Sabharwal and Miah *Journal of Big Data* (2022) 9:53
https://doi.org/10.1186/s40537-022-00605-3

Journal of Big Data

METHODOLOGY Open Access

An intelligent literature review: adopting inductive approach to define machine learning applications in the clinical domain

Renu Sabharwal* and Shah J. Miah

*Correspondence:
renu.sabharwal@uon.edu.au
Newcastle Business School,
The University of Newcastle,
Newcastle, NSW, Australia

Abstract

Big data analytics utilizes different techniques to transform large volumes of big datasets. The analytics techniques utilize various computational methods such as Machine Learning (ML) for converting raw data into valuable insights. The ML assists individuals in performing work activities intelligently, which empowers decision-makers. Since academics and industry practitioners have growing interests in ML, various existing review studies have explored different applications of ML for enhancing knowledge about specific problem domains. However, in most of the cases existing studies suffer from the limitations of employing a holistic, automated approach. While several researchers developed various techniques to automate the systematic literature review process, they also seemed to lack transparency and guidance for future researchers. This research aims to promote the utilization of intelligent literature reviews for researchers by introducing a step-by-step automated framework. We offer an intelligent literature review to obtain in-depth analytical insight of ML applications in the clinical domain to (a) develop the intelligent literature framework using traditional literature and Latent Dirichlet Allocation (LDA) topic modeling, (b) analyze research documents using traditional systematic literature review revealing ML applications, and (c) identify topics from documents using LDA topic modeling. We used a PRISMA framework for the review to harness samples sourced from four major databases (e.g., IEEE, PubMed, Scopus, and Google Scholar) published between 2016 and 2021 (September). The framework comprises two stages—(a) traditional systematic literature review consisting of three stages (planning, conducting, and reporting) and (b) LDA topic modeling that consists of three steps (pre-processing, topic modeling, and post-processing). The intelligent literature review framework transparently and reliably reviewed 305 sample documents.

Keywords: Machine learning, Clinical research, Systematic literature review, Latent Dirichlet Allocation, Topic modeling

Introduction

Organizations are continuously harnessing the power of various big data adopting different ML techniques. Captured insights from big data may create a greater impact to reshape their business operations and processes. As a vital technique, big

Springer Open

© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

data analytics methods are used to transform complicated and huge amounts of data, known as 'Big Data, in order to uncover hidden patterns, new learning, untold facts or associations, anomalies, and other perceptions [41]. Big Data alludes to the enormous amount of data that a traditional database management system cannot handle. In most of the cases, traditional software functions would be inadequate to analyze or process them. Big data are characterized by the 5 V's, which refers to volume, variety, velocity, veracity, and value [22]. ML is a vital approach to design useful big data analytics techniques, which is a rapidly growing sub-field in information sciences that deals with all these characteristics. ML employs numerous methods for machines to learn from past experiences (e.g., past datasets) reducing the extra burden of writing codes in traditional programming [7, 26]. Clinical care enterprises face a huge challenge due to the increasing demand of big data processing to improve clinical care outcomes. For example, an electronic health record contains a huge amount of patient information, drug administration, imaging data using various modalities. The variety and quantity of the huge data provide in the clinical domain as an ideal topic to appraise the value of ML in research.

Existing ML approaches, such as Oala et al. [35] proposed an algorithmic framework that give a path towards the effective and reliable application of ML in the healthcare domain. In conjunction with their systematic review, our research offers a smart literature review that consolidates a traditional literature review followed the PRISMA framework guidelines and topic modeling using LDA, focusing on the clinical domain. Most of the existing literature focused on the healthcare domain [14, 42, 49] are more inclusive and of a broader scope with a requisite of medical activities, whereas our research is primarily focused is clinical, which assist in diagnosing and treating patients as well as includes clinical aspects of medicine.

Since clinical research has developed, the area has become increasingly attractive to clinical researchers, in particular for learning insights of ML applications in clinical practices. This is because of its practical pertinence to clinical patients, professionals, clinical application designers, and other specialists supported by the omnipresence of clinical disease management techniques. Although the advantage is presumed for the target audience, such as self-management abilities (self-efficacy and investment behavior) and physical or mental condition of life amid long-term ill patients, clinical care specialists (such as further developing independent direction and providing care support to patients), their clinical care have not been previously assessed and conceptualized as a well-defined and essential sub-field of health care research. It is important to portray similar studies utilizing different types of review approaches in the aspect of the utilization of ML/DL and its value. Table 1 represents some examples of existing studies with various points and review approaches in the domain.

Although the existing studies included in Table 1 give an understanding of designated aspects of ML/DL utilization in clinical care, they show a lack of focus on how key points addressed in existing ML/DL research are developing. Further to this, they indicate a clear need towards an understanding of multidisciplinary affiliations and profiles of ML/DL that could provide significant knowledge to new specialists or professionals in this space. For instance, Brnabic and Hess [8] recommended a direction for future research

Table 1 Previous review studies utilizing various methodologies in the clinical domain

Existing studies	Literature methodologies	The outcome of the analysis
Brnabic and Hess (2021) [8]	34 articles, systematic review	Found a wide assortment of approaches, methods, techniques, software and validation procedures utilized in using ML/DL strategies to illuminate patient-provider decision making
Robles Mendo et al. (2021) [27]	20 articles, (PRISMA framework, systematic review	Reviewed commercial applications found in the best-known commercial platforms
Salazar-Reyna et al. (2020) [42]	576 articles, systematic review	Assessed and synthesized the published literature related to applying data analytics, big data, data mining, and ML to healthcare engineering systems
Verma et al. (2021) [47]	15 articles, systematic review	Utilized ML/DL strategies at various phases of exploiting datasets consisting of patient-detailed outcome measures for anticipating clinical outcomes, introducing the promising study and demonstrating the utility of patient-reported outcome measures data for developmental research, and personalized treatment and precision medicine with the help of ML-based decision-support systems

by stating that "*Future work should routinely employ ensemble methods incorporating various applications of machine learning algorithms*" (p. 1).

ML tools have become the central focus of modern biomedical research, because of better admittance to large datasets, exponential processing power, and key algorithmic developments allowing ML models to handle increasingly challenging data [19]. Different ML approaches can analyze a huge amount of data, including difficult and abnormal patterns. Most studies have focused on ML and its impacts on clinical practices [2, 9, 10, 24, 26, 34, 43]. Fewer studies have examined the utilization of ML algorithms [11, 20, 45, 48] for more holistic benefits for clinical researchers.

ML becomes an interdisciplinary science that integrates computer science, mathematics, and statistics. It is also a methodology that builds smart machines for artificial intelligence. Its applications comprise algorithms, an assortment of instructions to perform specific tasks, crafted to independently learn from data without human intercession. Over time, ML algorithms improve their prediction accuracy without a need for programming. Based on this, we offer an intelligent literature review using traditional literature review and Latent Dirichlet Allocation (LDA[1]) topic modeling in order to meet knowledge demands in the clinical domain. Theoretical measures direct the current study results because previous literature provides a strong foundation for future IS researchers to investigate ML in the clinical sector. The main aim of this study is to develop an intelligent literature framework using traditional literature. For this purpose, we employed four digital databases -IEEE, Google Scholar, PubMed, and Scopus then performed LDA topic modeling, which may assist healthcare or clinical researchers in analyzing many documents intelligently with little effort and a small amount of time.

[1] LDA is a probabilistic method for topic modeling in text analysis, providing both a predictive and latent topic representation.

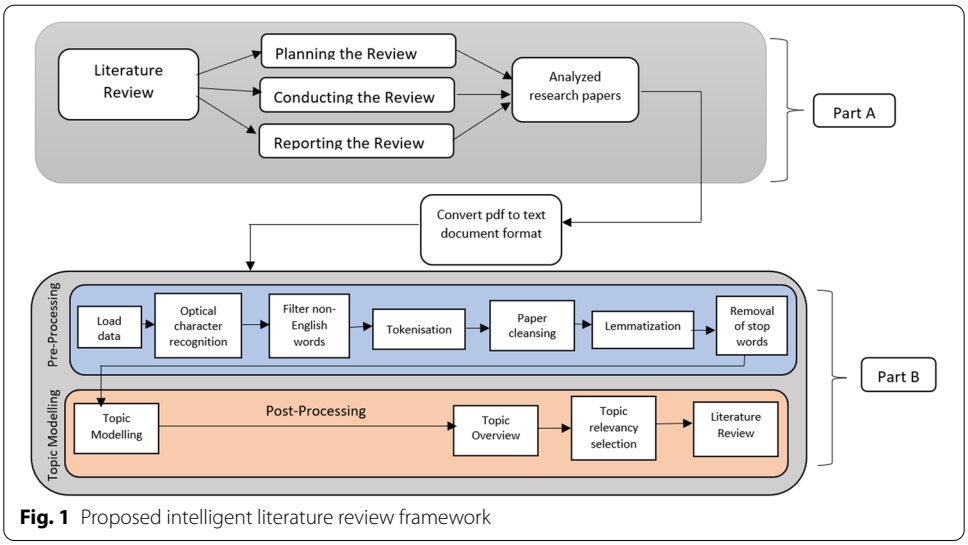

Fig. 1 Proposed intelligent literature review framework

Methodology

Traditional systematic literature is destined to be obsolete, time-consuming with restricted processing power, resulting in fewer sample documents investigated. Academic and practitioner-researchers are frequently required to discover, organize, and comprehend new and unexplored research areas. As a part of a traditional literature review that involves an enormous number of papers, the choice for a researcher is either to restrict the number of documents to review a priori or analyze the study using some other methods.

Framework

The proposed intelligent literature review approach consists of Part A and Part B, a combination of traditional systematic literature review and topic modeling that may assist future researchers in using appropriate technology, producing accurate results, and saving time. We present the framework below in Fig. 1.

The traditional literature review identified 534,327 articles embraces Scopus (24,498), IEEE (2558), PubMed (11,271), and Google Scholar (496,000) articles, which went through three stages–Planning the review, conducting the review, and reporting the review and analyzed 305 articles, where we performed topic modeling using LDA.

We follow traditional systematic literature review methodologies [25, 39, 40] including a PRISMA framework [37]. We review four digital databases and deliberately develop three stages entailing planning, conducting, and reporting the review (Fig. 2).

Planning the review

Research articles: the research articles are classified using some keywords mentioned below in Tables 2, 3.

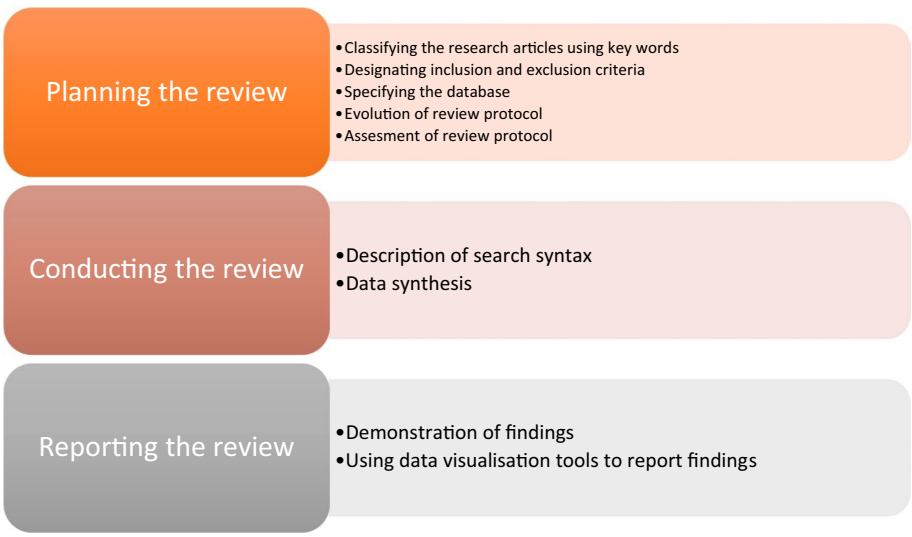

Fig. 2 Traditional literature review three stages

Table 2 Inclusion criteria

Inclusion criteria	Description
1	Keywords (or short phrases) include: "Machine Learning," "Machine Learning application," "Machine Learning algorithms," "Machine Learning techniques," "Clinical," "Clinical domain," "Clinical sector." Operators of search syntax are OR, AND. AND operator signifies that both keywords must be present in the search queries, and OR means that at least one keyword must be present in the queries searched
2	Research articles published between 2016 and 2021 (September)
3	Research articles published in English
4	Research limited to journal and conference articles
5	Only full-text articles

Table 3 Exclusion criteria

Exclusion criteria	Description
1	Exclude duplicate research articles with matching title and/or digital object identifiers (DOI)
2	Non-English research articles

Digital database: Four databases (IEEE, PubMed, Scopus, and Google Scholar) were used to collect details for reviewing research articles.

Review protocol development: We first used Scopus to search the information and found many studies regarding this review. We then searched PubMed, IEEE, and Google scholar for articles and extracted only relevant papers matching our keywords and review context based on their full-text availability.

Review protocol evaluation: To support the selection of research articles and inclusion and exclusion criteria, the quality of articles was explored and assessed to appraise their

Table 4 Search syntax for selected research articles

Database	Search syntax
Scopus	TITLEA ND ABSTRACT (machine learning in clinical) AND PUBYEAR > 2015 AND (LIMIT-TO (OA,"all")) AND (LIMIT-TO (LANGUAGE,"English")) AND (LIMIT-TO (PUBSTAGE,"final")) AND (LIMIT-TO (DOCTYPE,"ar") OR LIMIT-TO (DOCTYPE,"re") OR LIMIT-TO (DOCTYPE,"cp"))
IEEE	"Document Title and Abstract": machine learning in clinical
PubMed	(("machine"[All Fields] OR "machines"[All Fields]) AND "learning in clinical"[Title]) AND (2016:2021[pdat])
Google Scholar	allintitle: "machine learning in clinical"

suitability and impartiality [44]. Only articles with keywords "machine learning" and "clinical" in document titles and abstracts were selected.

Conducting the review

The second step is conducting the review, which includes a description of Search Syntax and data synthesis.

Search syntax Table 4 details the syntax used to select research articles.

Data synthesis We used a qualitative meta-synthesis technique to understand the methodology, algorithms, applications, qualities, results, and current research impediments. Qualitative meta-synthesis is a coherent approach for analyzing data across qualitative studies [4]. Our first search identified 534,327 papers, comprising Scopus (24,498), IEEE (2,558), PubMed (11,271), and Google Scholar (496,000) articles with the selected keywords. After subjecting this dataset to our inclusion and exclusion criteria, articles were reduced to Scopus (181), IEEE (62), PubMed (37), and Google Scholar (46) (Fig. 3).

Reporting the review

This section displays the result of the traditional literature review.

Demonstration of findings A search including linear literature and citation chaining was acted in digital databases, and the resulted papers were thoroughly analyzed to choose only the most pertinent articles, at last, 305 articles were included for the Part B review. Information of such articles were classified, organized, and demonstrated to show the finding.

Report the findings The word cloud is displayed on the selected 305 research articles which give an overview of the frequency of the word within those 305 research articles. The chosen articles are moved to the next step to perform the conversion of PDF files to text documents for performing LDA topic modeling (Fig. 4).

Conversion of pdf files to a text document

The Python coding is used to convert pdf files shared on GitHub https://github.com/MachineLearning-UON/Topic-modeling-using-LDA.git. The one text document is prepared with 305 research papers collected from a traditional literature review.

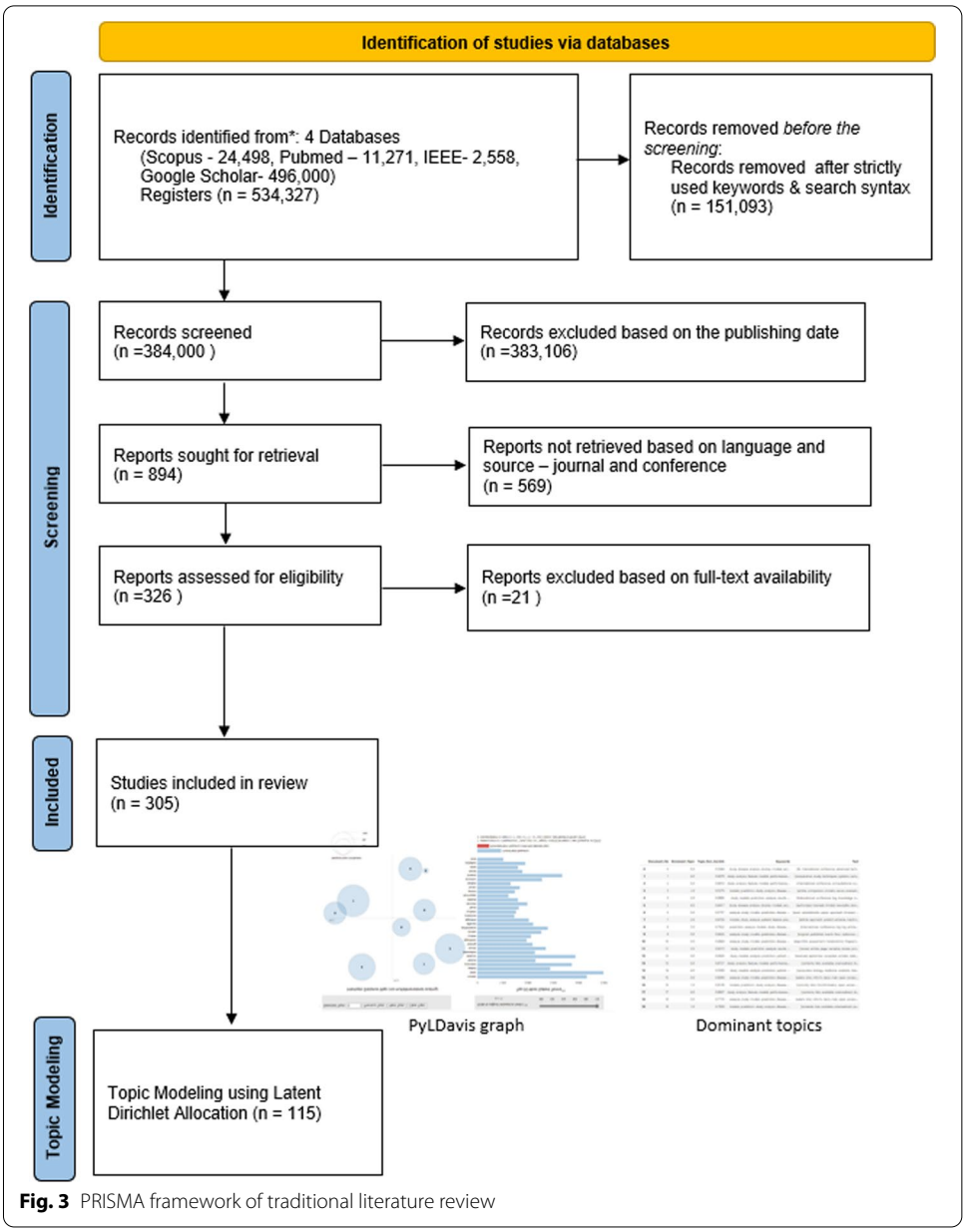

Fig. 3 PRISMA framework of traditional literature review

Topic modelling for intelligent literature review

Our intelligent literature review is developed using a combination of traditional literature review and topic modeling [22]. We use topic modeling—probability generating, a text-mining technique widely used in computer science for text mining and data recovery. Topic modeling is used in numerous papers to analyze [1, 5, 17, 36] and use various ML algorithms [38] such as Latent Semantic Indexing (LSI), Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA), Non-Negative Matrix Factorization (NMF), Parallel Latent Dirichlet Allocation (PLDA), and Pachinko Allocation Model (PAM). We developed the LDA-based methodological framework so it would be most widely and easily used [13, 17, 21] as a very elementary [6] approach. LDA is an unsupervised and probabilistic ML algorithm that discovers topics by calculating patterns of word

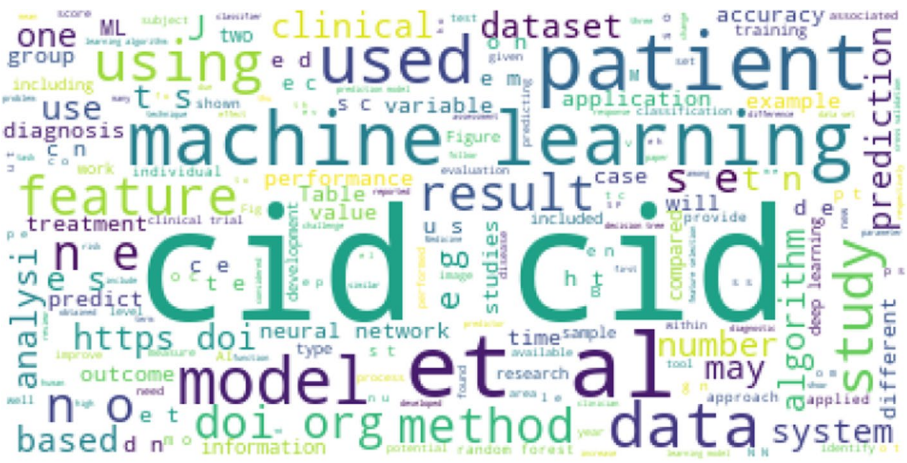

Fig. 4 Word cloud on 305 articles

co-occurrence across many documents or corpus [16]. Each LDA topic is distributed across each document as a probability.

While there are numerous ways of conducting a systematic literature review, most strategies require a high expense of time and prior knowledge of the area in advance. This study examined the expense of various text categorization strategies, where the assumptions and cost of the strategy are analyzed [5]. Interestingly, except manually reading the articles and topic modeling, all the strategies require prior knowledge of the articles' categories and high pre-examination costs. However, topic modeling can be automated, alternate the utilization of researchers' time, demonstrating a perfect match for the utilization of topic modeling as a part of an Intelligent literature review. Topic modeling has been used in a few papers to categorize research papers presented in Table 5.

The articles/papers in the above table analyzed are speeches, web documents, web posts, press releases, and newspapers. However, none of those have developed the framework to perform traditional literature reviews from digital databases then use topic modeling to save time. However, this research points out the utilization of LDA in academics and explores four parameters—text pre-processing, model parameters selection, reliability, and validity [5]. Topic modeling identifies patterns of the repetitive word across a corpus of documents. Patterns of word co-occurrence are conceived as hidden 'topics' available in the corpus. First, documents must be modified to be machine-readable, with only their most informative features used for topic modeling. We modify documents in a three-stage process entailing pre-processing, topic modeling, and post-processing, as defined in Fig. 1 earlier.

The utilization of topic modeling presents an opportunity for researchers to use advanced technology for the literature review process. Topic modeling has been utilized online and requires many statistical skills, which not all researchers have. Therefore, we have shared the codes in GitHub with the default parameter for future researchers.

Table 5 Topic modeling applications

Author details	Data type	Topic modeling used	Intended aim	Data size
DiMaggio et al. [13]	Newspapers	LDA	Identifying concepts in news coverage	8000
Grimmer [18]	Press release	Own implemented method	To develop a model	24,000
Koltsova and Koltcov [21]	Web posts	LDA	Explore the political agenda for live journal	1,300,000
Maier et al. [25]	Web documents	LDA	Explore the validity and reliability of the LDA model	186,557 web documents
Quinn et al. [38]	Legislative Speech	Own implemented method	To develop a statistical learning model	118,000 speeches (70,000,000 words)

Pre-processing

Székely and Brocke [46] explained that pre-processing is a seven-step process which explored below and mentioned in Fig. 1 as part B:

(1) Load data—The text data file is imported using the python command.
(2) Optical character recognition—using word cloud, characters are recognized.
(3) Filtering non-English words—non-English words are removed.
(4) Document tokenization—Split the text into sentences and the sentences into words. Lowercase the words and remove punctuation.
(5) Text cleaning—the text has been cleaned using portstemmer.
(6) Word lemmatization—words in the third person are changed to the first person, and past and future verb tenses are changed into the present.
(7) Stop word removal—All stop words are removed.

Topic modelling using LDA

Several research articles have been selected to run LDA topic modeling, explained in Table 5. LDA model results present the coherence score for all the selected topics and a list of the most frequently used words for each.

Post-processing

The goal of the post-processing stage is to identify and label topics and topics relevant for use in the literature review. The result of the LDA model is presented as a list of topics and probabilities of each document (paper). The list is utilized to assign a paper to a topic by arranging the list by the highest probability for each paper for each topic. All the topics contain documents that are like each other. To reduce the risk of error in topic identification, a combination of inspecting the most frequent words for each topic and a paper view is used. After the topic review, it will present in the literature review.

Following the intelligent literature review, results of the LDA model should be approved or validated by statistical, semantic, or predictive means. Statistical validation defines the mutual information tests of result fit to model assumptions; semantics validation requires hand-coding to decide if the importance of specific words varies significantly and as expected with tasks to different topics which is used in the current study to validate LDA model result; and predictive validation refers to checking if events that ought to have expanded the prevalence of particular topic if out interpretations are right, did so [6, 21].

LDA defines that each word in each document comes from a topic, and the topic is selected from a set of keywords. So we have two matrices:

(1) $\Theta td = P(t|d)$ which is the probability distribution of topics in documents
(2) $\Phi wt = P(w|t)$, which is the probability distribution of words in topics

And, we can say that the probability of a word given document, i.e., P(w|d), is equal to:

$$\sum_{t \in T} p(w|t,d)p(t|d)$$

where T is the total number of topics; likewise, let's assume there are W keywords for all the documents.

If we assume conditional independence, we can say that

$$P(w|t,d) = P(w|t)$$

And hence P(w|d) is equal to

$$\sum_{t=1}^{T} p(\omega|t)p(t|d)$$

that is the dot product of Θtd and Φwt for each topic t.

Results

Our systematic literature review identified 305 research papers after performing a traditional literature review. After executing LDA topic modeling, only 115 articles show the relevancy with our topic "machine learning application in clinical domain'. The following stages present LDA topic modeling process.

Pre-processing

The 305 research papers were stacked into a Python environment then converted into a single text file. The seven steps have been carried out, described earlier in Pre-processing.

Topic modeling

The two main parameters of the LDA topic model are the dictionary (id2word)-dictionary and the corpus—doc_term_matrix. The LDA model is created by running the command:

```
# Creating the object for LDA model using gensim library
LDA = gensim.models.ldamodel.LdaModel
# Build LDA model
lda_model = LDA(corpus=doc_term_matrix, id2word = dictionary, num_topics=20, random_state=100,
chunksize = 1000, passes=50,iterations=100)
```

In this model, 'num_topics' = 20, 'chunksize' is the number of documents used in each training chunk, and 'passes' is the total number of training passes.

Firstly, the LDA model is built with 20 topics; each topic is represented by a combination of 20 keywords, with each keyword contributing a certain weight to a topic. Topics are viewed and interpreted in the LDA model, such as Topic 0, represented as below:

(0, '0.005*"analysis" + 0.005*"study" + 0.005*"models" + 0.004*"prediction" + 0.003*"disease" + 0.003*"performance" + 0.003*"different" + 0.003*"results" + 0.003*"patient" + 0.002*"feature" + 0.002*"system" + 0.002*"accuracy" + 0.002*"diagnosis" + 0.002*"classification" + 0.002*"studies" + 0.002*"medicine" + 0.002*"value" + 0.002*"approach" + 0.002*"variables" + 0.002*"review"'),

Our approach to finding the ideal number of topics is to construct LDA models with different numbers of topics as K and select the model with the highest coherence value. Selecting the 'K' value that denotes the end of the rapid growth of topic coherence ordinarily offers significant and interpretable topics. Picking a considerably higher value can provide more granular sub-topics if the 'K' selection is too large, which can cause the repetition of keywords in multiple topics.

Model perplexity and topic coherence values are −8.855378536321144 and 0.3724024189689453, respectively. To measure the efficiency of the LDA model is lower the perplexity, the better the model is. Topics and associated keywords were then examined in an interactive chart using the pyLDAvis package, which presents the topics are 20 and most salient terms in those 20 topics, but these 20 topics overlap each other as shown in Fig. 5, which means the keywords are repeated in these 20 topics and topics are overlapped, which means so decided to use num_topics = 9 and presented PyLDAvis Figure below. Each bubble on the left-hand side plot represents a topic. The bigger the bubble is, the more predominant that topic is. A decent topic will have a genuinely big, non-overlapping bubble dispersed throughout the graph instead of grouped in one quadrant. A topic model with many topics will typically have many overlaps, small-sized bubbles clustered in one locale of the graph, as shown in Fig. 6.

Each bubble addresses a generated topic. The larger the bubble, the higher percentage of the number of keywords in the corpus is about that topic which can be seen on the GitHub file. Blue bars address the general occurrence of each word in the corpus. If no topic is selected, the blue bars of the most frequently used words are displayed, as depicted in Fig. 6.

The further the bubbles are away from each other, the more various they are. For example, we can tell that topic 1 is about patient information and studies utilized deep learning to analyze the disease, which can be seen in GitHub file codes (https://github.com/MachineLearning-UON/Topic-modeling-using-LDA.git) and presented in Fig. 7.

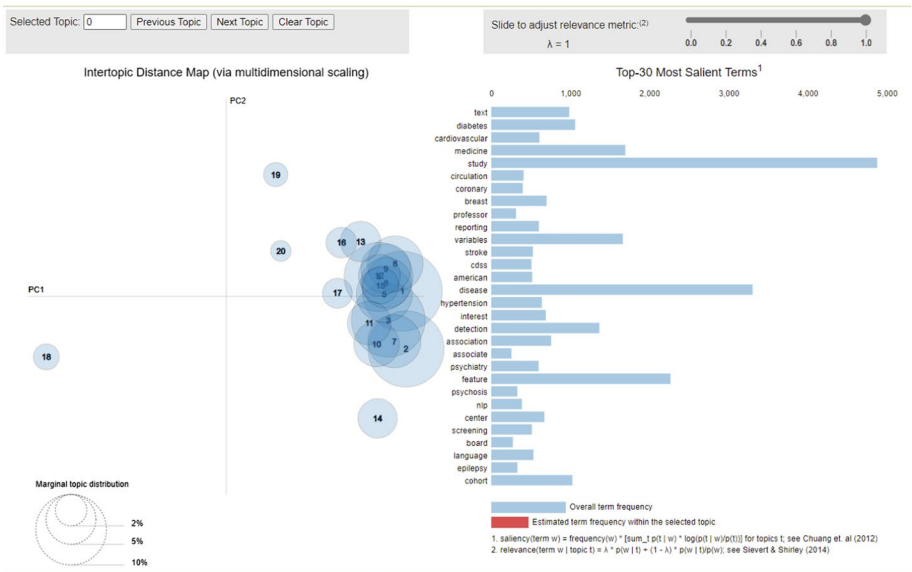

Fig. 5 PyLDAvis graph with 20 topics in the clinical domain

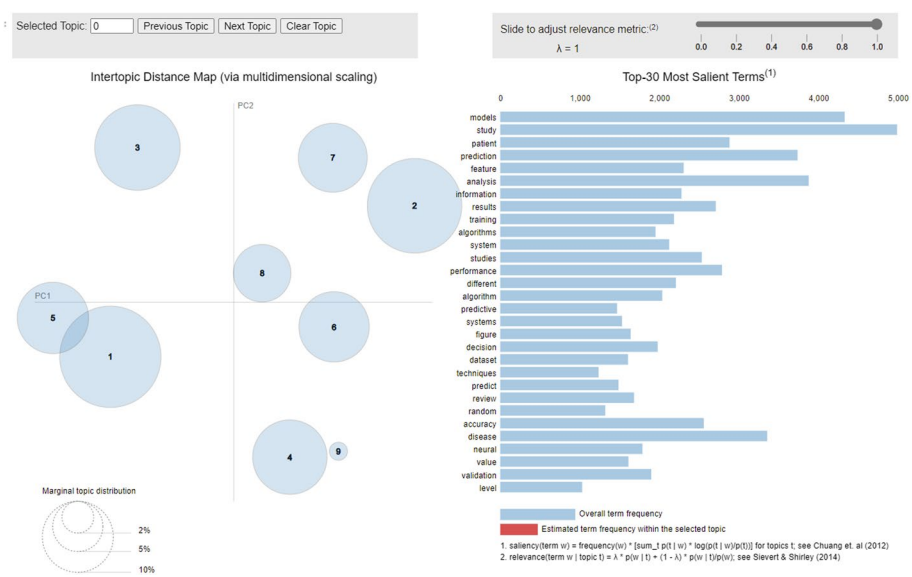

Fig. 6 PyLDAvis graph with nine vital topics in the clinical domain

Red bars give the assessed number of times a given topic produced a given term. As you can see from Fig. 7, there are around 4000 of the word 'analysis', and this term is utilized 1000 times inside topic 1. The word with the longest red bar is the most used by the keywords having a place with that topic.

A good topic model will have big and non-overlapping bubbles dispersed throughout the chart. As we can see from Fig. 6, the bubbles are clustered within one place. One of the practical applications of topic modeling is discovering the topic in a provided

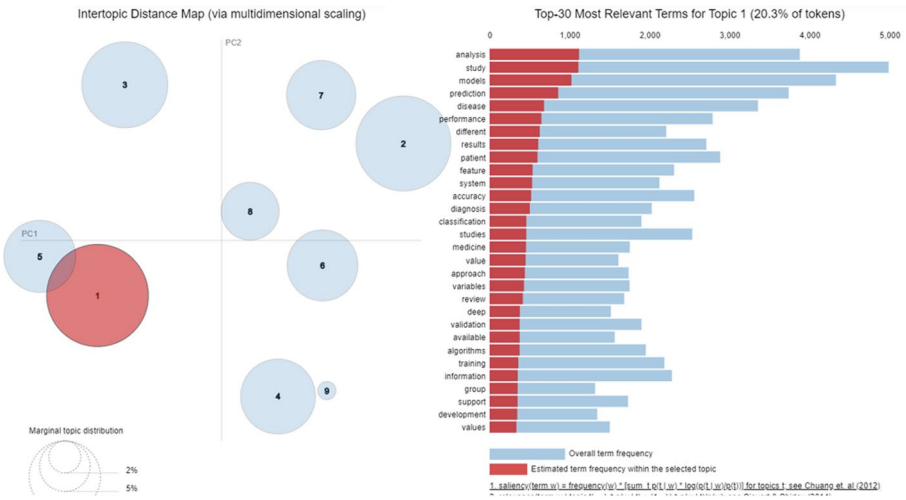

Fig. 7 PyLDAvis graph with topic 1

document. We find out the topic number with the highest percentage contribution in that document, as shown in Fig. 8.

Post-processing

The next stage is to process the discoveries and find a satisfactory depiction of the topics. A combination of evaluating the most continuous words utilized to distinguish the topic. For example, the most frequent words for the papers in topic 2 are "study" and "analysis", which indicate frequent words for ML usage in the clinical domain.

The topic name is displayed with the topic number from 0 to 8, which represents in the Table 6, which includes the Topic number and Topic words.

The result represents the percentage of the topics in all documents, which presents that topic 0 and topic 6 have the highest percentage and used in 58 and 57 documents, respectively, with 115 papers. The result of this research was an overview of the exploration areas inside the paper corpus, addressed by 9 topics.

Discussion

This paper presented a new methodology that is uncommon in scholarly publications. The methodology utilizes ML to investigate sample articles/papers to distinguish research directions. Even though the structure of the ML-based methodology has its restrictions, the outcomes and its ease of use leave a promising future for topic modeling-based systematic literature reviews.

The principal benefit of the methodological framework is that it gives information about an enormous number of papers, with little effort on the researcher's part, before time-exorbitant manual work is to be finished. By utilizing the framework, it is conceivable to rapidly explore a wide range of paper corpora and assess where the researcher's time and concentration should be spent. This is particularly significant for a junior researcher with minimal earlier information on a research field. If default boundaries and cleaning settings can be found for the steps in the framework, a completely programmed

	Document_No	Dominant_Topic	Topic_Perc_Contrib	Keywords	Text
0	0	8.0	0.5048	study, disease, analysis, studies, models, acc...	[th, international, conference, advanced, tech...
1	1	6.0	0.6078	study, analysis, feature, models, performance,...	[comparative, study, techniques, systems, cont...
2	2	6.0	0.9014	study, analysis, feature, models, performance,...	[international, conference, computational, sci...
3	3	1.0	0.5276	models, prediction, study, analysis, disease, ...	[article, comparison, models, versus, evaluati...
4	4	2.0	0.9884	study, models, prediction, analysis, results, ...	[international, conference, big, knowledge, ic...
5	5	8.0	0.6417	study, disease, analysis, studies, models, acc...	[authorized, licensed, limited, newcastle, dow...
6	6	0.0	0.5757	analysis, study, models, prediction, disease, ...	[open, amendments, paper, approach, forecast, ...
7	7	3.0	0.4755	models, study, analysis, patient, feature, pre...	[article, approach, predict, extreme, inactivi...
8	8	5.0	0.7622	prediction, analysis, models, study, disease, ...	[international, conference, big, big, article,...
9	9	0.0	0.6624	analysis, study, models, prediction, disease, ...	[original, published, march, fonc, radiomics, ...
10	10	0.0	0.6840	analysis, study, models, prediction, disease, ...	[algorithm, assessment, metabolomic, fingerpri...
11	11	2.0	0.6315	study, models, prediction, analysis, results, ...	[review, article, page, narrative, review, pro...
12	12	4.0	0.4534	study, models, analysis, prediction, patient, ...	[received, september, accepted, october, date,...
13	13	6.0	0.8737	study, analysis, feature, models, performance,...	[contents, lists, available, sciencedirect, bi...
14	14	4.0	0.5580	study, models, analysis, prediction, patient, ...	[computers, biology, medicine, contents, lists...
15	15	0.0	0.8048	analysis, study, models, prediction, disease, ...	[salem, bmc, inform, decis, mak, open, access,...
16	16	1.0	0.8108	models, prediction, study, analysis, disease, ...	[connolly, bmc, bioinformatics, open, access, ...
17	17	6.0	0.8837	study, analysis, feature, models, performance,...	[contents, lists, available, sciencedirect, bi...
18	18	0.0	0.7735	analysis, study, models, prediction, disease, ...	[salem, bmc, inform, decis, mak, open, access,...
19	19	1.0	0.7828	models, prediction, study, analysis, disease, ...	[contents, lists, available, sciencedirect, ps...

Fig. 8 Dominant topics with topic percentage contribution

Table 6 The number of topics and topic words in the LDA result

Topic number	Topic words
0	Analysis, study, models, prediction, disease, performance, different, results, patient, feature
1	Models, prediction, study, analysis, disease, results, patient, performance, studies, accuracy
2	Study, models, prediction, analysis, results, disease, performance, feature, neural, accuracy
3	Models, study, analysis, patient, feature, prediction, system, studies, results, training
4	Study, models, analysis, prediction, patient, disease, information, accuracy, training, results
5	Prediction, analysis, models, study, disease, patient, information, results, validation, training
6	Study, analysis, feature, models, performance, prediction, disease, studies, patient, accuracy
7	Study, models, disease, patient, prediction, results, performance, analysis, accuracy, algorithms
8	Study, disease, analysis, studies, models, accuracy, prediction, performance, decision, algorithm

gathering of papers could be empowered, where limited works have been introduced to accomplish an overview of research directions.

From a literature review viewpoint, the advantage of utilizing the proposed framework is that the inclusion and exclusion selection of papers for a literature review will be delayed to a later stage where more information is given, resulting in a more educated dynamic interaction. The framework empowers reproducibility, as every step can be reproduced in the systematic review process that ultimately empowers with transparency. The whole process has been demonstrated as a case concept on GitHub by future researchers.

The study has introduced an intelligent literature review framework that uses ML to analyze existing research documents or articles. We demonstrate how topic modeling can assist literature review by reducing the manual screening of huge quantities of literature for more efficient use of researcher time. An LDA algorithm provides default

parameters and data cleaning steps, reducing the effort required to review literature. An additional advantage of our framework is that the intelligent literature review offers accurate results with little time, and it comprises traditional ways to analyze literature and LDA topic modeling.

This framework is constructed in a step-by-step manner. Researchers can use it efficiently because it requires less technical knowledge than other ML algorithms. There is no restriction on the quantity of the research papers it can measure. This research extends knowledge to similar studies in this field [12, 22, 23, 26, 30, 46] which present topic modeling. The study acknowledges the inspiring concept of smart literature defined by Asmussen and Møller [3]. The researchers previously provided a brief description of how LDA is utilized in topic modeling. Our research followed the basic idea but enhanced its significance to broaden its scale and focus on a specific domain such as the clinical domain to produce insights from existing research articles. For instance, Székely and Vom [46] utilized natural language processing to analyze 9514 sustainability reports published between 1999 and 2015. They identified 42 topics but did not develop any framework for future researchers. This was considered a significant gap in the research. Similarly, Kushwaha et al. [22] used a network analysis approach to analyze 10-year papers without providing any clear transparent outcome (e.g., how the research step-by-step produces an outcome). Likewise, Asmussen and Møller [3] developed a smart literature review framework that was limited to analyzing 650 sample articles through a single method. However, in our research, we developed an intelligent literature review that combines traditional and LDA topic modeling, so that future researchers can get assistance to gain effective knowledge regarding literature review when it becomes a state-of-the-art in research domains.

Our research developed a more effective intelligent framework, which combines traditional literature review and topic modeling using LDA, which provides more accurate and transparent results. The results are shared via public access on GitHub using this link https://github.com/MachineLearning-UON/Topic-modeling-using-LDA.git.

Conclusion

This paper focused on creating a methodological framework to empower researchers, diminishing the requirement for manually scanning documents and assigning the possibility to examine practically limitless. It would assist in capturing insights of an enormous number of papers quicker, more transparently, with more reliability. The proposed framework utilizes the LDA's topic model, which gathers related documents into topics.

A framework employed topic modeling for rapidly and reliably investigating a limitless number of papers, reducing their need to read individually, is developed. Topic modeling using the LDA algorithm can assist future researchers as they often need an outline of various research fields with minimal pre-existing knowledge. The proposed framework can empower researchers to review more papers in less time with more accuracy. Our intelligent literature review framework includes a holistic literature review process (conducting, planning, and reporting the review) and an LDA topic modeling (pre-processing, topic modeling, and post-processing stages), which conclude the results of 115 research articles are relevant to the search.

The automation of topic modeling with default parameters could also be explored to benefit non-technical researchers to explore topics or related keywords in any problem domain. For future directions, the principal points should be addressed. Future researchers in other research fields should apply the proposed framework to acquire information about the practical usage and gain ideas for additional advancement of the framework. Furthermore, research in how to consequently specify model parameters could extraordinarily enhance the ease of use for the utilization of topic modeling for non-specialized researchers, as the determination of model parameters enormously affects the outcome of the framework.

Future research may be utilized more ML analytics tools as complete solution artifacts to analyze different forms of big data. This could be adopting design science research methodologies for benefiting design researchers who are interested in building ML-based artifacts [15, 28, 29, 31–33].

Abbreviations
IEEE: The Institute of Electrical and Electronics Engineers; ML: Machine learning; LDA: Latent Dirichlet Allocation; OC: Organizational Capacity; LSI: Latent Semantic Indexing; LSA: Latent Semantic Analysis; NPF: Non-Negative Matrix Factorization; PLDA: Parallel Latent Dirichlet Allocation; PAM: Pachinko Allocation Model.

Acknowledgements
Not applicable.

Author contributions
The first author conducted the research, while the second author has ensured quality standards and rewritten the entire findings linking to underlying theories. Both authors read and approved the final manuscript.

Funding
Not applicable.

Availability of data and materials
Data will be supplied upon request.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
Not applicable.

Received: 18 November 2021 Accepted: 6 April 2022
Published online: 28 April 2022

References

1. Abuhay TM, Kovalchuk SV, Bochenina K, Mbogo G-K, Visheratin AA, Kampis G, et al. Analysis of publication activity of computational science society in 2001–2017 using topic modelling and graph theory. J Comput Sci. 2018;26:193–204.
2. Adlung L, Cohen Y, Mor U, Elinav E. Machine learning in clinical decision making. Med. 2021;2(6):642–65.
3. Asmussen CB, Møller C. Smart literature review: a practical topic modeling approach to exploratory literature review. J Big Data. 2019;6(1):1–18.
4. Beck CT. A meta-synthesis of qualitative research. MCN Am J Mater Child Nurs. 2002;27(4):214–21.
5. Behera RK, Bala PK, Dhir A. The emerging role of cognitive computing in healthcare: a systematic literature review. Int J Med Informatics. 2019;129:154–66.
6. Blei DM. Probabilistic topic models. Commun ACM. 2012;55(4):77–84.
7. Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. J Mach Learn Res. 2003;3:993–1022.
8. Brnabic A, Hess LM. Systematic literature review of machine learning methods used in the analysis of real-world data for patient-provider decision making. BMC Med Inform Decis Mak. 2021;21(1):1–19.
9. Cabitza F, Locoro A, Banfi G. Machine learning in orthopedics: a literature review. Front Bioeng Biotechnol. 2018;6:75.
10. Chang C-H, Lin C-H, Lane H-Y. Machine learning and novel biomarkers for the diagnosis of Alzheimer's disease. Int J Mol Sci. 2021;22(5):2761.

11. Connor KL, O'Sullivan ED, Marson LP, Wigmore SJ, Harrison EM. The future role of machine learning in clinical transplantation. Transplantation. 2021;105(4):723–35.
12. Dias R, Torkamani A. Artificial intelligence in clinical and genomic diagnostics. Genome Med. 2019;11(1):1–12.
13. DiMaggio P, Nag M, Blei D. Exploiting affinities between topic modeling and the sociological perspective on culture: application to newspaper coverage of US government arts funding. Poetics. 2013;41(6):570–606.
14. Forest P-G, Martin D. Fit for Purpose: Findings and recommendations of the external review of the Pan-Canadian Health Organizations: Summary Report: Health Canada Ottawa, ON; 2018.
15. Genemo H, Miah SJ, McAndrew A. A design science research methodology for developing a computer-aided assessment approach using method marking concept. Educ Inf Technol. 2016;21(6):1769–84.
16. Greene D, Cross JP. Exploring the political agenda of the european parliament using a dynamic topic modeling approach. Polit Anal. 2017;25(1):77–94.
17. Grimmer J. A Bayesian hierarchical topic model for political texts: measuring expressed agendas in Senate press releases. Polit Anal. 2010;18(1):1–35.
18. Grimmer J, Stewart BM. Text as data: the promise and pitfalls of automatic content analysis methods for political texts. Polit Anal. 2013;21(3):267–97.
19. Hassan N, Slight R, Weiand D, Vellinga A, Morgan G, Aboushareb F, et al. Preventing sepsis; how can artificial intelligence inform the clinical decision-making process? A systematic review. Int J Med Inform. 2021;150:104457.
20. Hirt R, Koehl NJ, Satzger G, editors. An end-to-end process model for supervised machine learning classification: from problem to deployment in information systems. Designing the Digital Transformation: DESRIST 2017 Research in Progress Proceedings of the 12th International Conference on Design Science Research in Information Systems and Technology Karlsruhe, Germany 30 May-1 Jun; 2017: Karlsruher Institut für Technologie (KIT).
21. Koltsova O, Koltcov S. Mapping the public agenda with topic modeling: the case of the Russian live journal. Policy Internet. 2013;5(2):207–27.
22. Kushwaha AK, Kar AK, Dwivedi YK. Applications of big data in emerging management disciplines: a literature review using text mining. Int J Inf Manag Data Insights. 2021;1(2):100017.
23. Li S, Wang H. Traditional literature review and research synthesis. The Palgrave handbook of applied linguistics research methodology. 2018:123–44.
24. Magrabi F, Ammenwerth E, McNair JB, De Keizer NF, Hyppönen H, Nykänen P, et al. Artificial intelligence in clinical decision support: challenges for evaluating AI and practical implications. Yearb Med Inform. 2019;28(01):128–34.
25. Maier D, Waldherr A, Miltner P, Wiedemann G, Niekler A, Keinert A, et al. Applying LDA topic modeling in communication research: toward a valid and reliable methodology. Commun Methods Meas. 2018;12(2–3):93–118.
26. Mårtensson G, Ferreira D, Granberg T, Cavallin L, Oppedal K, Padovani A, et al. The reliability of a deep learning model in clinical out-of-distribution MRI data: a multicohort study. Med Image Anal. 2020;66:101714.
27. Mendo IR, Marques G, de la Torre DI, López-Coronado M, Martín-Rodríguez F. Machine learning in medical emergencies: a systematic review and analysis. J Med Syst. 2021;45(10):1–16.
28. Miah SJ. An ontology based design environment for rural business decision support. Nathan: Griffith University Nathan; 2008.
29. Miah SJ, A new semantic knowledge sharing approach for e-government systems. 4th IEEE International Conference on Digital Ecosystems and Technologies; 2010: IEEE.
30. Miah SJ, Camilleri E, Vu HQ. Big Data in healthcare research: a survey study. J Comput Inf Syst. 2021. https://doi.org/10.1080/08874417.2020.1858727.
31. Miah SJ, Gammack J, Kerr D, Ontology development for context-sensitive decision support. Third International Conference on Semantics, Knowledge and Grid (SKG 2007); 2007: IEEE.
32. Miah SJ, Gammack JG. Ensemble artifact design for context sensitive decision support. Australas J Inf Syst. 2014. https://doi.org/10.3127/ajis.v18i2.898.
33. Miah SJ, Gammack JG, McKay J. A metadesign theory for tailorable decision support. J Assoc Inf Syst. 2019;20(5):4.
34. Mimno D, Blei D, editors. Bayesian checking for topic models. Proceedings of the 2011 conference on empirical methods in natural language processing; 2011.
35. Oala L, Murchison AG, Balachandran P, Choudhary S, Fehr J, Leite AW, et al. Machine learning for health: algorithm auditing & quality control. J Med Syst. 2021;45(12):1–8.
36. Ouhbi S, Idri A, Fernández-Alemán JL, Toval A. Requirements engineering education: a systematic mapping study. Requir Eng. 2015;20(2):119–38.
37. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2020;372:n71.
38. Quinn KM, Monroe BL, Colaresi M, Crespin MH, Radev DR. How to analyze political attention with minimal assumptions and costs. Am J Polit Sci. 2010;54(1):209–28.
39. Rowley J, Slack F. Conducting a literature review. Management research news. 2004.
40. Rozas LW, Klein WC. The value and purpose of the traditional qualitative literature review. J Evid Based Soc Work. 2010;7(5):387–99.
41. Sabharwal R, Miah SJ. A new theoretical understanding of big data analytics capabilities in organizations: a thematic analysis. J Big Data. 2021;8(1):1–17.
42. Salazar-Reyna R, Gonzalez-Aleu F, Granda-Gutierrez EM, Diaz-Ramirez J, Garza-Reyes JA, Kumar A. A systematic literature review of data science, data analytics and machine learning applied to healthcare engineering systems. Management Decision. 2020.
43. Shah P, Kendall F, Khozin S, Goosen R, Hu J, Laramie J, et al. Artificial intelligence and machine learning in clinical development: a translational perspective. NPJ Digit Med. 2019;2(1):1–5.
44. Sone D, Beheshti I. Clinical application of machine learning models for brain imaging in epilepsy: a review. Front Neurosci. 2021;15:761.
45. Spasic I, Nenadic G. Clinical text data in machine learning: systematic review. JMIR Med Inform. 2020;8(3):e17984.

46. Székely N, Vom Brocke J. What can we learn from corporate sustainability reporting? Deriving propositions for research and practice from over 9,500 corporate sustainability reports published between 1999 and 2015 using topic modelling technique. PLoS ONE. 2017;12(4):e0174807.
47. Verma D, Bach K, Mork PJ, editors. Application of machine learning methods on patient reported outcome measurements for predicting outcomes: a literature review. Informatics; 2021: Multidisciplinary Digital Publishing Institute.
48. Weng W-H. Machine learning for clinical predictive analytics. Leveraging data science for global health. Cham: Springer; 2020. p. 199–217.
49. Yin Z, Sulieman LM, Malin BA. A systematic literature review of machine learning in online personal health data. J Am Med Inform Assoc. 2019;26(6):561–76.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen journal and benefit from:

► Convenient online submission
► Rigorous peer review
► Open access: articles freely available online
► High visibility within the field
► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com