JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGY
http://e-journal.uum.edu.my/index.php/jict

How to cite this article:
Ismail, N., & Yusof, U. K. (2022). Recent trends of machine learning predictions using open data: A systematic review. *Journal of Information and Communication Technology, 21*(3), 337-381. https://doi.org/10.32890/jict2022.21.3.3

Recent Trends of Machine Learning Predictions using Open Data: A Systematic Review

*[1]Norismiza Ismail & [2]Umi Kalsom Yusof
[1&2]School of Computer Sciences,
Universiti Sains Malaysia, Malaysia
[1]Digital Management and Development Centre,
Universiti Malaysia Perlis, Malaysia

*norismiza@unimap.edu.my,
umiyusof@usm.my
*Corresponding author

Received: 30/1/2022 Revised: 8/4/2022 Accepted: 11/4/2022 Published: 17/7/2022

ABSTRACT

Machine learning (ML) prediction determinants based on open data (OD) are investigated in this work, which is accomplished by examining current research trends over ten years. Currently, OD is commonly regarded as the most crucial trend for users to improve their ability to make decisions, particularly to the exponential expansion of social networking sites (SNSs) and open government data (OGD). The purpose of this study was to examine if there was an increase in the usage of OD in ML prediction techniques by conducting a systematic literature review (SLR) of the results of the trends. The papers published in major online scientific databases between 2011 and 2020, including ScienceDirect, Scopus, IEEE Xplore, ACM, and Springer, were identified and analysed. After various selection processes, according to SLR based on precise inclusion and exclusion criteria, a total of 302 articles were located. However, only 81 of them were included. The findings were presented and plotted based on the research questions (RQs). In conclusion, this research could be beneficial to organisations, practitioners, and researchers by providing information on current trends in the implementation of ML prediction using OD setting by mapping studies based on the RQs designed, the most recent growth, and the necessity for future research based on the findings.

Keywords: Machine learning, open data, prediction, systematic literature review.

INTRODUCTION

The proliferation of open data (OD) has resulted in a new generation of open datasets that are reusable, accessible, sustainable, and interoperable, exploring the possibility for OD principles to be implemented globally, and allowing multiple modules, frameworks, and organisations to collaborate (OKFN, 2014; OD, 2012; W3C, 2009). Since 2009, the open government data (OGD) movement has grown dramatically, when the United States (US) government committed to implementing the principle of openness by publishing millions of datasets initiated by Barack Obama, the former US President (Saxena, 2019). Later, the European Commission, Mexico, and Singapore opened the floodgates of publicly available information (Foulonneau et al., 2014). All stakeholders from a range of social, economic, environmental, and other backgrounds can benefit from the dataset's access, use, and interchange thanks to the creation of the OGD platform. Moreover, emerging OGDs are advantageous to these sectors as well as for scholarly debates, particularly in the context of service (MAMPU, 2017; Lindman et al., 2014). Web 2.0-based technologies, such as downloading raw data, using a transparent application programming interface (API), and accessing linked open data (LOD), are all options that have been used (Song et al., 2013).

Recently, a rising amount of user-generated content (UGC), such as reviews, commentaries, and previous experiences, in addition to OGD, provided through social networking services (SNSs) has made much of the OD material accessible (Pantano et al., 2017). Based on word-of-mouth communications and decision-making processes, SNSs have a significant impact (Chu & Kim, 2011), and users' interests must be successfully drawn and exploited. Besides, digital marketers are conscious that they must improve the usability of SNS by providing value-added services (Diffley et al., 2011). As a result, to satisfy the expectations of new social media experiences, social media operators are developing new capabilities by delivering a varied array of built-in applications (Jai et al., 2014) and personalised topic-specific virtual environments (e.g., Instagram, YouTube, Facebook, Twitter, and LinkedIn) to provide better UGC by incorporating comments, updates on prior experiences, and recommendations for future content (Turban et al., 2015).

Furthermore, the machine learning (ML) approach aims to learn unknown data concepts. OD has been implemented to forecast various attitudes or behaviours in decision-making processes in several studies using ML methodologies. As an example, OD was utilised to assist a traveller's procedure for deciding by profiling elements of various tourism locations throughout the world and locales using the Random Forest (RF) method of the ML technique (Pantano et al., 2017). Therefore, the goal of this paper is to review the latest ten-year OD-related articles to gain a general understanding of prediction using the ML method and to map the existing studies based on the designed research questions (RQs) through a systematic literature review (SLR). More precisely, the aim is to educate stakeholders about the current trends and practices and the bibliometric knowledge of the published articles in the field of prediction in OD and ML.

Several studies on prediction using ML are currently being performed, but none of them are directly relevant to the field of OD. Due to the rapid growth of OGD and SNSs, OD is now the most relevant trend for practitioners seeking to develop their prediction process. Nevertheless, a further study utilising SLR should be carried out to observe how well OD studies can predict behaviour related to a given interest using ML based on performance indicators. Theoretically, the goal of adopting SLR is to organise and summarise the current ten-year patterns in open datasets, which might greatly aid in prediction using various ML approaches and algorithms. This trend analysis will also present possible research gaps and challenges that will help other practitioners and researchers in this field.

An Overview of the Related Studies

Since there is no specific analysis of the recent trends in the ML prediction of using OD, reviews on related studies are presented in Table 1. The table depicts the contributions of six research studies (three systematic reviews and three other forms of reviews), addressing various aspects of OD using the predictive ML approach. These reviewed articles mainly focused on the novelty of open datasets in ML technique prediction. The evidence presents the current state and existing trends, including future research. The literature summarises the current state-of-the-art predictions using various ML approaches. The articles addressed issues and challenges, which show potential gaps and future directions.

Table 1

Contribution of Previous Review Studies on OD in ML Prediction

Ref.	Title	Year	Type	Contributions
Iskandaryan et al. (2020)	Air Quality Prediction in Smart Cities using ML Technologies based on Sensor Data: A Review	2020	Systematic Review 2002–2019 (41 articles)	Reviewed current approach to prediction concept Quantitative analysis and trends of evidence were presented. Highlighted prediction techniques, open dataset characteristics, and performance metrics analysis.
Butt et al. (2020)	Spatio-Temporal Crime Hot Spot Detection and Prediction: A Systematic Literature Review	2020	Systematic Review 2010–2019 (49 articles)	Demonstrated quantitative analysis and trends evidence. Highlighted prediction and detection techniques, dataset characteristics, and performance measurement analysis. Showed potential gaps, challenges, and future research direction.
Goldstein et al. (2019)	A Review of ML Applications to Coastal Sediment Transport and Morphodynamics	2019	Review	Evaluated the implementation of ML in experiments on supervised regression tasks. Described a selection of best practices for using ML techniques. Suggested potential areas for future study, the use of new ML methods, and open data exploration.

(continued)

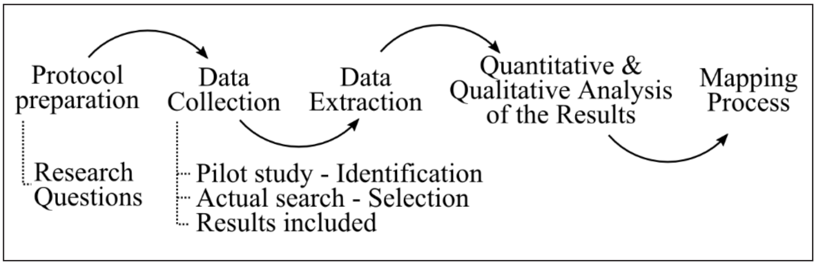
Ref.	Title	Year	Type	Contributions
Gutierrez-Osorio and Pedraza (2020)	Modern Data Sources and Techniques for Analysis and Forecast of Road Accidents: A Review	2019	Review	Provided an overview of the state-of-the-art prediction through ML algorithms and advanced information analysis techniques. Proposed a classification of ML according to its origin and characteristics. Suggestions on how to improve precision and accuracy.
Tamada et al. (2019)	Predicting and Reducing Dropout in Virtual Learning using Machine Learning Techniques: A Systematic Review	2019	Systematic Review 2015–2018 (199 articles)	Demonstrate quantitative analysis and trends in ML. Highlighted the evolution of publications on ML techniques used, dataset characteristics, and proposed solutions. Provided guide for future studies and tool development.
Al-Garadi et al. (2019)	Predicting Cyberbullying on Social Media in the Big Data Era using Machine Learning Algorithms: Review of Literature and Open Challenges	2019	Review	Reviewed prediction models and issues. Emphasis on features of algorithm selection and using various ML algorithms for prediction. Highlighted issues and challenges.

METHODOLOGY

The systematic literature review (SLR) method was employed to gain access to a large number of possible publications and to gain a thorough understanding of the literature in numerous research streams (Kitchenham & Charters, 2007; Bizer et al., 2011). SLR is effective in researching and providing a research area's overview in the sense of OD prediction, displaying the quantity of proof, and generating specific research facts. The SLR's findings aid in identifying research priorities within the field required. In general, the review procedure follows the steps outlined in Figure 1 (Davis et al., 2006; Maglyas et al., 2011). The results of the above method are known after the study has been completed and all the findings have been published.

Figure 1

The Study Selection Processes


Protocol preparation
Research Questions
Data Collection
Pilot study - Identification
Actual search - Selection
Results included
Data Extraction
Quantitative & Qualitative Analysis of the Results
Mapping Process

The publications were reviewed twice, as according to Budgen et al. (2011), this ensures consistency in the process of inclusion and exclusion of the papers (Budgen et al., 2011). The first round of analysis was performed to identify the study's specific topic using titles, abstracts, and keywords (Yin, 2013), based on research questions (RQs), excluding non-related studies. In the second phase, the entire texts of the papers were scrutinised, and non-related publications were once again discarded. Consequently, any fresh and relevant information of the RQs was gathered. The above-mentioned articles were thoroughly reviewed and fine-tuned as necessary. To determine the current status and trend, the mapping approach was used.

Research Questions (RQs)

The SLR's main objective is to identify all applicable studies for the RQs in light of Table 2's criteria. The RQs were then divided into two categories: bibliometric research questions (BRQs) and content research questions (CRQs) (Sadoughi et al., 2020).

Table 2

Research Questions Criteria

Open Data	Type of datasets that have been used
Machine Learning	Techniques/methods that have been implemented
Prediction	Accuracy of prediction and effective predicting techniques
Research Novelty	Potential research gaps, limitations, and challenges

Bibliometric Research Questions (BRQs)

To guide in the search and results presentation, the following RQs were investigated using the chosen papers on:

1. How many prediction papers have been published in the fields of OD and ML?
2. How has the trend changed over time?

Content Research Questions (CRQs)

Following the determination of the BRQs, a more extensive investigation of the publications' complete text was necessary to respond to the RQs below:

1. What are the ML approaches for a prediction that have been reported in existing OD research?
2. What are the accuracies or performance measures of the predictions when using OD and ML?
3. In this analysis, what were the characteristics of the open datasets used?
4. From studies related to the development of a robust prediction model, what are the possible challenges and study gaps highlighted?

Data Collection

The findings of the literature review were strongly influenced by keywords and the digital databases used in performing the search (Kitchenham & Charters, 2007). The articles were obtained from the selected databases using the search strategy to answer the RQs created.

Selection of Database and Search Queries

The inquiry began with a preliminary search focused on the nature of OD and prediction on 1st June 2020, utilising Google Scholar to locate keywords and develop an understanding of both available and crucial papers. Google Scholar was chosen to deliver scholarly literature metadata or full-text indexes (journal articles, conference papers, and workshops) (Halevi et al., 2017) as shown in Table 3 because of its usability as a web search engine and citations monitoring tool in the majority of online peer-reviewed journals.

Table 3

Preliminary Search Results of Articles Found by Using Google Scholar on 1st June 2020

Google Scholar	
Search Keywords	No. of Articles
Open Data AND prediction AND Machine Learning	2,700,000
“Open Data” AND “prediction” AND “Machine Learning”	18,200
“Open Data” AND (“prediction” OR “predict*” OR “forecast*”) AND “Machine Learning”	23,500

However, a brief check revealed that the phrases “predict” and “forecast” were equivalent to “prediction” and were used in some of the literature after several repetitions of combining and searching particular keywords. According to previous assessments, the most useful databases in the computer science (CS) and information technology (IT) fields are IEEE, ACM, and ScienceDirect (Bizer et al., 2011). The reason for choosing IEEE was that it is a significant organisation for advanced technology excellence (Madarash-Hill & Hill, 2004), while ACM is still the world’s largest CS database (Zelevinsky et al., 2008). Scopus was chosen in the meantime because it provides access to the world’s abstract literature and citation database with the most peer-reviewed abstracts and a complete overview of research output (Boyle & Sherman, 2006), with Springer serving as a digital database supplement. In conclusion, to find relevant articles, the search strategy comprised the decisions as shown in Table 4, and database searches were performed using titles, keywords, and abstracts as mentioned in Table 5. Table 6, based on a total of 302 ML prediction-related articles using OD publications, shows the distribution of articles from particular digital databases, with Scopus yielding the most results throughout the search procedure. Searches done within the ScienceDirect and ACM databases, meanwhile, provided the lowest results with 16 and 15 articles, respectively.

Table 4

Search Strategy Decisions

Criteria	Description
Databases	Scopus (https://www.scopus.com/), ScienceDirect (https://www.sciencedirect.com/), IEEE Xplore (https://ieeexplore.ieee.org/), Springer (https://link.springer.com/), and ACM (https://dl.acm.org/)
Items	Journal papers, conference papers, magazines, and workshops
Search applied on	Full-text papers, within article or document title, keywords, and abstract
Publication period	Between January 2011 and December 2020

Table 5

Search Keywords for Each Database

Databases	Search Keywords
Scopus	TITLE-ABS-KEY (“Open Data” AND (“prediction” OR “predict*” OR “forecast*”) AND “Machine Learning”)
ScienceDirect	Title, abstract, keywords: (“Open Data” AND (“prediction” OR “predict” OR “forecast”) AND “Machine Learning”)
IEEE Xplore	“Open Data” AND (“prediction” OR “predict*” OR “forecast*”) AND “Machine Learning”
Springer	“Open Data” AND (“prediction” OR “predict*” OR “forecast*”) AND “Machine Learning”
ACM	“Open Data” AND (“prediction” OR “predict*” OR “forecast*”) AND “Machine Learning”

Table 6

Publications Distribution (n = 302)

Scientific Databases	No. of Articles
Scopus (http://scopus.com)	175
Springer (https://link.springer.com/)	59
IEEE Xplore (http://ieeexplore.ieee.org/Xplore/home.jsp)	37
ScienceDirect (http://www.sciencedirect.com)	16
ACM (http://dl.acm.org)	15

Criteria for Study Selection

It was discovered that the number of articles collected (n = 302) using the search terms was quite large and that some were duplicate articles and were the same articles found in different databases. Therefore, they were subsequently removed and thus resulting in a final number of 240 relevant articles. From these journals, the abstracts of the articles were then reviewed, and 144 articles remained after removing concerns, which were irrelevant based on Table 7's list of criteria for inclusion and exclusion.

Table 7

Criteria for Inclusion and Exclusion

Inclusion Criteria	Exclusion Criteria
• Include primary research on the RQs. • Research articles or journal issue closely related to the topic of RQs. • Articles explaining "open data"AND "prediction"AND "machine learning." • Industry, government, and any academic research or study. • Full-text publications are available.	• Secondary studies (e.g., systematic literature, survey, review). • A copy of a research study that is identical to the original. • Publications that do not define OD, prediction, or machine learning. • Papers are written in languages other than English. • Articles on business (general business issue).

Results Included

The review's content was restricted based on the title, abstract, and availability of the papers. In other words, the selected papers were only approved after the complete texts had been checked and mapped systematically to the current study. After the title and abstract screening, 7 articles were eliminated for non-scholarly papers, and 56 articles were removed for not answering the RQs. After completing all stages shown in Figure 2, 81 articles were selected from the final review.

Figure 3 depicts the distribution of publications by scientific databases in more detail between 2011 and 2020, in which Scopus had the highest number with 34 articles, followed by IEEE Xplore with 25 articles, and ACM with 10 articles. ScienceDirect and Springer were the lowest with 8 and 4 articles, respectively. Among the 81 selected articles, 51 were conference articles, 29 were journals, and 1 was a workshop paper.

Figure 2

The Final Articles Included (n = 81)

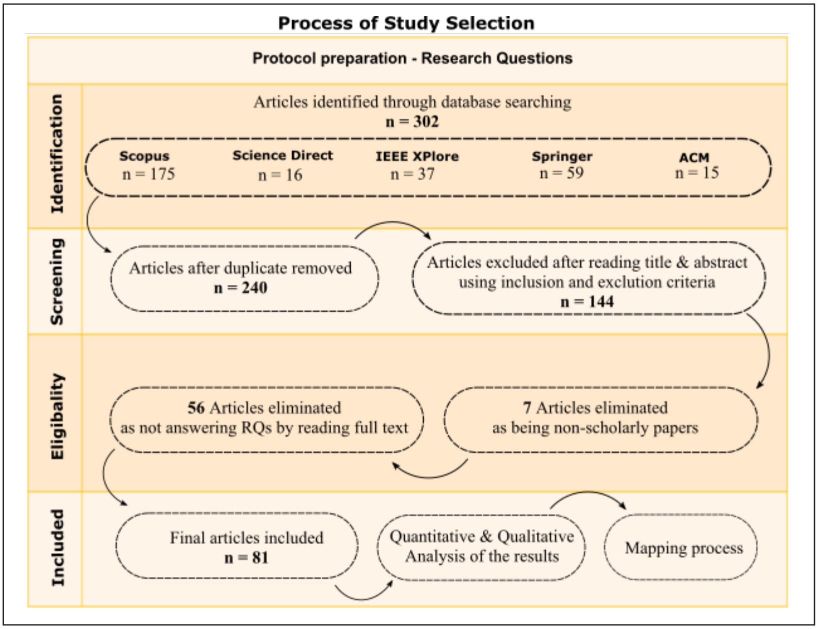
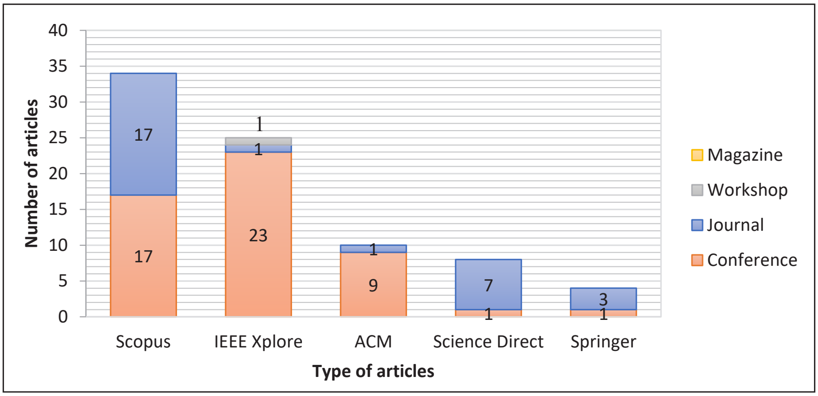

Process of Study Selection
Protocol preparation - Research Questions
Identification
Articles identified through database searching
n = 302
Scopus
n = 175
Science Direct
n = 16
IEEE XPlore
n = 37
Springer
n = 59
ACM
n = 15
Screening
Articles after duplicate removed
n = 240
Articles excluded after reading title & abstract using inclusion and exclution criteria
n = 144
Eligibality
56 Articles eliminated
as not answering RQs by reading full text
7 Articles eliminated
as being non-scholarly papers
Included
Final articles included
n = 81
Quantitative & Qualitative
Analysis of the results
Mapping process

Figure 3

The Publications Distributed by the Type of Articles (n = 81)



Data Extraction

The entire contents of the publications were evaluated in detail at this stage, and the data obtained were categorised according to the search method used to respond to each RQ. The papers were analysed in-depth using elicitation procedures established through a systematic examination of empirical evidence (Davis et al., 2006). All the articles were analysed using the CRQs and BRQs that had been identified as explanations for the findings. At the end of the process, the publishing frequency was determined.

Bibliometric Research Questions (BRQs)

To respond to the RQs, the bibliographic data from the publications were analysed and compiled in this part.
BRQs1 – Articles' publication range and trend: It is crucial to keep track of whether the quantity of papers has increased or decreased over time. From the BRQs, the novelty of this focus research could be seen from the trend presented.

Content Research Questions (CRQs)

The substance of the papers was assessed at this stage, and information for RQs was acquired.
CRQs1 – Machine Learning Techniques: ML is a young branch in the field of Artificial Intelligence (AI), which belongs to one of the core research topics of AI and neural computing (Xue, 2020). ML approaches have only recently become a widely used method for data mining, creating multiple conclusions for prediction purposes (Alyahyan & Düştegör, 2020). The papers were derived based on the research method discussed in the articles or by determining the research design through the evaluation of knowledge used in the articles for non-stated approaches.

CRQs2 – Accuracy or performance measure of the predictions: Performance tests were conducted to determine the prediction's accuracy.
CRQs3 – Characteristics of open datasets: It is crucial to consider how the sorts of datasets listed in the articles, which have been utilised in ML, affect prediction results.
CRQs4 – Potential challenges and research gaps: The highlighted

potential challenges and research gaps in existing studies of OD used ML prediction approaches.

RESULTS AND DISCUSSIONS

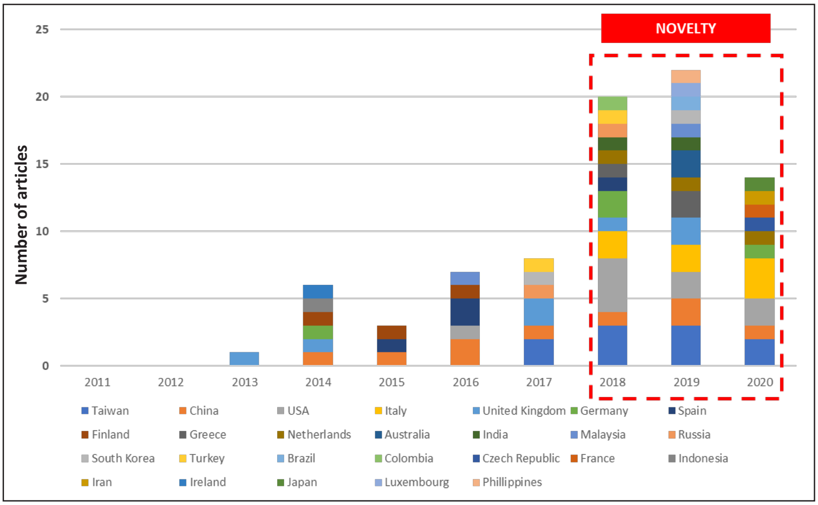
To better understand the trends of the investigations, the data from the SLR were mapped and the results were scrutinised and compared to each of the RQs.

Bibliometric Research Questions (BRQs)

BRQs1 – Articles' publication range and trend:
In this section, for each year, a quantitative study of OD and forecast papers was conducted to evaluate whether there was an increasing or declining trend. The distribution articles' publication range of all 81 mentioned publications spanning the years of 2011 to 2020 and involving 26 countries is as shown in Figure 4.

Figure 4

The Publications Distributed by Years (n = 81)



In 2013, only one relevant article was identified, and no publications were recorded in 2011 and 2012. Nevertheless, this field of research has been shown to have experienced exponential growth since 2014.

This suggests that many studies or analyses regarding this research were conducted within this period. This drastic change was the result of the governments and the practitioners themselves being empowered and inspired to make their datasets accessible and public. In 2013, the European Commission, for example, released a new version of the Public Sector Information Directive, which supplied the entire cultural heritage data in the form of public data that European public agencies may access (Schultz & Shatter, 2013). Besides, federal agencies under the supervision of the White House are producing more open data and machine-readable government data, such as open APIs, which will be used by the government and private developers (Gray, 2014; Weerakkody et al., 2020). In addition, in response to user demand for convenient, structured, and access to the OGD platform that is simple to use, Germany's policymakers, public authorities, commercial sector, the Dresden Agreement was adopted by researchers (Hunnius et al., 2014).

Nevertheless, as seen from the trend, publishing has grown dramatically since 2019. It signifies that more investigations or assessments of the original research were undertaken within this period, which shows the research novelty of this topic on ML prediction using OD. Besides, the number is expected to increase in the coming years. The advancement of ML can disseminate data model architecture and link data silos with data from other organisations to increase data quality and efficiency. However, there was a drastic drop in the number of articles in 2020 that might be affected by the coronavirus disease 2019 (Covid-19) pandemic, through the imposition of 24-hour curfews and closing of schools and universities. Research shows that this pandemic led to the decreasing numbers of non-Covid-19 articles, including the research area of OD and ML prediction (Raynaud et al., 2021) in 2020.

Overall, it can be seen that 26 countries worldwide actively participated and published in this research area with Taiwan, China, the United States of America (USA), Italy, and the United Kingdom being the first five active countries. Over the years, the rise of publishing-producing countries could be aided by government memorandums (Gray, 2014; Wright, 2014) and also by the guidelines of OD principles (Nugroho et al., 2015), which are believed to encourage public knowledge openness and interoperability without barriers to its reuse and consumption. Taiwan debuted its first OD portal in 2013 as compared to the other governments (Chen & Hsu, 2019).

However, the trend showed that OD is still limited and underutilised in several countries including Malaysia. According to Husin et al. (2019), even though OD is a necessity in developing countries, the usage of OD was found to be low, which gained the researchers' interest to identify factors that influenced OD adoption among Malaysian users. Considering that OD consists of free access to the public, this could benefit government agencies to improve their OD in certain areas that can be used by the users. From the trend, Malaysia is still lacking OD initiatives as compared to other countries due to the low support from data providers (Stagars, 2016). Some countries continue to restrict data openness because they believe certain data are too sensitive to be shared with users. As a result, it can be seen that not so many papers are utilising OD in ML prediction models, especially in the Malaysian scenario.

Content Research Questions (CRQs)

CRQs1 - Machine Learning Techniques:
The origins of publication were drawn and identified from the selected articles using the ML approaches and algorithms discussed above, with the results displayed in Figure 5. In brief, from the selected articles, the main ML techniques can be categorised into Supervised, Unsupervised, and Semi-Supervised Learning with appropriate algorithms (Mahesh, 2020; Castanon, 2019; Krishna Sharma & Wang, 2018; Kononenko & Kukar, 2007; Zawacki-Richter et al., 2019). However, from the selected articles, there are several other techniques that have not been mentioned such as Reinforcement Learning and Instance-Based Learning.

Over the previous ten years, the number of publications had increased dramatically. It presented that with 50 publications (62.0%), the approaches for Supervised Learning had received numerous attention and aided the development of prediction utilising OD. Semi-Supervised Learning, a combination of Supervised and Unsupervised Learning methods, was at the second highest with 31.0%, having 25 articles as compared to the other techniques. The trend demonstrated that Supervised Learning and Semi-Supervised Learning had grown considerably since 2013 and are anticipated to grow much more in the upcoming years, as seen in Figure 6.

Figure 5

ML Techniques Classified from the Articles Included (%)

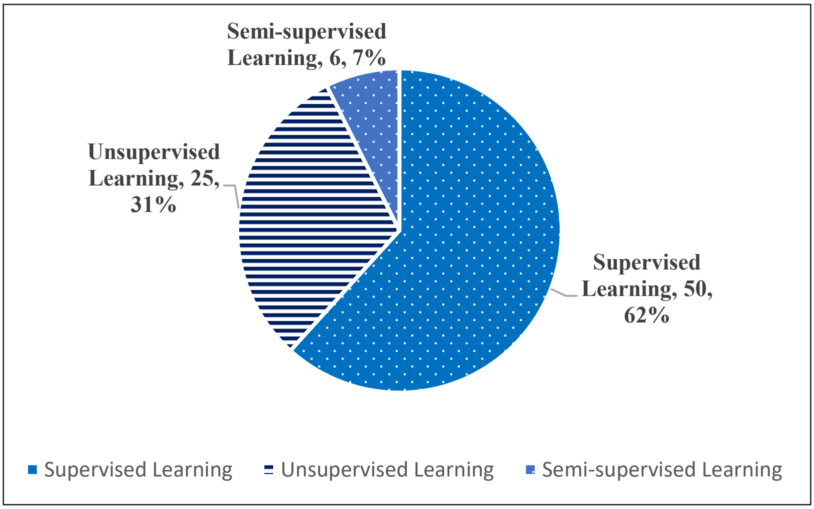
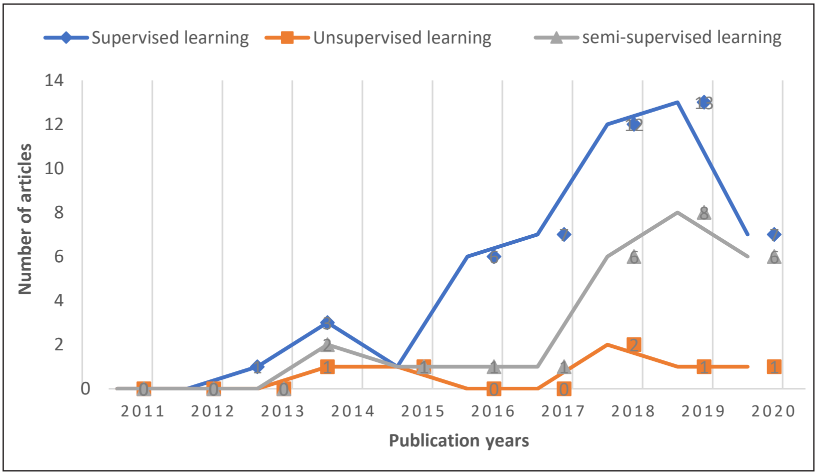


Figure 6

The Trend of ML Techniques Distribution by Year



The classification strategy in Supervised Machine Learning was found to be more popular in the review, with 36 papers, as compared to the regression approach, which had just 7 articles. A variety of algorithms can be used to accomplish the classification strategy, which predicts a discrete value of output, including Naïve Bayes (NB), Logical Regression (LR), Random Forest (RF), Support Vector Machine (SVM), Bayesian Model, Gradient Boosting Decision Tree (GBDT), Decision Tree (DT), J48, and K-Nearest Neighbour Algorithm (KNN), according to the literature. However, Linear Regression, Decision Tree Regressor, Ridge Regression, and Support Vector Regressor (SVR) can be used in a regression technique that predicts a continuous value output.

Recently, Deep Learning-based approaches, such as Time Series Analysis, Artificial Neural Network (ANN), Multi-Layer Perceptron (MLP), Long Short-Term Memory (LSTM), Deep Neural Network (DNN), and Auto-Regressive Integrated Moving Averages (ARIMA), have been introduced as a substitute to Clustering and Classification approaches, which failed to provide results for some research areas. Several clustering algorithms, such as Hidden Markov Models (HMM), K-Medoids, Fuzzy, and K-means, have been used even though Unsupervised Learning techniques were hardly given in the literature. Unsupervised Learning is beneficial when there are no labels assigned to the data. It seeks for previously undetected trends using the bare minimum of human inspection. In addition, a few papers recorded Semi-Supervised Learning with 6.17 percent (6 articles), in which Supervised Learning (labelled data) and Unsupervised Learning (unlabelled data) were combined.

There have been various studies on the multiple ML approach and ensemble method, both of which were innovative prominent strategies for improving prediction. It can be a mix of several algorithms, such as Supervised Learning, Unsupervised Learning, and Deep Learning Techniques. A more detailed comparison of the techniques with their different algorithms together with the most accurate techniques is displayed in Table 8.

From the Classical Classification approach mentioned above, RF was reported to extensively outperform all the other models and achieved high accuracy scores in several studies (Pradhan et al., 2019; Rocca et al., 2016; Kim & Cho, 2019; Dias et al., 2015). Furthermore, for the Deep Learning-based approach, most of the articles reported that LSTM, a derivative model of Recurrent Neural Network (RNN), is the distinguished technique for their research with the best performance measures (Chen et al., 2016; Lee et al., 2020; Awan et al., 2020). Regression techniques, such as Linear Regression, Support Vector Regressor, KNN Regression, and RF Regression, have also been reported in several publications (Violos et al., 2019; Shidik et al., 2014; Boeke et al., 2019; Cocca et al., 2020). Time Series Analysis, particularly ARIMA, has been compared to the other techniques together with ANN and Exponential Smoothing State Space (ETS) (Kamath & Kamat, 2018). However, the ARIMA technique seemed promising, outperforming the other techniques with the best performance and model fit.

In addition, new current ML tools, such as WEKA (Waikato Environment for Knowledge Analysis), have been reported for testing ML algorithms, including SVM, Decision Tree, ANN, Linear Regression, and J48 (Derguech et al., 2014; Sarker et al., 2013; Li et al., 2015). WEKA offers several data visualisation and predictive modelling tools and algorithms, together with graphical user interfaces for easy access. The other tool that was demonstrated was Mathematica™, which can be used to run experiments with datasets using ML algorithms like RF (Pantano et al., 2017).

Table 8

Most Accurate ML Prediction Techniques/Algorithms Using OD

Ref.	Technique's Comparison	Preeminent Techniques	Results/Description
Mohammad et al. (2019)	LR vs ANN vs RF	LR	LR classifier shows the best results with the highest accuracy of 100%
Pradhan et al. (2019)	NB vs DT vs RF vs KNN vs MLR	RF	Accuracy or Log Loss score: 2.276
Codeluppi et al. (2020)	ANN vs BP	ANN	RMSE of 1.50 °C, MAPE: 4.91%, and R2: 0.965
Caparino et al. (2019)	NB vs One-R vs KNN vs C4.5 vs SVM	SVM	Precision, ROC, and Accuracy: 0.97628706
Prabakar et al. (2018)	MLR vs ANN	ANN	The average R2 is close to 1, and RME values are less than 3.8%
Chen et al. (2018)	LR vs DT vs RF vs SVM	SVM	Accuracy: 99.81%
Chen et al. (2019)	(GBDT, LR, and AE) vs LR vs DT vs (GBDT and LR)	GBDT, LR, & AE	AUC: GBDT_AE_LR - 0.858, LR - 0.790, DT - 0.780, and GBDT_LR - 0.847
Yu et al. (2018)	MLP vs AdaBoost vs DT VS LR vs NB vs RF vs SVM	MLP	93% accuracy and AUC with 0.9290
Pohjankukka et al. (2016)	KNN vs MLP vs Ridge regression	MLP	Accuracy: 80%
Roth et al. (2020)	Lasso Regression vs RF vs GBC vs SVM	RF	MSE: RF - 0.293, Regression - 0.312, GB - 0.343, and SVM - 0.316
Li et al. (2019)	KNN vs LR vs NB vs SGD vs RF vs DNN	RF	Accuracy: (0.82, 0.85, 0.86, and 0.88) and F1-scores: (0.82,0.72, 0.90, and 0.88)
Kim and Cho (2019)	Proposed semi-supervised method vs TSVM	Proposed method	Accuracy: 76.79% and F1-score: 86.47%
Celebi et al. (2018)	LR vs NB vs RF	RF	AUC: 0.932 and F-score: 0.860

(continued)

Ref.	Technique’s Comparison	Preeminent Techniques	Results/Description
Awan et al. (2020)	LSTM RNN	LSTM RNN	MAE: 0.214 and MSE: 0.60
Shidik et al. (2014)	BPNN vs Linear Regression vs SVM	Linear Regression & SVM	Linear Regression: MSE - 0.065 and RMSE - 0.255 SVM: MSE - 0.043 and RMSE - 0.207
Celebi et al. (2017)	LR vs KNN vs RFVC GBC	GBC	AUC: 0.88
Arabameri et al. (2020)	Ensemble approaches RSJ48 vs RJ48 vs MJ48 vs J48	RSJ48	AUC: 0.931, PRC: 0.951, E: 0.89, sensitivity: 0.87, and TSS: 0.78
Gao et al. (2019)	MLR vs SVM vs DT vs RF vs GBRT DNN	RF	Accuracy:(0.82, 0.85, 0.86, and 0.88) and F1-scores: (0.82, 0.72, 0.90, and 0.88)
Kamath and Kamat (2018)	ANN vs ETS vs ARIMA	ARIMA	RMSE and model fit: 127.6744
Boeke et al. (2019)	Regression: Linear Regression vs RF Regressor vs SVR and Classification: RF Classifier vs SVC vs KNN	Regression: SVR & Classification: SVC	Regression: MAE SVR 6.83 percentage points and Classification: SVC recall score - 88%, 75%, and 81.82%
Rao and Clarke (2018)	MLR vs Regression Trees vs DNN	DNN	R2: 0.71
Dias et al. (2015)	RF vs ARIMA	RF	Accuracy: 86%

Abbreviations:
MSE, Mean Squared Error; MAE, Mean Absolute Error; MAPE, Mean Absolute Percentage Error, RTAE, Relative Total Absolute Error; RF, Random Forest, SVM, Support Vector Machine; SVR, Support Vector Regression; ANN, Artificial Neural Network; DNN, Deep Neural Network; NN, Neural Network; RNN, Recurrent Neural Network; MLP, Multi-Layer Perceptron; GBDT, Gradient Boosting Decision Tree; NB, Naïve Bayes; DT, Decision Tree; KNN, K-Nearest Neighbour; DCNN, Deep Convolutional Neural Network; AdaBoost, Adaptive Boosting; GBC, Gradient Boosting Classifier; RLS, Regularised Least Squares Regression; BN, Bayes Network; DL, Deep Learning; BPNN, Back Propagation Neural Network; LSTM, Long Short-Term Memory; LR, Logistic Regression; HMM, Hidden Markov Models; SGD, Stochastic Gradient Descent; GBRT, Gradient Boosted Regression Trees; CF, Conditional Inference Forest; F,Fourier Series; KM, K-Means; KP, KM-Polynomial; SP, Shift & Phase; TS, Time Series; MLR, Multiple Linear Regression; LMR, Linear Multiple Regression; BRANNs, Bayesian Regularised ANNs; KRR, Kernel Ridge Regression; DLR, Deep Learning Regression; SGD, Stochastic Gradient Descent; TSVM, Transductive Support Vector Machine; RSJ48, Random Subspace J48; RJ48, Real AdaBoost J48 vs MJ48, MultiBoosting J48; SVC, Support Vector Classifier

CRQs2 – Performance measure of the predictions:
In evaluating a model's efficiency, as shown in Table 9, researchers tried different kinds of approaches to improve prediction and produce more precise results. Nevertheless, some studies did not include any performance measures, and some did not compare their work with other techniques. ML performance techniques can be evaluated using more than only one performance measurement to generate a more accurate prediction.

Accuracy and F1-score have been demonstrated to be used extensively in the articles reviewed. Furthermore, most studies have employed a combination of Accuracy, AUC, F1-score, Precision, and Recall in evaluating their models. Some authors have recommended the Root Mean Squared Error (RMSE), Mean Squared Error (MSE), R-squared, Mean Average Error (MAE), and Mean Absolute Percentage Error (MAPE) as evaluation measures for predicting the best models. However, the accuracy of the results also depends on the precision of the input data (Belesiotis et al., 2018). Moreover, the model's accuracy would improve if the data had more features (Rocca et al., 2016). Feature selection, which is also known as attribute selection, is an essential process to prediction analysis, especially in real OD that consists of a large number of attributes (Basir et al., 2018). This study showed significant results on the manipulation of a bio-inspired algorithm to reduce feature sets. In conclusion, it is difficult to nominate the best performance measure because every technique has its context and novelty.

CRQs3 - Characteristics of open datasets:
It is equally important to see the state-of-the-art open datasets that have been used in ML prediction during the entire SLR process. Figure 7 depicts the distribution of the types of datasets investigated, with the transportation dataset accounting for 19 percent (15) of the total publications.

The spectrums of the transportation dataset were traffic congestion, traffic accidents, traffic flow forecast, primary delay in urban railways, electric vehicles, car-sharing system, parking slot or street parking, etc. With 13 publications (16%), the second highest was Environmental, Climate, and Meteorology. Specifically, meteorology, air temperature, weather forecasting, climate, soil, rainfall, typhoons, floods, air pollution, forestry, and wildland were all covered by the datasets. Then, with 10 publications (12%), the data collected as part of scientific research, such as biology, disease, chemistry, medicines, drugs, life sciences, and healthcare and biomedicine, came in third place. A total of 9 publications (11%) were on energy relating to any energy or power consumption, solar, and water data. Not so different from the energy dataset, the Commerce, Finance, and Economy category had 8 publications (10%) consisting of purchasing behaviour, customer income level, food export, price trend of stocks, finance, and credit risk prediction data. In comparison to the other types of datasets, the lowest were Crime and Citizen Safety, Social and Community, Entertainment and Tourism, Entity-profiling, Geospatial, Education, and Smart Home.

Figure 7

Proportions of the Dataset Types Employed in the Studies (n = 81)

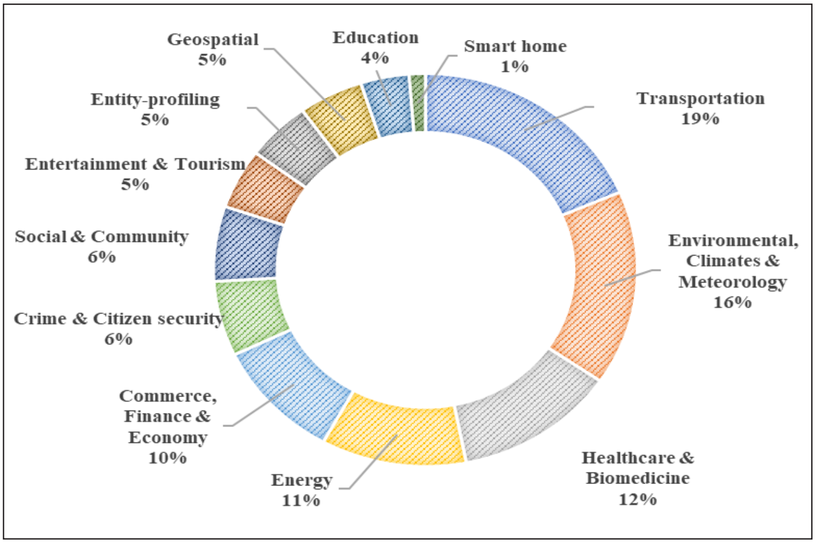


Table 9

Performance Measure Trend for OD in ML Approach Prediction

Ref.	Techniques	ACC	F1	PR	RE	AUC	MSE	RMSE	R	MAE	MAPE
Belesiotis et al. (2018)	Regression, Ridge Regression, RF, SVR	X									
Mohammad et al. (2019)	LR, ANN, RF	X									
Pradhan et al. (2019)	NB, DT, RF, KNN, Multinomial LR	X									
Rocca et al. (2016)	Multiple LR, RF	X									
Pohjankukka et al. (2016)	KNN, MLP, Ridge Regression	X									
Goel et al. (2019)	SVM, NB, KNN, NN	X									
Stolfi et al. (2020)	Polynomial Fitting, F, KM, KP, SP, TS	X									
Lee et al. (2020)	LSTM	X									
Zou and Ergan (2018)	HMM	X									
Bhatia et al. (2018)	NB, LR, RF, SVM, NN (MLP)	X									
Piscopo et al. (2017)	RF	X									
Wu et al. (2017)	DT	X									
Lee and Park (2017)	ANN	X									
Yu et al. (2018)	MLP, AdaBoost, DT, LR, NB, RF, SVM	X				X					
Celebi et al. (2017)	LR, KNN, RF, GBC	X				X					
Kim and Cho (2019)	Transductive SVM	X	X								
Tuke et al. (2020)	Bayesian Model	X	X	X	X						
Ma et al. (2017)	LR, SVM, NB, KNN, BN	X	X	X	X						
Gao et al. (2019)	MLR, SVM, DT, RF, GBRT, DNN	X	X	X	X						
Zainudin and Shamsuddin (2016)	DT	X	X	X	X						

(continued)

Ref.	Techniques	ACC	F1	PR	RE	AUC	MSE	RMSE	R	MAE	MAPE
Cuenca et al. (2018)	GBDT, DL, NB	X	X	X							
Celebi et al. (2018)	LR, NB, RF		X			X					
Menezes et al. (2019)	DNN		X								
Li et al. (2019)	KNN, LR, NB, SGD, RF, DNN		X								
Yang et al. (2018)	GBDT		X								
Gochoo et al. (2018)	DCNN		X								
Xue (2020)	RF, LightGBM Algorithm		X			X					
Chen et al. (2016)	LSTM, MLP, SVM		X	X	X						
Nechaev et al. (2018)	Linear SVM		X	X	X						
Weigert et al. (2020	RF, SVM, NB, KNN, ANN		X	X	X						
Chen et al. (2018)	Logistic Regression, DT, RF, SVM		X	X	X						
Chen et al. (2014)	KNN, RF								X	X	X
Awan et al. (2020)	DNN						X			X	
Shen et al. (2020)	Prophet Forecasting Model (PFM)						X	X		X	
Wu et al. (2019)	LSTM, CNN							X		X	
Utsumi et al. (2020)	Nonlinear Regression							X			X
Codeluppi et al. (2020)	ANN							X			X

Abbreviations: ACC, Accuracy; F1, F1 Score; PR, Precision; RE, Recall; R, R-squared; MSE, Mean Squared Error; MAE, Mean Absolute Error; MAPE, Mean Absolute Percentage Error, RTAE, Relative Total Absolute Error; RF, Random Forest, SVM, Support Vector Machine; SVR, Support Vector Regression; ANN, Artificial Neural Network; DNN, Deep Neural Network; NN, Neural Network; RNN, Recurrent Neural Network; MLP, Multi-Layer Perceptron; GBDT, Gradient Boosting Decision Tree; NB, Naïve Bayes; DT, Decision Tree; KNN, K-Nearest Neighbour; DCNN, Deep Convolutional Neural Network; AdaBoost, Adaptive Boosting; GBC, Gradient Boosting Classifier; RLS, Regularised Least Squares Regression; BN, Bayes Network; DL, Deep Learning; BPNN, Back Propagation Neural Network; LSTM, Long Short-Term Memory; LR, Logistic Regression; HMM, Hidden Markov Models; SGD, Stochastic Gradient Descent; GBRT, Gradient Boosted Regression Trees; CF, Conditional Inference Forest; F,Fourier Series; KM, K-Means; KP, KM-Polynomial; SP, Shift & Phase; TS, Time Series; MLR, Multiple Linear Regression; LMR, Linear Multiple Regression

The type of the datasets and their characteristics were examined from the datasets that had been cited in the articles and several more detailed dataset examples presented in Table 10. However, some articles did not cite their datasets due to information sensitivity and security to the data providers or stakeholders, such as data provided by the police department or data related to personal data protection. The types of datasets and their characteristics with respect to the source of the dataset can provide information to the researchers and practitioners in selecting and evaluating the prediction models that are most suited for their studies. Correct and reliable features of datasets will increase the accuracy of prediction and performance.

Limited experimental data and unknown relevant variables may pose a challenge in some studies. As a result, OD can give useful data to confirm current data, increase the applicability of indicator variables, and improve forecast validity (Noymanee et al., 2017). Another challenge of OD from the government portal is an imbalance dataset, which leads to low prediction performance (Zainudin & Shamsuddin, 2016).

Table 10

Example of OD Datasets and their Characteristics Used in ML Prediction

Ref.	Type	Characteristics	URL link
Belesiotis et al. (2018)	Crime	Data included crime data, points of interest, demographics, transportation and mobility, land use.	https://data.gov.uk/dataset/lower_layer_super_output_area_lsoa_boundaries, https://data.police.uk/about/, http://www.ons.gov.uk/census/2011census, http://www.openstreetmap.org, http://wikimapia.org, https://foursquare.com, https://tfl.gov.uk/info-for/open-data-users/our-open-data
Derguech et al. (2014)	Climate	Weather forecast information.	https://openweathermap.org/
Menezes et al. (2019)	Entity profile	Named Entity Recognition (NER) profiling information on entity's class (person, organisation, location), ID of the page (Wiki ID), title of the page, and names of the entity.	Wikipedia and Dbpedia
Chen et al. (2019)	Finance	Dataset from "Give Me Some Credit" Kaggle competition with 150,000 samples for credit risk prediction.	https://www.kaggle.com/
Xue (2020)	Finance	6,720 financial data samples with deadline of 30th October 2018.	https://xunlei.com/, https://www.ppmoney.com, http://www.webhome.org/
Chen et al. (2016)	Transportation	Traffic congestion between 28th December 2014 and 3rd February 2015. Data were collected every five minutes, covering 1,649 segments in Beijing, China.	http://ditu.amap.com/
Devyatkin et al. (2018)	Economy	Food export gain data from 2008–2016.	http://www.fao.org/faostat/en/, https://comtrade.un.org/data/, http://www.imf.org/en/

(continued)

Ref.	Type	Characteristics	URL link
Badii et al. (2018)	Transportation	Data used for the parking slot prediction during the period from 5th January 2017 to 26th March 2017.	https://www.km4city.org/webapp/
Arango et al. (2016)	Geospatial	Landsat 8 satellite radiometric data valuable for precision agriculture.	http://earth.esa.int/
Tuke et al. (2020)	Tourism	Twitter data was collected using the public API between 21st July 2017 and 14th February 2018.	https://twitter.com/?lang=en
Roth et al. (2020)	Energy consumption	Data for building energy estimation in 2016 included 15,000 buildings in New York City (NYC) energy consumption.	https://www1.nyc.gov/html/gbee/html/plan/ll84_scores.shtml https://www1.nyc.gov/site/planning/data-maps/open-data/dwn-pluto-mappluto.page, https://openei.org/doe-opendata/dataset/commercial-and-residential-hourly-load-profiles-for-all-tmy3-locations-in-the-united-states, https://openweathermap.org/history, http://www.energyonline.com/Data/GenericData.aspx?DataId = 13
Noymanee et al. (2017)	Climate	Flood forecasting from datasets of Pattani Basin telemeter project with hourly data in the period of 2015–2016 (training dataset) and 15th January–15th February 2017 (testing dataset).	http://www.telepattani.com.
Kim and Cho (2019)	Social & community	Data of social lending between loan, borrowers, investors, and credit histories with 69 attributes after normalisation and binary dummies.	https://www.lendingclub.com/info/statistics.action

(continued)

Ref.	Type	Characteristics	URL link
Stolfi et al. (2020)	Transportation	Data included from the cities of Birmingham, Nottingham, Glasgow, and the county of Norfolk, all in the United Kingdom. Under the UK Open Government Licence (OGL) or Creative Commons Attribution, data were published.	http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/, http://opendefinition.org/licenses/cc-by/ https://data.birmingham.gov.uk/dataset/birmingham-parking https://data.glasgow.gov.uk/dataset/car-park-feeds https://data.gov.uk/dataset/norfolk-county-council-live-car-park-data, http://www.opendatanottingham.org.uk/dataset.aspx?id=55 http://mallba4.lcc.uma.es/parking/
Dorado et al. (2018)	Climate	The climate and soil data.	http://conservacion.cimmyt.org/es/hubs/683
Wood (2019)	Energy consumption	A compiled dataset (8,784 data records) of hourly-averaged solar power generation (MW) for Germany in 2016 integrated eight influencing weather, environmental, and market price variables.	https://data.open-power-system-data.org/
Zou and Ergan (2018)	Social & Community	Data included from 311 Service Requests from 2010 to 2017.	https://data.cityofnewyork.us/Social-Services/311-Service-Requests-from-2010-to-Present/erm2-nwe9
Awan et al. (2020)	Transportation	Road traffic open data from Madrid City, Spain collected and normalised for one year of observation.	https://datos.madrid.es/portal/site/egob
Shidik et al. (2014)	Environmental	Open government time series data of forest fires from 1,960 until 2008 were included from OGD United States of America.	http://data.gov, http://data-gov.tw.rpi.edu/wiki/

(continued)

Ref.	Type	Characteristics	URL link
Bhatia et al. (2018)	Entity-profile	Sentiment or emotion information by Stanford University between June 2009 to December 2009, 467 million Twitter posts from 20 million users.	https://snap.stanford.edu/data/twitter7.html.
Pantano et al. (2017)	Tourism	Five-hundred online reviewers were chosen from the TripAdvisor database - a profile including interest in the 18 topics (foodie, shopping fanatic, urban explorer, nightlife seeker etc.)	https://www.tripadvisor.com/

CRQs4 – Potential challenges and research gap highlighted in existing studies

In the present study, there are some challenges and research gaps that have been highlighted in the publications to improve the proposed model by increasing the accuracy performance. Further research may be required to fill in the research gap.

Most of the articles suggested exploring and evaluating other state-of-the-art ML models that could potentially improve the prediction models. The performance of DNN was compared to other approaches, such as the Convolutional Neural Network (CNN) feature extraction and the Recurrent Convolutional Neural Network (RCNN), which have been used to improve prediction accuracy (Chen et al., 2019). A combination of DL and ML prediction has also been carried out on the Persian sentiment analysis by using Tweet OD for the first time; however, future research should include new features to the classification to boost the performance (Nezhad & Deihimi, 2020). Research shows that manipulating of the bio-inspired search algorithms can be considered in OD for future studies, as it can demonstrate the best setup for more promising results (Basir et al., 2018). Nevertheless, some researchers also showed that conventional ML models could work better than DL techniques over time. Therefore, it is a research opportunity to explore other options and select the better ML or DL model in prediction (Awan et al., 2020).

Another key point raised by the researchers is the need to include more OD variances in the prediction model, which could help the performance index (Chen et al., 2018). Some studies also found that incorporating data from multiple dataset sources improved prediction model accuracy significantly (Belesiotis et al., 2018). Existing researchers mentioned that the robust prediction model development was heavily influenced by the accuracy and reliability of datasets to be trained in the model. Moreover, it was noted that certain predictions were not accurate due to the instability of the database itself in which the dataset would require more features and sufficient data (Jai et al., 2014). In addition, another research gap is the study on how the various databases impacted the prediction results (Devyatkin et al., 2018). The literature also proposed that training and test datasets can be prepared without shuffling, such as using the 2011–2015 training data and the 2016 data as the test dataset (Prabakar et al., 2018). One more limitation reported is the training data, which did not have a “contextual consistency” feature and thus affected the accuracy of the results.

UGC and SNSs, such as Twitter, Facebook, Instagram, and many more, have been widely identified as OD platforms, and the current study remarked that Twitter might not be the best medium for detecting event features of the dataset. Likely, combining multiple datasets (e.g., Facebook posts or web searches) will improve predictions and framework flexibility. Tweets referenced as features in predictive models can be further investigated using network characteristics. The use of these networks’ structural characteristics will minimise “noise” in the data used for prediction and provide better-quality evidence for future events (Awan et al., 2020). In addition, videos are shared on social media as it is one of the OD platforms that provide comments/ reviews features to predict effective video concepts by using ML prediction such as the Self-Organising Map (SOM) technique (Thabet et al., 2021).

Furthermore, the models predict whether the same strategy can be generalised and applied for different contexts or tasks of research that should be done, which is termed as Transfer Learning (Cocca et al., 2020). Extension research could be performed, for example, by transferring the proposed prototype of predictors by creating a mobile app or by implementing parallel graphics processing unit-based (GPU) computation on the prediction for millions of users and billions of items (Stolfi et al., 2020; Pradhan et al., 2019). New services can also be developed by implementing big data processing techniques with more data streams combined (Lee & Park, 2017).

Limitations and Threats to Validity

Limitations: The findings of this study are based on the following limitations: (a) publications that were available after December 2020 were not accounted for; (b) results may be subjected to the limitations of each digital library’s automated search engines (IEEE, ScienceDirect, ACM, Springer, and Scopus); (c) only studies published in English have been chosen; and (d) a single researcher carried out the whole review.

Threats to validity: In this study, papers that did not have OD, prediction, and ML in their titles, keywords, and abstracts were excluded. Because of its sensitivity, certain datasets were found to be missing in the literature and were referred to as grey literature, such as scientific reports. This may also lead to a negative connotation that SLR could not discuss such important datasets and their scientific contribution.

CONCLUSIONS AND AREAS FOR FURTHER RESEARCH

The goal of this research was to present the findings of the SLR that highlight the enormous potential of OD sources in ML-based prediction to influence users' attitudes and behaviours. In practice, ML prediction tools can help anticipate outcomes for various fields in decision-making. This study could help organisations, practitioners, and researchers by giving information on current trends in the OD setting and mapping studies based on the RQs designed, the most recent developments, and the necessity for additional research based on the information supplied. In this systematic review, 81 selected articles published from January 2019 to December 2020 were examined. The trends showed that ML prediction techniques using OD increased since 2014 and are expected to be more in the coming years. Since the opening of datasets by the governments, 26 countries worldwide actively participated and published articles in this research area. However, in the analysis, some countries, including Malaysia, are not ranked in any one of the top countries contributing to the ML prediction using OD. This is probably due to a lack of skills and competencies among government agencies in Malaysia on leveraging AI and ML. This can be investigated in more detail and is worthy of being identified as one of the implementation gaps.

Various state-of-the-art ML techniques applied in several sub-fields in the prediction model have been mentioned in these existing studies, but all of them are still in their infancy. Each ML technique has its state-of-the-art or novelty. Most of the studies reviewed compared the proposed prediction model with other techniques to achieve the most accurate and robust model. There are more than ten performance measures that can be used for ML techniques in this scope of research.

However, choosing the optimal performance measurement is difficult because each technique has its own context and novelty. As the accuracy of the prediction model depends on the accuracy of the input datasets, this study researched the characteristics of the open datasets that have been used in the literature and their impact on the accuracy of the results. The type of datasets and their characteristics in comparison with the source of the datasets can provide information to the researchers and practitioners in selecting and evaluating which prediction models should be proposed in their studies. In addition, the articles have identified various challenges and research gaps that must be addressed to improve the proposed prediction model in terms of increasing the accuracy of the results.

This study has certain flaws as not much research has been done on this area, and thus more research is needed to fill in the research gap. Consideration of a longer publication period of reviewed journals may show a more prominent trend in ML prediction using OD.

ACKNOWLEDGMENT

The authors gratefully acknowledge Universiti Sains Malaysia (USM) and Universiti Malaysia Perlis (UniMAP) for the support they have extended in the completion of this research.

REFERENCES

Al-Garadi, M. A., Hussain, M. R., Khan, N., Murtaza, G., Nweke, H. F., Ali, I., ... & Gani, A. (2019). Predicting cyberbullying on social media in the big data era using machine learning algorithms: Review of literature and open challenges. *IEEE Access*, *7*, 70701–70718. https://doi.org/10.1109/access.2019.2918354

Alyahyan, E., & Düştegör, D. (2020). Predicting academic success in higher education: Literature review and best practices. *International Journal of Educational Technology in Higher Education*, *17*(1), 1–21. https://doi.org/10.1186/s41239-020-0177-7

Arabameri, A., Saha, S., Mukherjee, K., Blaschke, T., Chen, W., Ngo, P. T. T., & Band, S. S. (2020). Modeling spatial flood using novel ensemble artificial intelligence approaches in northern Iran. *Remote Sensing*, *12*(20), 3423. https://doi.org/10.3390/rs12203423

Arango, R. B., Campos, A. M., Combarro, E. F., Canas, E. R., & Díaz, I. (2016). Mapping cultivable land from satellite imagery with clustering algorithms. *International Journal of Applied Earth Observation and Geoinformation*, *49*, 99–106. https://doi.org/10.1016/j.jag.2016.01.009

Awan, F. M., Minerva, R., & Crespi, N. (2020). Improving road traffic forecasting using air pollution and atmospheric data: Experiments based on LSTM recurrent neural networks. *Sensors*, *20*(13), 3749. https://doi.org/10.3390/s20133749

Badii, C., Nesi, P., & Paoli, I. (2018). Predicting available parking slots on critical and regular services by exploiting a range of open data. *IEEE Access*, *6*, 44059–44071. https://doi.org/10.1109/access.2018.2864157

Basir, M. A., Yusof, Y., & Hussin, M. S. (2018). Optimization of attribute selection model using bio-inspired algorithms. *Journal of Information and Communication Technology*, *18*(1), 35–55. https://doi.org/10.32890/jict2019.18.1.3

Belesiotis, A., Papadakis, G., & Skoutas, D. (2018). Analyzing and predicting spatial crime distribution using crowdsourced and open data. *ACM Transactions on Spatial Algorithms and Systems (TSAS)*, *3*(4), 1–31. https://doi.org/10.1145/3190345

Bhatia, A., Hagras, H., & Lepley, J. J. (2018, September). Machine learning approach to extracting emotions information from open source data for relative forecasting of stock prices. In *2018 10th Computer Science and Electronic Engineering (CEEC)* (pp. 142–147). IEEE. https://doi.org/10.1109/ceec.2018.8674180

Bizer, C., Heath, T., & Berners-Lee, T. (2011). Linked data: The story so far. In *Semantic services, interoperability and web applications: Emerging concepts* (pp. 205–227). IGI global. https://doi.org/10.4018/978-1-60960-593-3.ch008

Boeke, S., van den Homberg, M. J. C., Teklesadik, A., Fabila, J. L. D., Riquet, D., & Alimardani, M. (2019). Towards predicting rice loss due to typhoons in the Philippines. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, *42*, 63–70. https://doi.org/10.5194/isprs-archives-xlii-4-w19-63-2019

Boyle, F., & Sherman, D. (2006). Scopus™: The product and its development. *The Serials Librarian*, *49*(3), 147–153. https://doi.org/10.1300/j123v49n03_12

Budgen, D., Burn, A. J., Brereton, O. P., Kitchenham, B. A., & Pretorius, R. (2011). Empirical evidence about the UML: A systematic literature review. *Software: Practice and Experience*, *41*(4), 363–392. https://doi.org/10.1002/spe.1009

Butt, U. M., Letchmunan, S., Hassan, F. H., Ali, M., Baqir, A., & Sherazi, H. H. R. (2020). Spatio-temporal crime HotSpot detection and prediction: A systematic literature review. *IEEE Access*, *8*, 166553–166574. https://doi.org/10.1109/access.2020.3022808

Capariño, E. T., Sison, A. M., & Medina, R. P. (2019, February). Application of the modified imputation method to missing data to increase classification performance. In *2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS)* (pp. 134–139). IEEE. https://doi.org/10.1109/ccoms.2019.8821632

Castañón, J. (2019). Machine learning methods that every data scientist should know. *Consultado em Outubro*, *16*.

Celebi, R., Erten, Ö., & Dumontier, M. (2017, December). Machine learning based drug indication prediction using linked open data. In *Proceedings of the 10th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences (SWAT4LS)*.

Celebi, R., Yasar, E., Uyar, H., Gumus, O., Dikenelli, O., & Dumontier, M. (2018). Evaluation of knowledge graph embedding approaches for drug-drug interaction prediction using linked open data. https://doi.org/10.1186/s12859-019-3284-5

Chen, C. L., Huang, F. M., Liu, Y. H., & Wu, D. E. (2018, October). Artificial intelligence and mobile phone sensing based user activity recognition. In *2018 IEEE 15th International Conference on e-Business Engineering (ICEBE)* (pp. 164–171). IEEE. https://doi.org/10.1109/icebe.2018.00034

Chen, H., Hu, Q., & He, L. (2014, November). Clairvoyant: An early prediction system for video hits. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management* (pp. 2054–2056). https://doi.org/10.1145/2661829.2661847

Chen, I. C., & Hsu, I. C. (2019). Open Taiwan government data recommendation platform using DBpedia and semantic web based on cloud computing. *International Journal of Web Information Systems*, *15*(2), 236–254. https://doi.org/10.1108/ijwis-02-2018-0015

Chen, S., Wang, Q., & Liu, S. (2019, June). Credit risk prediction in peer-to-peer lending with ensemble learning framework. In *2019 Chinese Control And Decision Conference (CCDC)* (pp. 4373–4377). IEEE. https://doi.org/10.1109/ccdc.2019.8832412

Chen, Y. Y., Lv, Y., Li, Z., & Wang, F. Y. (2016, November). Long short-term memory model for traffic congestion prediction with online open data. In *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)* (pp. 132–137). IEEE. https://doi.org/10.1109/itsc.2016.7795543

Chu, S. C., & Kim, Y. (2011). Determinants of consumer engagement in electronic word-of-mouth (eWOM) in social networking sites. *International Journal of Advertising*, *30*(1), 47–75. https://doi.org/10.2501/ija-30-1-047-075

Cocca, M., Teixeira, D., Vassio, L., Mellia, M., Almeida, J. M., & Couto da Silva, A. P. (2020). On car-sharing usage prediction with open socio-demographic data. *Electronics*, *9*(1), 72. https://doi.org/10.3390/electronics9010072

Codeluppi, G., Cilfone, A., Davoli, L., & Ferrari, G. (2020, November). AI at the edge: A smart gateway for greenhouse air temperature forecasting. In *2020 IEEE International Workshop on Metrology for Agriculture and Forestry (MetroAgriFor)* (pp. 348–353). IEEE. https://doi.org/10.1109/metroagrifor50201.2020.9277553

Cuenca, L. G., Puertas, E., Aliane, N., & Andres, J. F. (2018, September). Traffic accidents classification and injury severity prediction. In *2018 3rd IEEE International Conference on Intelligent Transportation Engineering (ICITE)* (pp. 52–57). IEEE. https://doi.org/10.1109/icite.2018.8492545

Davis, A., Dieste, O., Hickey, A., Juristo, N., & Moreno, A. M. (2006, September). Effectiveness of requirements elicitation techniques: Empirical results derived from a systematic review. In *14th IEEE International Requirements Engineering Conference (RE'06)* (pp. 179–188). IEEE. https://doi.org/10.1109/re.2006.17

Derguech, W., Bruke, E., & Curry, E. (2014, December). An autonomic approach to real-time predictive analytics using open data and internet of things. In *2014 IEEE 11th Intl Conf on Ubiquitous Intelligence and Computing and 2014 IEEE 11th Intl Conf on Autonomic and Trusted Computing and 2014 IEEE 14th Intl Conf on Scalable Computing and Communications and Its Associated Workshops* (pp. 204–211). IEEE. https://doi.org/10.1109/uic-atc-scalcom.2014.137

Devyatkin, D., Suvorov, R., Tikhomirov, I., & Otmakhova, Y. (2018, September). Neural networks for food export gain forecasting. In *2018 International Conference on Intelligent Systems (IS)* (pp. 312–317). IEEE. https://doi.org/10.1109/is.2018.8710561

Dias, G. M., Bellalta, B., & Oechsner, S. (2015, November). Predicting occupancy trends in Barcelona's bicycle service stations using open data. In *2015 sai intelligent systems conference (intellisys)* (pp. 439–445). IEEE. https://doi.org/10.1109/intellisys.2015.7361177

Diffley, S., Kearns, J., Bennett, W., & Kawalek, P. (2011). Consumer behaviour in social networking sites: Implications for marketers. *Irish Journal of Management*.

Dorado, H., Delerce, S., Jimenez, D., & Cobos, C. (2018, October). Finding optimal farming practices to increase crop yield through global-best harmony search and predictive models, a data-driven approach. In *Mexican International Conference on Artificial Intelligence* (pp. 15–29). Springer, Cham. https://doi.org/10.1007/978-3-030-04497-8_2

Foulonneau, M., Martin, S., & Turki, S. (2014, February). How open data are turned into services?. In *International Conference on Exploring Services Science* (pp. 31–39). Springer, Cham. https://doi.org/10.1007/978-3-319-04810-9_3

Gao, S., Li, M., Liang, Y., Marks, J., Kang, Y., & Li, M. (2019). Predicting the spatiotemporal legality of on-street parking using open data and machine learning. *Annals of GIS*, *25*(4), 299–312. https://doi.org/10.1080/19475683.2019.1679882

Gochoo, M., Tan, T. H., Liu, S. H., Jean, F. R., Alnajjar, F. S., & Huang, S. C. (2018). Unobtrusive activity recognition of elderly people living alone using anonymous binary sensors and DCNN. *IEEE Journal of Biomedical and Health Informatics*, *23*(2), 693–702. https://doi.org/10.1109/jbhi.2018.2833618

Goel, M., Sharma, N., & Gurve, M. K. (2019, September). Analysis of global terrorism dataset using open source data mining tools. In *2019 International Conference on Computing, Power and Communication Technologies (GUCON)* (pp. 165–170). IEEE.

Goldstein, E. B., Coco, G., & Plant, N. G. (2019). A review of machine learning applications to coastal sediment transport and morphodynamics. *Earth-Science Reviews*, *194*, 97–108. https://doi.org/10.1016/j.earscirev.2019.04.022

Gray, J. (2014, September). Towards a genealogy of open data. In *The paper was given at the General Conference of the European Consortium for Political Research in Glasgow*. https://doi.org/10.2139/ssrn.2605828

Gutierrez-Osorio, C., & Pedraza, C. (2020). Modern data sources and techniques for analysis and forecast of road accidents: A review. *Journal of Traffic and Transportation Engineering*

(English edition), *7*(4), 432–446. https://doi.org/10.1016/j.jtte.2020.05.002

Halevi, G., Moed, H., & Bar-Ilan, J. (2017). Suitability of Google Scholar as a source of scientific information and as a source of data for scientific evaluation—Review of the literature. *Journal of Informetrics*, *11*(3), 823–834. https://doi.org/10.1016/j.joi.2017.06.005

Hunnius, S., Krieger, B., & Schuppan, T. (2014, September). Providing, guarding, shielding: Open government data in Spain and Germany. In *European Group for Public Administration Annual Conference, Speyer, Germany*.

Husin, N. N. F. A., Zakaria, N. H., & Dahlan, H. M. (2019, December). Factors influencing open data adoption in Malaysia based on users perspective. In *2019 6th International Conference on Research and Innovation in Information Systems (ICRIIS)* (pp. 1–5). IEEE. https://doi.org/10.1109/icriis48246.2019.9073396

Iskandaryan, D., Ramos, F., & Trilles, S. (2020). Air quality prediction in smart cities using machine learning technologies based on sensor data: A review. *Applied Sciences*, *10*(7), 2401. https://doi.org/10.3390/app10072401

Jai, T. M. C., & Burns, L. D. (2014). Attributes of apparel tablet catalogs: Value proposition comparisons. *Journal of Fashion Marketing and Management*. https://doi.org/10.1108/jfmm-12-2012-0073

Kamath, R. S., & Kamat, R. K. (2018). Time-series analysis and forecasting of rainfall at Idukki district, Kerala: Machine learning approach. *Disaster Advances*, *11*(11), 27–33.

Kim, A., & Cho, S. B. (2019). An ensemble semi-supervised learning method for predicting defaults in social lending. *Engineering Applications of Artificial Intelligence*, *81*, 193–199. https://doi.org/10.1016/j.engappai.2019.02.014

Kitchenham, B., & Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering. https://doi.org/10.1016/j.infsof.2008.09.009

Kononenko, I., & Kukar, M. (2007). *Machine learning and data mining*. Horwood Publishing. https://doi.org/10.1533/9780857099440

Krishna Sharma, S., & Wang, X. (2018). Towards massive machine type communications in ultra-dense cellular IoT networks: Current issues and machine learning-assisted solutions. *arXiv e-prints*, arXiv-1808. https://doi.org/10.1109/comst.2019.2916177

Lee, J., & Park, G. L. (2017, September). Temporal data stream analysis for EV charging infrastructure in Jeju.

In *Proceedings of the International Conference on Research in Adaptive and Convergent Systems* (pp. 36–39). https://doi.org/10.1145/3129676.3129717

Lee, M. C., Liao, J. S., Yeh, S. C., & Chang, J. W. (2020, January). Forecasting the short-term price trend of Taiwan stocks with deep neural network. In *Proceedings of the 2020 11th International Conference on E-Education, E-Business, E-Management, and E-Learning* (pp. 296–299). https://doi.org/10.1145/3377571.3377608

Li, M., Gao, S., Liang, Y., Marks, J., Kang, Y., & Li, M. (2019, November). A data-driven approach to understanding and predicting the spatiotemporal availability of street parking. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (pp. 536–539). https://doi.org/10.1145/3347146.3359366

Li, R., Xiong, H., & Zhao, H. (2015, November). More than address: Pre-identify your income with the open data. In *2015 International Conference on Cloud Computing and Big Data (CCBD)* (pp. 193–200). IEEE. https://doi.org/10.1109/ccbd.2015.51

Lindman, J., Kinnari, T., & Rossi, M. (2014, January). Industrial open data: Case studies of early open data entrepreneurs. In *2014 47th Hawaii international conference on system sciences* (pp. 739–748). IEEE. https://doi.org/10.1109/hicss.2014.99

Ma, C., Yao, B., Ge, F., Pan, Y., & Guo, Y. (2017, September). Improving prediction of student performance based on multiple feature selection approaches. In *Proceedings of the 2017 International Conference on E-Education, E-Business and E-Technology* (pp. 36–41). https://doi.org/10.1145/3141151.3141160

Madarash-Hill, C., & Hill, J. B. (2004). Enhancing access to IEEE conference proceedings: A case study in the application of IEEE Xplore full text and table of contents enhancements. *Science & Technology Libraries*, *24*(3-4), 389–399. https://doi.org/10.1300/j122v24n03_09

Maglyas, A., Nikula, U., & Smolander, K. (2011, August). What do we know about software product management? - A systematic mapping study. In *2011 Fifth International Workshop on Software Product Management (IWSPM)* (pp. 26–35). IEEE. https://doi.org/10.1109/iwspm.2011.6046201

Mahesh, B. (2020). Machine learning algorithms - A review. *International Journal of Science and Research (IJSR). [Internet]*, *9*, 381–386.

MAMPU (2017), “Our open data policy.” https://www.data.gov.my, Last accessed 2019-09-13.

Menezes, D., Milidiu, R., & Savarese, P. (2019, October). Building a massive corpus for named entity recognition using free open data sources. In *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)* (pp. 6–11). IEEE. https://doi.org/10.1109/bracis.2019.00011

Mohammad, N. I., Ismail, S. A., Kama, M. N., Yusop, O. M., & Azmi, A. (2019, August). Customer churn prediction in telecommunication industry using machine learning classifiers. In *Proceedings of the 3rd International Conference on Vision, Image and Signal Processing* (pp. 1–7). https://doi.org/10.1145/3387168.3387219

Nechaev, Y., Corcoglioniti, F., & Giuliano, C. (2018, October). Type prediction combining linked open data and social media. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management* (pp. 1033–1042). https://doi.org/10.1145/3269206.3271781

Nezhad, Z. B., & Deihimi, M. A. (2020). Sarcasm detection in Persian. *Journal of Information and Communication Technology*, *20*(1), 1–20. https://doi.org/10.32890/jict.20.1.2021.6249

Noymanee, J., Nikitin, N. O., & Kalyuzhnaya, A. V. (2017). Urban pluvial flood forecasting using open data with machine learning techniques in Pattani basin. *Procedia Computer Science*, *119*, 288–297. https://doi.org/10.1016/j.procs.2017.11.187

Nugroho, R. P., Zuiderwijk, A., Janssen, M., & de Jong, M. (2015). A comparison of national open data policies: Lessons learned. *Transforming Government: People, Process and Policy*. https://doi.org/10.1108/tg-03-2014-0008

Open Data Handbook (OD). (2012). *What is open data?* https://opendatahandbook.org/guide/en/what-is-open-data/. Last accessed 2019-04-01.

Open Knowledge Foundation (OKFN). (2014). “*What is open data?*. http://okfn.org/opendata/. Last accessed 2019-04-01.

Pantano, E., Priporas, C. V., & Stylos, N. (2017). ‘You will like it!’ Using open data to predict tourists’ response to a tourist attraction. *Tourism Management*, *60*, 430–438. https://doi.org/10.1016/j.tourman.2016.12.020

Piscopo, A., Siebes, R., & Hardman, L. (2017). Predicting sense of community and participation by applying machine learning to open government data. *Policy & Internet*, *9*(1), 55–75. https://doi.org/10.1002/poi3.145

Pohjankukka, J., Riihimäki, H., Nevalainen, P., Pahikkala, T., Ala-Ilomäki, J., Hyvönen, E., ... & Heikkonen, J. (2016). Predictability of boreal forest soil bearing capacity by machine learning. *Journal of Terramechanics*, *68*, 1–8. https://doi.org/10.1016/j.jterra.2016.09.001

Prabakar, A., Wu, L., Zwanepol, L., Van Velzen, N., & Djairam, D. (2018, October). Applying machine learning to study the relationship between electricity consumption and weather variables using open data. In *2018 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe)* (pp. 1–6). IEEE. https://doi.org/10.1109/isgteurope.2018.8571430

Pradhan, I., Potika, K., Eirinaki, M., & Potikas, P. (2019, June). Exploratory data analysis and crime prediction for smart cities. In *Proceedings of the 23rd International Database Applications & Engineering Symposium* (pp. 1–9). https://doi.org/10.1145/3331076.3331114

Publications, Insights (2014). *What executives should know about open data*. http://www.mckinsey.com/insights/high_tech_telecoms_internet/what_executives_should_know_about_open_data. Last accessed 2019-04-01.

Rao, A. R., & Clarke, D. (2018, July). A comparison of models to predict medical procedure costs from open public healthcare data. In *2018 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–8). IEEE. https://doi.org/10.1109/ijcnn.2018.8489257

Raynaud, M., Goutaudier, V., Louis, K., Al-Awadhi, S., Dubourg, Q., Truchot, A., ... & Loupy, A. (2021). Impact of the COVID-19 pandemic on publication dynamics and non-COVID-19 research production. *BMC Medical Research Methodology*, *21*(1), 1–10. https://doi.org/10.1186/s12874-021-01404-9

Rocca, G. B., Castillo-Cara, M., Levano, R. A., Herrera, J. V., & Orozco-Barbosa, L. (2016, November). Citizen security using machine learning algorithms through open data. In *2016 8th IEEE Latin-American Conference on Communications (LATINCOM)* (pp. 1–6). IEEE. https://doi.org/10.1109/latincom.2016.7811562

Roth, J., Martin, A., Miller, C., & Jain, R. K. (2020). SynCity: Using open data to create a synthetic city of hourly building energy estimates by integrating data-driven and physicsbased methods. *Applied Energy*, *280*, 115981. https://doi.org/10.1016/j.apenergy.2020.115981

Saxena, S. (2016). Open government data (OGD) usage in India: A conceptual framework using TOE & UTAUT frameworks. *SOCRATES*, *4*(3), 124–144. https://doi.org/10.1108/fs-02-2017-0003

Sadoughi, F., Behmanesh, A., & Sayfouri, N. (2020). Internet of things in medicine: A systematic mapping study. *Journal of Biomedical Informatics*, *103*, 103383. https://doi.org/10.1016/j.jbi.2020.103383

Sarker, F., Tiropanis, T., & Davis, H. C. (2013). Students' performance prediction by using institutional internal and external open data sources. https://doi.org/10.5220/0004383006390646

Schultz, M., & Shatter, A. (2013). Directive 2013/37/EU of the European Parliament and of the Council of 26 June 2013 amending Directive 2003/98/EC on the re-use of public sector information. *Official Journal of the European Union, Brussels*. https://doi.org/10.5040/9781509923205.0008

Shen, J., Valagolam, D., & McCalla, S. (2020). Prophet forecasting model: A machine learning approach to predict the concentration of air pollutants (PM2. 5, PM10, O3, NO2, SO2, CO) in Seoul, South Korea. *PeerJ*, *8*, e9961. https://doi.org/10.7717/peerj.9961

Shidik, G. F., Ashari, A., Abdullah, S. B., Anandhi, R., Chitra, D., Mamun, A., ... & Khader, A. T. (2014). Linked open government data as background knowledge in predicting forest fire. *J. Inform*.

Song, S. H., & Kim, T. D. (2013, January). A study on the open platform modeling for linked open data ecosystem in public sector. In *2013 15th International Conference on Advanced Communications Technology (ICACT)* (pp. 730–734). IEEE.

Stagars, M. (2016). *Open data in Southeast Asia: Towards economic prosperity, government transparency, and citizen participation in the ASEAN*. Springer.

Stolfi, D. H., Alba, E., & Yao, X. (2020). Can I park in the city center? Predicting car park occupancy rates in smart cities. *Journal of Urban Technology*, *27*(4), 27–41. https://doi.org/10.1080/10630732.2019.1586223

Tamada, M. M., de Magalhães Netto, J. F., & de Lima, D. P. R. (2019, October). Predicting and reducing dropout in virtual learning using machine learning techniques: A systematic review.

In *2019 IEEE Frontiers in Education Conference (FIE)* (pp. 1–9). IEEE. https://doi.org/10.1109/fie43999.2019.9028545

Thabet, M., Ellouze, M., & Zaied, M. (2021). A new approach for video concept detection based on user comments. *Journal of Information and Communication Technology*, *20*(4), 629–649. https://doi.org/10.32890/jict2021.20.4.7

Tuke, J., Nguyen, A., Nasim, M., Mellor, D., Wickramasinghe, A., Bean, N., & Mitchell, L. (2020). Pachinko Prediction: A Bayesian method for event prediction from social media data. *Information Processing & Management*, *57*(2), 102147. https://doi.org/10.1016/j.ipm.2019.102147

Turban, E., King, D., Lee, J.K., Liang, T.P., & Turban, D.C. (2015). Social commerce: Foundations, social marketing, and advertising. In *Electronic commerce* (pp. 309–364). Springer, Cham. https://doi.org/10.1007/978-3-319-10091-3_7

Utsumi, M., Shigemori, I., & Watanabe, T. (2020, September). Forecasting Electricity Demand with Dynamic Characteristics Based on Signal Analysis and Machine Learning. In *2020 59th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE)* (pp. 1049–1054). IEEE. https://doi.org/10.23919/sice48898.2020.9240281

Violos, J., Pelekis, S., Berdelis, A., Tsanakas, S., Tserpes, K., & Varvarigou, T. (2019, February). Predicting visitor distribution for large events in smart cities. In *2019 IEEE International Conference on Big Data and Smart Computing (BigComp)* (pp. 1–8). IEEE. https://doi.org/10.1109/bigcomp.2019.8679181

W3C(e-Gov). (2009). *eGovernment at W3C: improving access to government through better use of the web*. http://www.w3.org/2007/eGov/. Last accessed 2019-04-01.

Weerakkody, V., Sivarajah, U., Mahroof, K., Maruyama, T., & Lu, S. (2021). Influencing subjective well-being for business and sustainable development using big data and predictive regression analysis. *Journal of Business Research*, *131*, 520–538. https://doi.org/10.1016/j.jbusres.2020.07.038

Weigert, A., Hopf, K., Weinig, N., & Staake, T. (2020). Detection of heat pumps from smart meter and open data. *Energy Informatics*, *3*(1), 1–14. https://doi.org/10.1186/s42162-020-00124-6

Wood, D. A. (2019). German solar power generation data mining and prediction with transparent open box learning network integrating weather, environmental and market variables. *Energy conversion and management*, *196*, 354–369. https://doi.org/10.1016/j.enconman.2019.05.114

Wright, F. (2014). Data. gov.

Wu, C. H., Kao, S. C., & Kan, M. H. (2017, July). Knowledge discovery in open data of dengue epidemic. In *Proceedings of the 4th Multidisciplinary International Social Networks Conference*(pp.1–8).https://doi.org/10.1145/3092090.3092093

Wu, J., Zhou, L., Cai, C., Dong, F., Shen, J., & Sun, G. (2019, October). Towards a general prediction system for the primary delay in urban railways. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)* (pp. 3482–3487). IEEE. https://doi.org/10.1109/itsc.2019.8916868

Xue, J. (2020, February). Financial risk prediction and evaluation model of P2P network loan platform. In *2020 12th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA)* (pp. 1060–1064). IEEE. https://doi.org/10.1109/icmtma50254.2020.00227

Yang, F., Han, X., Lang, J., Lu, W., Liu, L., Zhang, L., & Pan, J. (2018, October). Commodity Recommendation for Users Based on E-commerce Data. In *Proceedings of the 2nd International Conference on Big Data Research* (pp. 146–149). https://doi.org/10.1145/3291801.3291803

Yin, R. K. (2013). Validity and generalization in future case study evaluations. *Evaluation*, *19*(3), 321–332. https://doi.org/10.1177/1356389013497081

Yu, P. F., Huang, F. M., Yang, C., Liu, Y. H., Li, Z. Y., & Tsai, C. H. (2018, October). Prediction of crowdfunding project success with deep learning. In *2018 IEEE 15th International Conference on e-Business Engineering (ICEBE)* (pp. 1–8). IEEE. https://doi.org/10.1109/icebe.2018.00012

Zainudin, Z., & Shamsuddin, S. M. (2016). Predictive analytics in Malaysian dengue data from 2010 until 2015 using BigML. *International Journal of Advances in Soft Computing and its Applications*, *8*(3).

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education–where are the educators? *International Journal of Educational Technology in Higher Education*, *16*(1), 1–27. https://doi.org/10.1186/s41239-019-0171-0

Zelevinsky, V., Wang, J., & Tunkelang, D. (2008, October). Supporting exploratory search for the ACM digital library. In *Workshop on Human-Computer Interaction and Information Retrieval (HCIR 2008)* (pp. 85–88). https://doi.org/10.1061/9780784481301

Zou, Z., & Ergan, S. (2018). Impact of construction projects on urban quality of life: A data analysis method. In *Construction Research Congress 2018: Sustainable Design and Construction and Education, CRC 2018* (pp. 34–44). American Society of Civil Engineers (ASCE). https://doi.org/10.1061/9780784481301.004