
	IEEE International Conference on Data Mining Workshops	ICDM	A(ERA)
	IEEE BigComp 2017	BigComp	-
	proceeding of Cybernetics and Informatics	K$I	
	International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management	IC3K	C(ERA)
	Not selected papers		
	IEEE International Congress on Big Data	BigData Congress	-
	MATEC Web of Conferences		-
	IEEE International Workshop on Behavioral Implications of Contextual Analytics		-
	International Conference on Image, Vision and Computing	ICIVC	-
	International Conference on Cloud Computing and Big Data	CBDCom	-
	International Conference on Information Science and Control Engineering	ICISCE	-
	IEEE International Conference on Intelligent Engineering Systems	INES	-
	IEEE Winter Conference on Applications of Computer Vision Workshops	WACV	-
	International Conference on Reliability, Infocom Technologies and Optimization	ICRITO	-
	IEEE International Conference on Cyber Technology in Automation, Control and Intelligent Systems	CYBER	-
	Joint Conference on Digital Libraries	JCDL	A2(qualis)
	International Conference on Audio, Language and Image Processing	ICALIP	-
	IEEE International Symposium on Multimedia	ISM	B2(qualis)

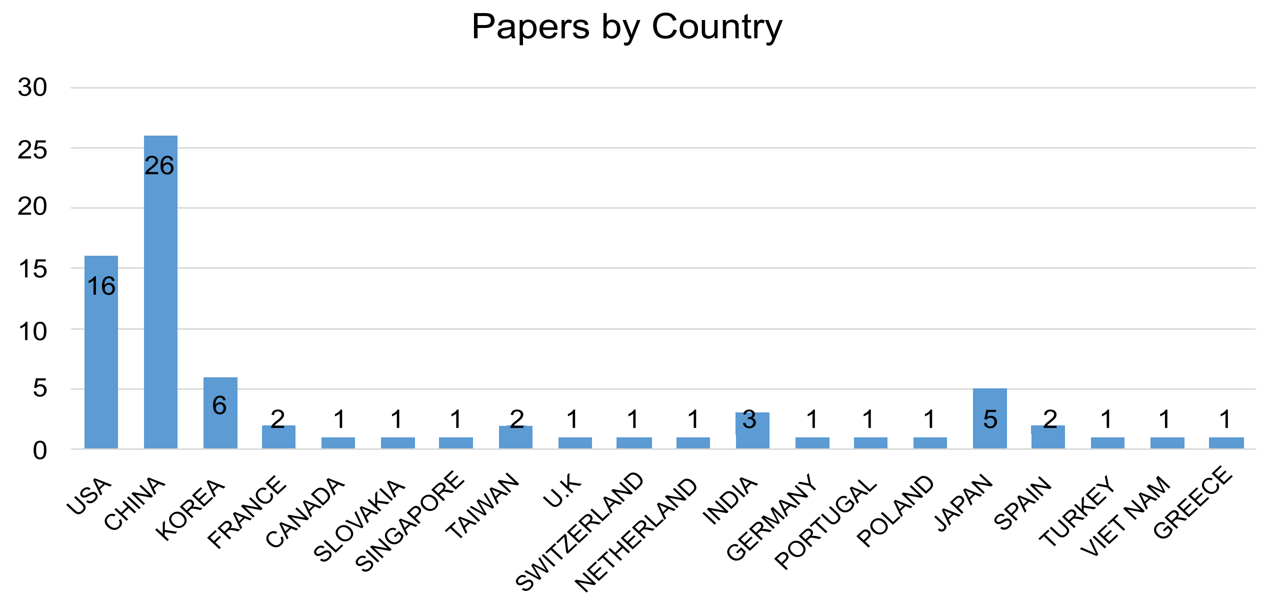


Figure 7: Publication of papers by country.

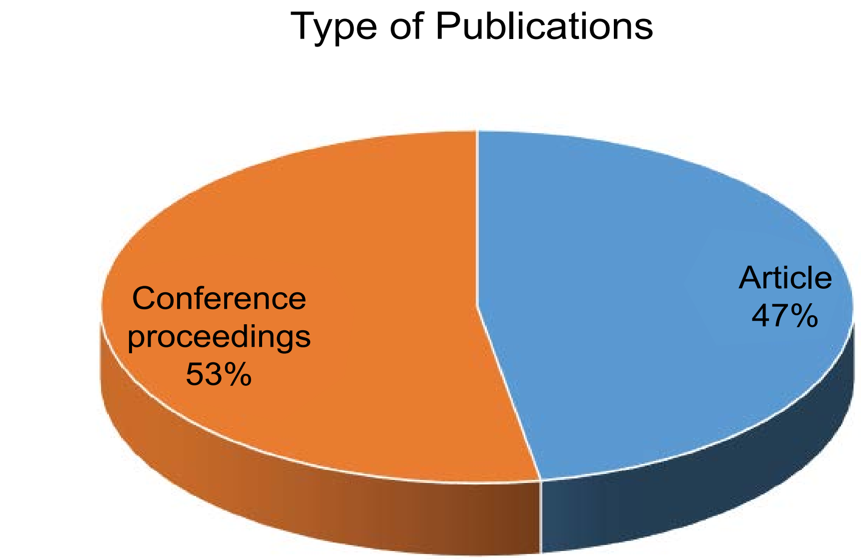


Figure 8: Publication type.

Table 11: Deep learning algorithms used on big data.

Algorithms	Description of the algorithms	No of FS Applied
CNN	Convolutional Neural Networks	[FS4], [FS7], [FS8], [FS13], [FS21], [FS29], [FS30], [FS32], [FS16], [FS19], [F5], [F10], [F23]
DNN	Deep neural networks	[FS8], [FS13], [FS15], [FS22], [FS25], [FS27], [FS26], [FS14]
DBN	Deep Belief Networks	[FS20], [FS6], [FS24], [FS31]
DCNN	Deep Convolutional Neural Networks	[FS33], [FS25], [FS23]
RNN	Recurrent Neural Networks	[FS13], [FS28], [FS7], [FS5], [FS18]
C-RNN	Convolutional-recursive neural network	[FS18]
RCNN	Region-based Convolutional Network	[FS3], [F29]
ELM, H-ELM	Extreme learning and hierarchical extreme learning	[FS2]
LEML	Metric learning method to learn	[FS4]
MDB	Spark-based Deep Learning Framework for MBD Analytics	[FS11]
H2O Deep Learning	Deep learning with H2O features	[FS12]
TDML	TENSOR DEEP LEARNING MODEL	[FS17]
DCH	Deep concept hierarchies	[FS18]

trending, few articles have been published to strictly deal with big data and deep learning tetchiness (Figure 8).

Analyzing the Content of the Final Selected Papers

In this section, we answer all of our research questions from the final selected papers (33 papers). We synthesized the data to answer these questions.

RQ1: What are the relevant techniques, methods, and algorithms of deep learning in big data analysis?

According to our review, a number of algorithms are used on big data. Some popular algorithms are associated with deep learning, whether big data or not. As deep learning algorithms are an aspect of machine learning, these algorithms cannot be overlooked. Table 11, which was formulated based on the final selected papers, shows the deep learning algorithms used on big data.

From Table 11, we can see how well these algorithms have been adopted. They have also been modified for various uses and applications, and some algorithms have even been deduced from the concepts of deep learning algorithms. In other words, some algorithms from our final selected papers used these deep learning algorithms as the foundation for the implementation of their own algorithms. This is similar to [FS1], which proposed a "shifu" algorithm, which was developed based on the CNN algorithms.

According to [23], deep learning algorithms can be classified into four major types:

- Convolution neural network (CNN)
- Deep neural network (DNN)
- Recurrent neural network (RNN)

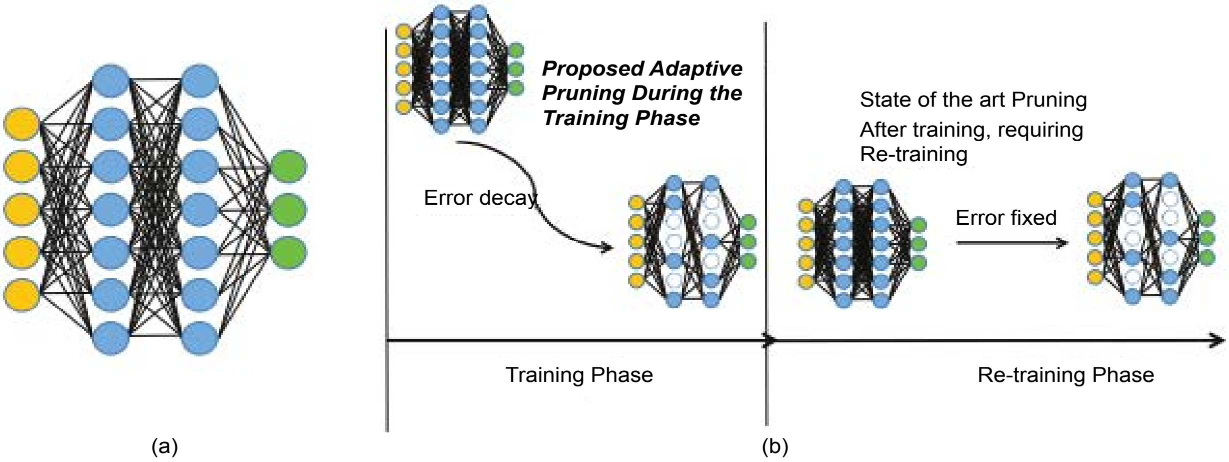

Proposed Adaptive Pruning During the Training Phase
Error decay
State of the art Pruning After training, requiring Re-training
Error fixed
Training Phase
Re-training Phase
(a)
(b)

Figure 9: An example of a Deep Neural Network (DNN) with two hidden layers from [31].

- Q-learning

Algorithm synthesis: Based on our final selected papers, three major algorithms are commonly implemented on big data: CNN, DNN, and RNN.

- **CNN**

CNN is the type of neural network that uses network structures such as "convolutional layers, spatial pooling layers, local response normalization layers and fully connected layers" [24]. CNNs have been widely used in many images, face identification, and text recognition [25]. The characteristics of big data, such as volume and diversity, are necessary for training robust deep learning models [25]. It is noticeable that one deep learning model trained based on data with sufficient diversity tends to outperform data with limited variability [25]. CNNs have been shown to be a significant trend in feature learning [24].

CNNs are among the most common deep learning algorithms used in big data analytics due to their hierarchical and neural structures. [FS4], [FS7], [FS8], [FS13], [FS21], [FS29], [FS30], [FS32], [FS16], and [FS19] have used the CNN algorithms to either improve a particular algorithm, or they have been used to classify data. Pouyanfar and Chen [26] in [FS7] used CNN to conduct experiments on a challenging "multimedia task, namely concept and image classification". [FS7], [FS8], [FS13], [FS21], [FS29], [FS30] papers had also developed their own algorithms from the CNN to analyze and classify data, as well as provide the solutions to their investigated problems. For instance, according to [24], CNN algorithms designed for image recognition consist of two components: "a multiple-layer architecture composed of several layers that gradually learns image representations from raw pixels, and a loss layer that propagates supervision cues back and fine-tunes the deep network to learn better representations for the specific tasks" [24]. Based on our final selected papers, we realized that data have been analyzed and algorithms have been proposed from a CNN based on the hierarchical relationships among the data.

- **DNN**

DNN algorithms perform well in big data-that is, where there is a high number of features [27]. DNN uses multilayer architecture to learn, classify, and represent [27]. This algorithm extracts high-level features from data that are necessary for classification [27]. DNNs are one of the famous machine learning classifiers due to their good feature-extraction techniques and good performance in regard to solving practical problems [28].

DNNs are said to give the best performance in "terms of accuracy and an acceptable model size" [29]. Generally, deep learning algorithms extract high-level, complex abstractions of the data [30]. The major benefit of DNN algorithms is that when the number of samples to be trained increases, classification accuracy also improves [30]. Figure 9 shows an example of DNN architecture with two hidden layers [31].

This algorithm has been used in [FS8], [FS13], [FS15], [FS22], [FS25], [FS24], [FS27], [FS26], and [FS14] to improve classification among data. From [FS15], [30] stated that although deep learning algorithms achieve excellent results in classification tasks, one of the major drawbacks for multiple sclerosis (MS) lesion classification can be traced back to the number of training sets used on the networks; this means that "the number of MS lesions samples is much lower than the number of samples drawn from the surrounding WM tissue" and is referred to as "The Class Imbalance Problem".

- **RNN**

RNN, which is another deep learning algorithm, is used for sequence generation and labeling due to its rich set of dynamic models [32]. According to [33] RNN is an advanced model that performs well when dealing with variables and length. RNNs also performed well on image and video captioning, language modeling, and machine translation, depending on the models within a time series [33].

From our final selected papers, [FS13], [FS28], [FS5] used RNN on the data they have, either by applying it directly to the data or by further developing a more sophisticated algorithm due to the limitations of RNN. For instance, from [FS28], the Long Short-Term Memory (LSTM) algorithm was designed from RNN architecture to improve "storing and accessing information compared to classical RNNs". This

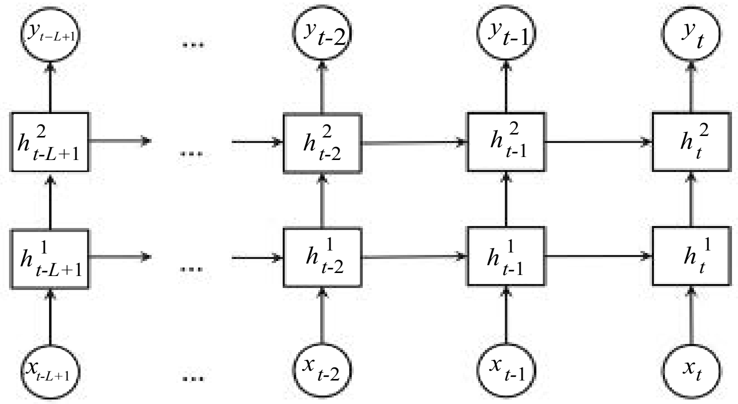


Figure 10: A many-to-many deep recurrent neural network prediction architecture [32].

LSTM has been applied to many sequence-modeling tasks, such as handwriting recognition, character generation, and sentiment analysis [33]. [FS5] also adapted RNN as a "pooling model to encode the variable length fashion items". With the fashion outfit dataset used, RNN maintains a state and performs the update for each fashion item. Figure 10 shows the predication architecture of deep RNN (DRNN) algorithm.

Based on our final selected papers, other algorithms have been applied to these big datasets. These algorithms are mostly derived from the modification of these common deep learning algorithms due to the limitation of these algorithms in order to perform certain tasks or for the purpose of the experiment. These algorithms are deep CNNs (DCNNs), deep belief networks (DBNs), HBPNNs, hierarchical SVMs (HSVMs) derived from DNNs, DRNN, LSTMs from RNNs, regCuCD-1 and cuCD-1 from DBNs, and droid deep from CNNs.

Common frameworks or methods:

- **Tensor Flow**

The TensorFlow (TF) framework, "is an open source software library for machine learning tasks" [30]. TF is said to be a data flow-based deep learning software package developed by Google Inc. in 2015 [23]. The TF framework has been recognized as implementing different versions that implement the RNN, DNN, and CNN algorithms [23]. Even though the TF framework is the most applied to the datasets or algorithms from deep learning, research shows that it is being limited to fields such as image recognition and speech recognition [23].

TF has been one of the most popular frameworks of deep learning algorithms used when applying them to big data or when building algorithms from deep learning or machine learning algorithms for big data. Based on our final selected papers, [FS5], [FS19], [FS8], [FS13], and [FS15] have used the TF framework in their proposed solutions.

For instance, [FS5] used TF in their implementation process, and their entire framework was implemented using TF [33]. [FS19] used the TF deep learning model to build the packing algorithms for their experiments [34]. [FS15] used the TF∗ framework to build a network that was trained on a set of dataset that yields a good mapping result for the experiment [30]. TF has been used by [FS13] as a deep learning neural network for predicting traffic flow conditions using real-time traffic data [23].

- **Hierarchical Layer**

Across all the final selected studies used for our SLR, the deep learning algorithms proposed or used in our study have shown different layers, such as the input layer, hidden layer, and output layer, which have been seen as characteristics of the deep learning algorithm. These layers can be referred to as the hierarchical layer. According to [6], hierarchical layer is defined as learning multiple of layers and classified as an algorithm that learns from the lower-level features in order to retrieve the higher-level features of the data via a hierarchical learning process [6]. Based on our final selected papers, almost all the reviews in which the deep learning algorithms were used highlighted that one of the most powerful attributes of the deep learning algorithm is its hierarchical attributes. According to [FS25], algorithms such as the DCNN can break an image from low- to high-level features with its hierarchical structure [35].

RQ2: What are the most common datasets used for validation?

Based on our final selected papers, various big datasets have been used for validation. These datasets vary from social media data to image datasets, medical datasets, traffic flow datasets, etc. Our research question seeks to determine the kind of big data to which the deep learning algorithms are applied. We prepared a table from our final selected papers to show the names of these datasets, how big they are, and their descriptions.

Table 12 shows the types of datasets used for validation, how they were collected, their sizes, and their descriptions.

Dataset application: From our final selected papers, we have been able to classify the datasets into specific areas of study and application. This has helped us to identify the most common areas in which deep learning algorithms are applied to big data. We classified them into areas of study, such as human relations, medicine, mechanics, multimedia, image recognition, face recognition, food processing, banking, hydrology, fashion, traffic flow, and logistics. We classified them by focusing on the content of the data and not the type

Table 12: Common datasets used for validation.

No of FS	Tittle	Name of dataset	Description of dataset	Size of dataset
[FS1]	Shifu: Deep Learning Based Advisor-advisee Relationship Mining in Scholarly Big Data.	PhDTree DBLP dataset.	A deep-learning-based advisor-advisee relationship identification method which takes into account both the local properties and network characteristics	Consist of Inferred 1,111,513 advisor- advisee pairs
[FS2]	GPU-Accelerated Parallel Hierarchical Extreme Learning Machine on Flink for Big Data	MINST,GISETTE, ADULT, WINE	Dataset of Flink cluster	Average running time of large-scale training samples with 2,000,000
[FS3]	Learning deep representation with large-scale attributes	ILSVRC 2014	large-scale object attribute database	Contains rich attribute annotations (over 300 attributes) for ~180k samples and 494 object classes
[FS4]	Large-scale deep learning for computer-aided detection of mammographic lesions	mammographic data set	Dataset that consist of an X-rays with diagnosed tumors of the breasts	The train and validation set comprise 44,090 mammographic views, from which we used 39,872 for training and 4218 for validation
[FS5]	Mining Fashion Outfit Composition Using an End-to-End Deep Learning Approach on Set Data	Outfit dataset	large-scale fashion outfit dataset with	195K outfits and 368K fashion items
[FS6]	A novel sparse representation classification face recognition based on deep learning	FERET face database,	Database of a face images	1400 images, including a total of 200 different people (seven of each face image)
[FS7]	T-LRA: Trend-Based Learning Rate Annealing for Deep Neural Networks	Multimedia dataset	Multimedia dataset with different images	This dataset includes 60,000 images (50,000 for training and 10,000 for testing)
[FS8]	Adaptive neuron apoptosis for accelerating deep learning on large scale systems	Higgs Boson data set	Not specified	Classification dataset (11M samples) and ImageNet classification datasets (≈1.3M images)
[FS9]	Retrieval from and Understanding of Large-Scale Multimodal Medical Datasets : A Review	Clinical data	Medical datasets in general. Some are Generated through xray	(small: < 1000 data items, medium, or large: > 10000), 3D microscopy images (17,107 images), 600 images in 2004 and 300,000 images in 2013
[FS10]	Big Data and Deep Analytics Applied to the Common Tactical Air Picture (CTAP) and Combat Identification (CID)	(NVESD)	Visible and IR imagery collected by the US Army Night Vision and Electronic Sensors Directorate	207 GB of IR imagery and 106 GB of visible imagery 4500 total images per test
[FS11]	Mobile Big Data Analytics Using Deep Learning and Apache Spark	Actitracker dataset	Which includes accelerometer samples of 6 conventional activities (walking, jogging, climbing stairs, sitting, standing, and lying down) from 563 crowdsourcing users	Both labeled and unlabeled data of 2,980,765 and 38,209,772 samples, respectively
[FS12]	Comparison between Multi-Class Classifiers and Deep Learning with Focus on Industry 4.0	H_2O cluster	datasets to the H_2O cluster	First 16,000 items from dataset. Other 4,000 items from dataset
[FS13]	Deep Neural Networks for Traffic Flow Prediction	The traffic flow condition data	The traffic flow condition data	The central server produces 6,260,603 traffic condition datasets every five minutes, 288 times per day. That is, the traffic flow condition data consists of 180,305,364 values (6,260,603-by-288 matrix) for each day
[FS14]	Gender Classification by Deep Learning on Millions of Weakly Labelled Images	Face images dataset	Dataset of face images	Five million weakly labeled face images

[FS15]	Large Deep Neural Networks for MS Lesion Segmentation	(CLIMB)	The Partners Multiple Sclerosis center enrolled in the CLIMB	3000 patients enrolled in the CLIMB study
[FS16]	Deep net architectures for visual-based clothing image recognition on large database	Clothing datasets	Clothing dataset	Dataset with 80,000 images
[FS17]	Deep Computation Model for Unsupervised Feature Learning on Big Data	CUAVE Dataset, SNAE2 Dataset, INEX 2007 Dataset	Not specified	500 training images, 800 testing images 100,000 unlabeled images 1,800 video clips grouped into four categories
[FS18]	Social Network Analysis of TV Drama Characters via Deep Concept Hierarchies	Data of TV drama	Data of TV drama	Adopting approximately 4400-minute data of TV drama
[FS19]	Small boxes big data: A deep learning approach to optimize variable sized bin packing	Real-world customer produced logistics orders of Walmart e-Commerce.	Real-world customer produced logistics orders of Walmart e-Commerce	4 million data, 4,278,645 instances
[FS20]	Big-Data-Generated Traffic Flow Prediction Using Deep Learning and Dempster-Shafer Theory	(PeMS), CityPulse Dataset	Traffic flow datasets. Caltrans Performance Measurements Systems (PeMS), CityPulse Dataset	There are 47 roads, where the traffic flow of each road is calculated based on the average of all the loop detectors in that particular road
[FS21]	Deep learning in remote sensing scene classification: a data augmentation enhanced convolutional neural network framework	SAT-4 and SAT-6	Were extracted from the NASA National Agriculture Imagery Program dataset	SAT-4 consists of 5,00,000 images, SAT-6 consists of 4,05,000 images
[FS22]	DP-miRNA: An improved prediction of precursor microRNA using deep learning model	pre-miRNA	The human pre-miRNA sequence	The negative dataset consists of 8494 pseudo hairpins
[FS23]	Predicting the Success of Bank Telemarketing using Deep Convolutional Neural Network	Bank marketing data	Bank marketing data	45,211 instances of whether acceptance or rejection to the phone call proposal for the given deposit option are collected in 2008~2010, by the Portuguese banking institution
[FS24]	A Deep Learning Approach to Android Malware Feature Learning and Detection	Benign apps	Not specified	In an experiment with 3,986 benign apps and 3,986 malware
[FS25]	Mass detection in digital breast tomosynthesis: Deep convolutional neural network with transfer learning from mammography	Mammographic dataset	Mammographic dataset	After data augmentation, a total of 45,072 mammographic ROIs and 37,450 DBT ROIs were obtained
[FS26]	A Novel Multimode Fault Classification Method Based on Deep Learning	Case Western Reserve University Bearing	Case Western Reserve University Bearing	This paper selects 200 samples in each fault type; each sample contains 2048 observation points. 100 samples are randomly selected as the training data, and the other 100 samples as the testing data
[FS27]	Learning Transportation Modes from Smartphone Sensors Based on Deep Neural Network	Accelerometer, magnetometer, and gyroscope measurements database	Accelerometer, magnetometer, and gyroscope measurements data base	The proposed mechanism is evaluated on a database that contains more than 1000 h of accelerometer, magnetometer, and gyroscope measurements from five transportation modes, including still, walk, run, bike, and vehicle
[FS28]	Automated IT system failure prediction: A deep learning approach	(WSC) and (MSC)	The dataset has been collected from two large enterprise systems, a web server cluster (WSC) and a mailer server, cluster (MSC)	(WSC) 1,885,022 (81.4%) (MSC) 4,536,360 (96.7%)

[FS29]	Deep Convolutional Computation Model for Feature Learning on Big Data in Internet of Things	CUAVE, SNAE2, and, STL-10	The SNAE2 is collected from YouTube	It contains more than 1800 pieces, whose theme focuses on sport, news, advertisement and entertainment. The STL-10 dataset contains more than 1300 images, where 500 images
[FS30]	Fast auto-clean CNN model for online prediction of food materials	(MLC dataset)	The dataset is collected by a large food supply chain platform in China (www.mealcome.com), it includes nearly	1000 restaurants and above 12 000 food supplies
[FS31]	Large-scale restricted Boltzmann machines on single GPU	RBM, regCuCD-1, cuCD-1	Not specified	No of parameters 32M 64M 128M 256M 512M 1B 2B
[FS32]	Weakly Semi-supervised Deep Learning for Multi-label Image Annotation	NUS-WIDE, MS COCO 2014	Not specified	Labeled images 82; 783 74; 320 64; 430 48; 265
[FS33]	A Novel Left Ventricular Volumes Prediction Method Based on Deep Learning Network in Cardiac MRI	ADSB datasets	Direct LV volumes prediction research using the new open accessible ADSB datasets	This datasets include 1140 subjects (more than 1026000 CMR images)

Table 13: Applications of dataset.

No	Applications	Counts
1.	Human activities	5
2.	Medical	5
3.	Pattern recognition (Face and image)	5
4.	Mechanics	3
5.	Multimedia	3
6.	Fashion	2
7.	Traffic flow	2
8.	Food processing	2
9.	Banking	1
10.	Sensor	1
11.	Malware	1
12.	Server	1
13.	Hydrology	1
14.	Logistics	1

of data collected, regardless of it being an image dataset or not. For instance, we classified data collected based on human interaction, behavior, actions, etc. Table 13 shows the applications and its count based on our final selected studies.

RQ3: What are the trends and future research directions?

Based on our study, we were able to identify the trend in this area of study. One way in which we were able to show how fast this area of study (big data, large data, and deep learning) is growing is by the number of studies published over the years from our primary and final selected papers, their applications, and the areas of study.

Distribution of final articles over the year: We showed the trends in the years of the studies-that is, after our study selection and quality assessment; it showed that research on this area is growing. There has been a rapid increase in research attention since 2015. Although the deep learning technology has been in existence for approximately 10 years, with regard to big data, its adaptation and implementation are now becoming a trend. Figure 11 shows the trend for these research areas.

Research groups per country: From our final selected papers, we were able to deduce the countries that contributed to research on big data, large data, and deep learning techniques as indicated in Figure 12. We considered only our final selected papers because they are the ones we analyzed in our study. The graph below shows that China has made the largest contribution to the area of study, followed by the United States and then Korea. We deduced that the population of a country might impact its contribution to the area of study. China, like the United States, has a very large population. More data might be generated from these two countries toward an area of research because there is a high probability that there will be people who are willing to contribute to the study, as well as more activities going on in these countries-for instance, e-commerce in China.

Research trends and direction: Big data analysis has recently been gaining more recognition; therefore, its challenges should be highlighted and focused on so that it can be more effective, as it is currently an active area of research. Our research also shows that more research will be done in this area in regard to the types of data to which these algorithms are being applied; for instance, soon, data types other than pictures, images, and text files will embrace these algorithms. According to our study, research is heading toward the modification of deep learning algorithms-that is, the modification of deep learning algorithms into other algorithms to solve specific problems, due to some of their limitations. In the future, deep learning algorithms, such as CNN, DNN, and RNN, might eventually turn into frameworks for researchers to build on. For instance, [33] proposed a fashion outfit scoring model based on a CNN algorithm. Figure 13 shows the architecture of their proposed model.

Threat to Validity

While conducting this SLR on big data and deep learning, we identified methods, algorithms, datasets, and trends in this area of study. However, validity is a major concern in

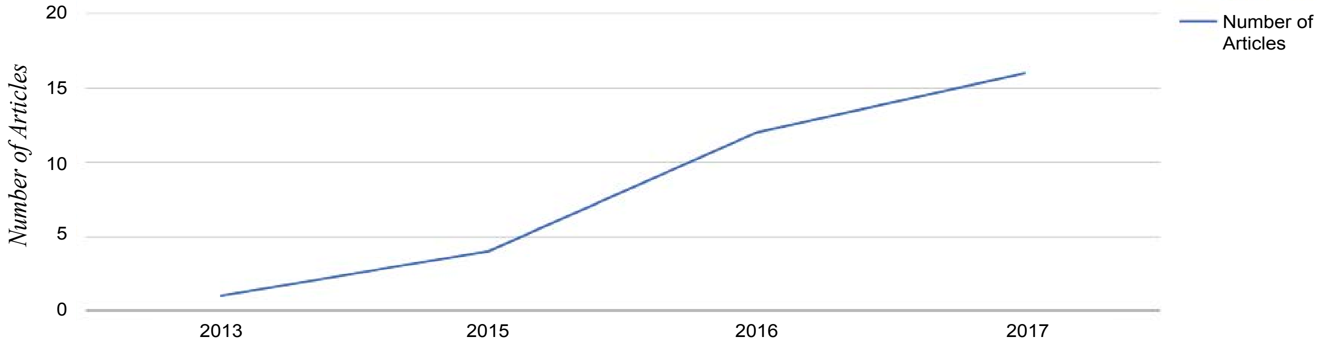


Figure 11: Final articles distribution over the years.

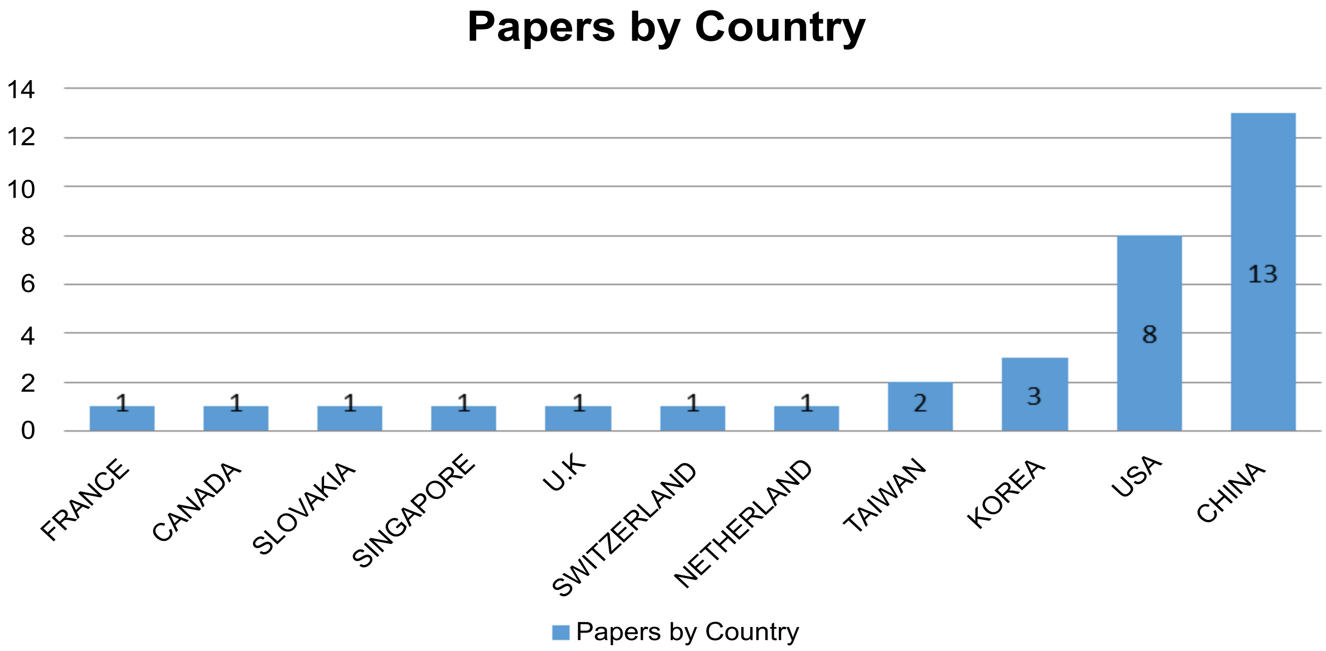


Figure 12: Research groups per country and number of publications.

such empirical studies. We identify our search and selection processes as the primary concern in this regard. The search terms and keywords were derived from the population and intervention used to form the research questions and were tested against a well-known list of research studies. However, the completeness and thoroughness of the terms are not always assured.

We were unable to gain access to some studies because the university had no access to them. They were, therefore, were excluded. Our exclusion and inclusion criteria may also be a factor that threatens the validity of this study, because, in some cases, we did not consider articles that were longer than 20 pages and shorter than two pages; the latter were considered posters. Although well-known digital libraries were used to search for the selected studies, other digital libraries may contain relevant studies that have not been taken into consideration [36-53].

We are also concerned about the data extraction. In the process of extracting data from the primary selected studies, we relied on our interpretation and analysis when the necessary data were not clearly stated. Regarding the data extraction process, some of the required data were missing from the final selected studies. This may be a threat to internal validity.

Our research questions and quality assessment were explicitly used to reduce the risk of generalizability of the outcomes. Furthermore, the SLR focused primarily on the common algorithms of deep learning used for big data, the common data used for validation, the trends in this area, how the research is growing, and the areas of study, considering only a predefined time for published articles: 2007 to 2017.

Conclusion

The reported SLR provides a well-detailed study of big data and deep learning. We identified common algorithms of deep learning used on big data, datasets in which deep learning

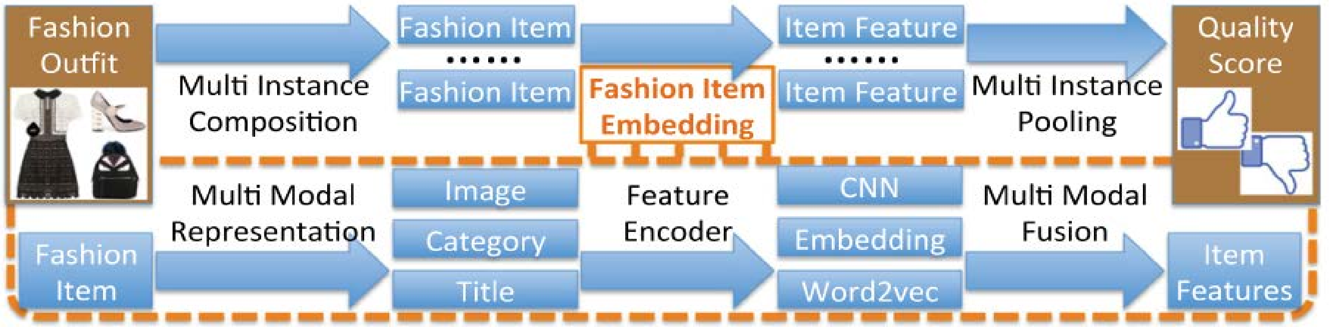


Figure 13: Proposed fashion outfit scoring model from [34].

algorithms are being applied, and the trends in this area of study. The review was conducted to assess the current state of the art in the area of deep learning and big data based on systematic procedures. The focus was on finding the answers to the predefined research questions, which were formulated based on the final selected papers. We identified 33 final selected papers with specific quality requirements and extracted the data to answer the research questions. From our study, we can identify the type of dataset that is most in need of the deep learning algorithm. This area is a current trend, but becuase we are barely in the world of big data and machine learning, improvising algorithms to solve big data problems might be challenging. We believe that there is a great deal of work to be done to improve the current state of research on the use of deep learning algorithms to solve big data problems, especially when the dataset is not an image dataset.

References

1. S Kaisler, F Armour, JA Espinosa, et al. (2013) Big data: Issues and challenges moving forward. 46th Hawaii Int Conf Syst Sci 995-1004.
2. IAT Hashem, I Yaqoob, NB Anuar, et al. (2015) The rise of 'big data' on cloud computing: Review and open research issues. Inf Syst 47: 98-115.
3. H Hu, Y Wen, TS Chua, et al. (2014) Toward scalable systems for big data analytics: A technology tutorial. IEEE Access 2: 652-687.
4. C Ji, Y Li, W Qiu, et al. (2012) Big data processing in cloud computing environments. 12th Int Symp Pervasive Syst Algorithms and Networks 17-23.
5. A Katal, M Wazid, RH Goudar (2013) Big data: Issues, challenges, tools and Good practices. 6th Int Conf Contemp Comput IC3 404-409.
6. NF Hordri, A Samar, SS Yuhaniz, et al. (2017) A systematic literature review on features of deep learning in big data analytics. Int J Adv Soft Compu Appl 9: 32-49.
7. CA Hammerschmidt, S Garcia, S Verwer, et al. (2017) Reliable machine learning for networking: Key issues and approaches. IEEE 42nd Conf Local Comput Networks 167-170.
8. XW Chen, X Lin (2014) Big data deep learning: Challenges and perspectives. IEEE ACCESS 2: 514-525.
9. B Kitchenham (2004) Procedures for performing systematic reviews. Keele Univ, Keele 33: 28.
10. (2015) Parsifal.
11. (2017) Mendeley.
12. A Fernandez, S del Rio, V Lopez, et al. (2014) Big data with cloud computing: An insight on the computing environment, mapreduce, and programming frameworks. Wiley Interdiscip Rev Data Min Knowl Discov 4: 380-409.
13. D Talia (2013) Clouds for scalable big data analytics. Computer 46: 98-101.
14. (2013) Big data in the cloud: Converging technologies. Intel IT Cent 12.
15. PC Neves, B Schmerl, J Bernardino, et al. (2016) Big data in cloud computing: Features and issues.
16. MD Assuncao, RN Calheiros, S Bianchi, et al. (2015) Big data computing and clouds: Trends and future directions. J Parallel Distrib Comput 3-15.
17. X Su, D Zhang, W Li, et al. (2016) A deep learning approach to android malware feature learning and detection. IEEE Trust 244-251.
18. P Li, Z Chen, LT Yang, et al. (2017) Deep convolutional computation model for feature learning on big data in internet of things. IEEE Trans Ind Informatics 790-798.
19. Caglar Gulcehre (2015) Welcome to deep learning.
20. K Jung, BT Zhang, P Mitra (2015) Deep learning for the web. WWW'15 Companion 1525-1526.
21. Bibault JE, Giraud P, Burgun A (2016) Big data and machine learning in radiation oncology: State of the art and future prospects. Cancer Lett 382: 110-117.
22. WJ Obidallah, B Raahemi Clustering and association rules for web service discovery and recommendation: A systematic literature review. 1-57.
23. Hongsuk Yi, HeeJin Jung, Sanghoon Bae (2017) Deep neural networks for traffic flow prediction. 2017 IEEE International Conference on Big Data and Smart Computing (BigComp) 328-331.
24. F Wu, Z Wang, Z Zhang, et al. (2015) Weakly semi-supervised deep learning for multi-label image annotation. IEEE Trans Big Data 1: 109-122.
25. X Yu, X Wu, C Luo, et al. (2017) Deep learning in remote sensing scene classification: A data augmentation enhanced convolutional neural network framework. GISCIENCE Remote Sens 54: 741-758.
26. S Pouyanfar, SC Chen (2017) T-LRA: Trend-based learning rate annealing for deep neural networks. 2017 IEEE 3rd Int Conf Multimed Big Data (BigMM) 50-57.

27. J Thomas, S Thomas, L Sael (2017) DP-miRNA: An improved prediction of precursor microRNA using deep learning model. 2017 IEEE International Conference on Big Data and Smart Computing (BigComp) 96-99.

28. KH Kim, CS Lee, SM Jo, et al. (2015) Predicting the success of bank telemarketing using deep convolutional neural network. 2015 7th International Conference of Soft Computing and Pattern Recognition (SoCPaR) 314-317.

29. SH Fang, YX Fei, Z Xu, et al. (2017) Learning transportation modes from smartphone sensors based on deep neural network. IEEE Sens J 17: 6111-6118.

30. JC Prieto, M Cavallari, M Palotai, et al. (2017) Large deep neural networks for MS lesion segmentation. Medical Imaging: Image Processing 10133.

31. C Siegel, J Daily, A Vishnu (2016) Adaptive neuron apoptosis for accelerating deep learning on large scale systems. 2016 IEEE International Conference on Big Data (Big Data) 753-762.

32. K Zhang, J Xu, MR Min, et al. (2016) Automated IT system failure prediction: A deep learning approach. 2016 IEEE International Conference on Big Data (Big Data) 1291-1300.

33. Y Li, L Cao, J Zhu, et al. (2017) Mining fashion outfit composition using an end-to-end deep learning approach on set data. IEEE Trans Multimed 19: 1946-1955.

34. F Mao, E Blanco, M Fu, et al. (2017) Small boxes big data: A deep learning approach to optimize variable sized bin packing. 2017 IEEE Third International Conference on Big Data Computing Service and Applications 80-89.

35. Samala RK, Chan HP, Hadjiiski L, et al. (2016) Mass detection in digital breast tomosynthesis: Deep convolutional neural network with transfer learning from mammography. Med Phys 43: 6654-6666.

36. W Wang, J Liu, F Xia, et al. (2017) Shifu: Deep learning based advisor-advisee relationship mining in scholarly big data. 303-310.

37. C Chen, K Li, A Ouyang, et al. (2017) GPU-Accelerated parallel hierarchical extreme learning machine on flink for big data. IEEE Trans Syst, Man, Cybern: Syst 47: 2740-2753.

38. W Ouyang, H Li, X Zeng, et al. (2015) Learning deep representation with large-scale attributes. IEEE Int Conf Comput Vis 1895-1903.

39. Kooi T, Litjens G, van Ginneken B, et al. (2017) Large scale deep learning for computer aided detection of mammographic lesions. Med Image Anal 35: 303-312.

40. J Zeng, Y Zhai, J Gan (2015) A novel sparse representation classification face recognition based on deep learning. 2015 IEEE 12th International Conference on Ubiquitous Intelligence and Computing, 2015 IEEE 12th International Conference on Advanced and Trusted Computing, 2015 IEEE 15th International Conference on Scalable Computing and Communications 20: 1520-1523.

41. H Muller, D Unay (2017) Retrieval from and understanding of large-scale multi-modal medical datasets: A review. IEEE Trans Multimed 19: 2093-2104.

42. Y Zhao, T Kendall, B Johnson (2016) Big data and deep analytics applied to the Common Tactical Air Picture (CTAP) and Combat Identification (CID). 1: 443-449.

43. MA Alsheikh, D Niyato, S Lin, et al. (2016) Mobile big data analytics using deep learning and apache spark. IEEE Netw 30: 22-29.

44. M Miskuf, I Zolotova (2016) Comparison between multi-class classifiers and deep learning with focus on industry 4.0. 2016 Cybernetics & Informatics (K&I).

45. S Jia, T Lansdall-Welfare, N Cristianini (2017) Gender classification by deep learning on millions of weakly labelled images. 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW) 462-467.

46. JC Chen, CF Liu (2017) Deep net architectures for visual-based clothing image recognition on large database. Soft Comput 21: 2923-2939.

47. Q Zhang, LT Yang, Z Chen (2016) Deep computation model for unsupervised feature learning on big data. IEEE Trans Serv Comput 9: 161-171.

48. C Nan, K Kim, B Zhang (2015) Social network analysis of tv drama characters via deep concept hierarchies. 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) 831-836.

49. R Soua, A Koesdwiady, F Karray (2016) Big-data-generated traffic flow prediction using deep learning and dempster-shafer theory. 2016 International Joint Conference on Neural Networks (IJCNN), 3195-3202.

50. F Zhou, Y Gao, C Wen (2017) A novel multimode fault classification method based on deep learning. Journal of Control Science and Engineering 2017: 442-452.

51. H Chen, J Xu, G Xiao, et al. (2017) Fast auto-clean CNN model for online prediction of food materials. J Parallel Distrib Comput 117: 218-227.

52. Y Zhu, Y Zhang, Y Pan (2013) Large-scale restricted boltzmann machines on single GPU. 2013 IEEE Int Conf Big Data 169-174.

53. G Luo, G Sun, K Wang, et al. (2016) A novel left ventricular volumes prediction method based on deep learning network in cardiac MRI. 2-5.

Appendix 1: Final selected papers and their references.

ID	Title	References
[FS1]	Shifu: Deep Learning Based Advisor-advisee Relationship Mining in Scholarly Big Data.	[36]
[FS2]	GPU-Accelerated Parallel Hierarchical Extreme Learning Machine on Flink for Big Data	[37]
[FS3]	Learning deep representation with large-scale attributes	[38]
[FS4]	Large-scale deep learning for computer-aided detection of mammographic lesions	[39]
[FS5]	Mining Fashion Outfit Composition Using An End-to-End Deep Learning Approach on Set Data	[33]
[FS6]	A novel sparse representation classification face recognition based on deep learning	[40]
[FS7]	T-LRA: Trend-Based Learning Rate Annealing for Deep Neural Networks	[26]
[FS8]	Adaptive neuron apoptosis for accelerating deep learning on large scale systems	[31]
[FS9]	Retrieval from and Understanding of Large-Scale Multimodal Medical Datasets : A Review	[41]
[FS10]	Big Data and Deep Analytics Applied to the Common Tactical Air Picture (CTAP) and Combat Identification (CID)	[42]
[FS11]	Mobile Big Data Analytics Using Deep Learning and Apache Spark	[43]
[FS12]	Comparison between Multi-Class Classifiers and Deep Learning with Focus on Industry 4.0	[44]
[FS13]	Deep Neural Networks for Traffic Flow Prediction	[23]
[FS14]	Gender Classification by Deep Learning on Millions of Weakly Labelled Images	[45]
[FS15]	Large Deep Neural Networks for MS Lesion Segmentation	[30]
[FS16]	Deep net architectures for visual-based clothing image recognition on large database	[46]
[FS17]	Deep Computation Model for Unsupervised Feature Learning on Big Data	[47]
[FS18]	Social Network Analysis of TV Drama Characters via Deep Concept Hierarchies	[48]
[FS19]	Small boxes big data: A deep learning approach to optimize variable sized bin packing	[34]
[FS20]	Big-Data-Generated Traffic Flow Prediction Using Deep Learning and Dempster-Shafer Theory	[49]
[FS21]	Deep learning in remote sensing scene classification: a data augmentation enhanced convolutional neural network framework	[25]
[FS22]	DP-miRNA: An improved prediction of precursor microRNA using deep learning model	[27]
[FS23]	Predicting the Success of Bank Telemarketing using Deep Convolutional Neural Network	[28]
[FS24]	A Deep Learning Approach to Android Malware Feature Learning and Detection	[17]
[FS25]	Mass detection in digital breast tomosynthesis: Deep convolutional neural network with transfer learning from mammography	[35]
[FS26]	A Novel Multimode Fault Classification Method Based on Deep Learning	[50]
[FS27]	Learning Transportation Modes From Smartphone Sensors Based on Deep Neural Network	[29]
[FS28]	Automated IT system failure prediction: A deep learning approach	[32]
[FS29]	Deep Convolutional Computation Model for Feature Learning on Big Data in Internet of Things	[18]
[FS30]	Fast auto-clean CNN model for online prediction of food materials	[51]
[FS31]	Large-scale restricted Boltzmann machines on single GPU	[52]
[FS32]	Weakly Semi-supervised Deep Learning for Multi-label Image Annotation	[24]
[FS33]	A Novel Left Ventricular Volumes Prediction Method Based on Deep Learning Network in Cardiac MRI	[53]