Using AI on Open Data: A Systematic Literature Review

Sara Kijanović, Darko Stefanović, Dušanka Dakić, Teodora Vučković

University of Novi Sad, Faculty of Technical Sciences, Serbia

Department of Industrial Engineering and Engineering Management

Trg Dositeja Obradovica 6, Novi Sad, 21000

{sara.kijanovic, darko.stefanovic, dakic.dusanka, teodora.vuckovic}@uns.ac.rs

Abstract. *Open Data, publicly available in machine-readable formats, is a crucial resource driving transformative changes globally. This research uses a systematic literature review to explore the synergies between Open Data and the growth of artificial intelligence, including machine learning and deep learning. The goal was to investigate how integrating Open Data with AI/ML/DL techniques creates intelligent industry solutions. Using the PRISMA framework, the paper analyzes key findings, highlighting Open Data's impact on technological innovation. It also identifies primary data sources for these studies and provides background insights, methodology, results, and conclusions based on the review's insights.*

Keywords. *open data, open government data, artificial intelligence, machine learning, deep learning, literature review*

1 Introduction

According to the Law on Electronic Administration of the Republic of Serbia (*Zakon o elektronskoj upravi | Paragraf Lex propisi*, 2018), and also the Open Data Handbook (*What Is Open Data?*, 2024), Open Data (OD) are data that are publicly available for reuse, along with metadata, in a machine-readable and open format. Accessible worldwide, open data empowers governments, businesses, and entrepreneurs to ignite healthcare, education, and agriculture revolutions. These freely available data, accessible online, play a pivotal role in steering social and economic transformations. Innovations propelled by data contribute to heightened government transparency and efficiency, nurture business prospects, generate substantial economic benefits, and address urgent social challenges (*Open Data and AI: A Symbiotic Relationship for Progress | Data.Europa.Eu*, 2023).

We are witnessing the exponential development of artificial intelligence and various segments of AI, such as machine learning (ML), deep learning (DL), natural language processing (NLP) etc. Today, these technologies enhance and boost all aspects of human life. In the context of large amounts of data, advanced methods and techniques can be used for the analysis of such data, obtaining meaningful outputs, as well as intelligent solutions that can find practical applications in everyday life. For instance, advanced technologies can significantly support business processes and process mining, given that this field requires extracting knowledge from vast amounts of data (Dakic et al., 2019).

As mentioned, open data holds tremendous potential as a source of information for creating solutions that can improve all aspects of life. In scientific literature, many authors encounter the problem of data scarcity, which slows down their research. In many scientific papers within the context of humanities and medical sciences, authors such as (Adler et al., 2022; Norori et al., 2021), emphasize the importance of sharing data and materials for the easier implementation of technologies and the detection of anomalies, leading to faster treatments. The authors (Rosés et al., 2021) emphasize the importance of opening various datasets for further research to make progress in the field of crime detection more efficient. This is done to enable the development of spatial crime simulations that contribute to understanding the mechanisms that drive crime (Rosés et al., 2021).

This paper represents a systematic literature review (SLR) providing insights into the current state of the field in the context of open data usage, exploring where researchers gather data and which contemporary techniques from the AI, ML, and DL domains they employ to create intelligent solutions. To achieve the defined research objectives, SLR was conducted using Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA). The main contribution of this paper is to gain insights into the previous application of AI/ML/DL methods and techniques on OD, as well as to identify the primary data sources for these studies over 5 years (2018-2023). Consequently, this may be significant for researchers in the future when it comes to easier identification of suitable datasets. Key research questions were formulated, along with inclusion and exclusion criteria for the works. The results are presented in the paper.

The rest of the paper is organized into the following sections: Section 2 introduces the background, providing an overview of the fundamental concept of OD and advanced methods and techniques in the context of AI/ML/DL. Section 3 presents related work, Section 4 describes the methodology, Section 5 presents the research results, then Section 6 presents a discussion. Finally, in Section 7, conclusions are drawn based on the results obtained and the review conducted.

2 Background

As mentioned previously, Open Data (OD) represents data that is publicly available, accompanied by metadata, in formats that enable its use and reuse. The OD initiative fosters the development of various sectors and promotes data accessibility for everyone's use. Its goal is to encourage transparency and enable equal participation in problem-solving and innovation for all. OD can be used in many fields, and most importantly in the public sector, like public finance researched in (Komosar & Kijanovic, 2023), and education (Kijanovic & Komosar, 2023). Many papers, like those mentioned, provide background on the importance of using OD and highlight the significance of utilizing OD for various purposes.

Governments worldwide often release data, known as Open Government Data (OGD). Many researchers utilize open data from various platforms, and the authors in the paper mention that frequently used platforms include Kaggle, GitHub, Socrata, CKAN, DKAN, etc. (Ali et al., 2022).

The latest reports from the European Union emphasize the immense potential of artificial intelligence, especially when combined with open data. This synergy unlocks new possibilities, not only for extracting fresh insights from open data but also for leveraging AI in novel applications (*Open Data and AI: A Symbiotic Relationship for Progress | Data.Europa.Eu*, 2023).

Open data enables researchers to utilize extensive datasets for testing and scaling their studies and algorithms, overcoming the challenge of accessing real databases. The lack of access or insufficient data can significantly hinder the development of data-driven technologies while harnessing this power can enhance various industrial processes.

Data and technology enhance the industry. Authors (Lopes et al., 2024) stress the importance of precise production data for digital twin models. They tackled data scarcity by simulating production lines and generating artificial data to support deep neural network algorithms. Open data not only promote transparency and collaboration in the research community but also serves as a valuable foundation for advancing industry technologies.

In the subsequent work, insights into the current state of utilizing AI, ML, or DL with OD, using publicly available and open datasets, will be aimed to be gained.

3 Related Work

The paper (Hurbean et al., 2021) conducts an SLR on the integration of OD and ML in smart cities (SC). It examines ML applications across six smart city domains using OD, identifying trends, challenges, and potential solutions. It concludes that OD-based ML applications show promise in addressing urban challenges but also highlight data quality and consistency challenges.

The authors in (Wirtz et al., 2022) explore the link between open government data (OGD) and the digital economy, emphasizing OGD's potential for innovation and economic growth. Conducting an SLR synthesizes existing knowledge into a theoretical framework and suggests a research agenda for information systems and digital business research. Its conclusions stress OGD's importance for driving innovation and economic value, highlighting a research gap in integrating OGD concepts with digital business research.

The paper (Brinkhaus et al., 2023) concludes that the future of AI applications in chemistry relies heavily on open data initiatives. It emphasizes the necessity of open data infrastructures. Additionally, it underscores the importance of open-source software and molecular string representations in advancing AI research in chemistry. The study suggests that sharing curated data and trained models with the public will accelerate growth in the field. It calls for maintaining high data standards and leveraging public cloud infrastructures to expedite advancements in AI-powered molecular informatics.

The presented papers provide insights and answers to research questions that are concretized through specific examples. Given that no literature review has been conducted so far that could cover general questions applicable to each area, one has been carried out below.

4 Methodology

A systematic literature review (SLR) serves the purpose of systematically analyzing, synthesizing, and evaluating available research and papers in a specific field. This approach allows researchers to create a comprehensive overview of relevant studies and literature on a particular topic. An SLR aids in identifying key findings and shortcomings in research and provides a foundation for further investigation.

For this SLR, we used methodology based on Barbara Kitchenham's guidelines for performing Systematic Literature Reviews (Kitchenham &

Charters, 2007), and Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA). According to (Moher et al., 2009), it consists of four phases: Identification, Screening, Eligibility, and Included phase.

4.1 Search Strategy

We used the Scopus index database for this literature review with a query string:

```
("artificial intelligence" OR "AI" OR "Machine learning" OR "ML" OR "Deep Learning" OR "DL")
AND
("Open Data" OR "Open Government Data")
AND
PUBYEAR > 2017
```

Access was granted to open-access papers, meaning they are under a specific Creative Commons license.

4.2 Research questions and inclusion and exclusion criteria

To gain a clear understanding and achieve the objective of this review, we have formulated four research questions:

1. Which methods and techniques of AI/ML/DL are predominantly used with OD? (RQ1)
2. Which databases are most commonly used for downloading OD? (RQ2)
3. What are the strengths and limitations of different AI techniques in OD? (RQ3)
4. What is the main purpose of research and the use of AI/ML/DL methods and techniques on the utilized open data? (RQ4)

To enhance the SLR's effectiveness, we defined inclusion and exclusion criteria. The inclusion criteria are:

1. The paper must primarily focus on applying various AI, ML, and DL methods and techniques to OD. (IC1)
2. The paper must contain the source of the Open Dataset, e.g. reference, name of official source/platform, or similar. (IC2)
3. The paper must describe a method or technique applied to OD, provide insight into the algorithm, or include a link to the code in open form. (IC3)
4. The paper that qualitatively describes the measured performance of the applied method or technique. (IC4)
5. The paper must be published after 2017. (IC5)
6. The paper must be written in English. (IC6)
7. The paper must be an article or conference paper. (IC7)
8. Paper must be open accessed. (IC8)

The exclusion criteria are:

1. Papers that do not contain the materials needed to answer the research questions. (EC1)
2. If the paper is centered around ML/AI/DL or another technique but does not present a solution related to OD, it will be excluded. (EC2)

4.3 SLR Workflow based on PRISMA Workflow

Based on the query string, we conducted the PRISMA workflow shown in Fig. 1. The query result on the Scopus database listed 705 papers, and we excluded 77 based on paper type and 16 based on language. In the screening phase, we had 612 papers, where we excluded 272 papers based on abstract and title. In the next step of this phase, we screened full-text papers, 340 ones, and excluded 44 papers based on open access even though we marked open access on Scopus. And last, we excluded 260 papers based on exclusion and inclusion criteria. In the end, a total of 36 papers were included in the review.

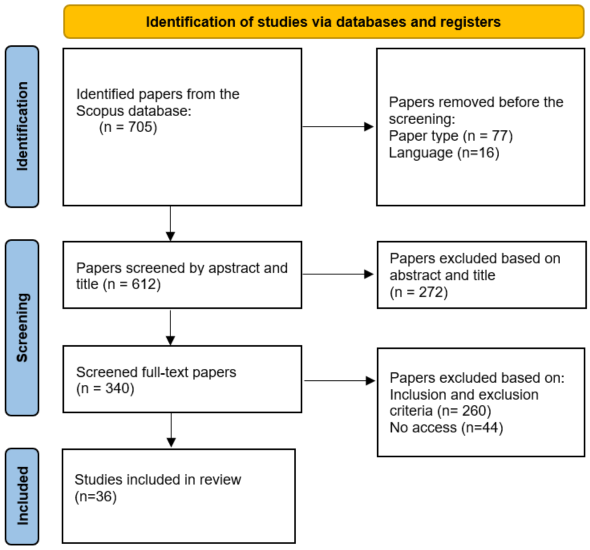

Figure 1. PRISMA Workflow diagram

5 Results

Within this section, the results obtained through the literature review will be presented. A significant increase in the number of papers is observed, especially in the year 2023 (33,3%), where the majority of selected papers for review were published in that year. Additionally, most of the papers are Journal Articles. This can be correlated with the development of AI/ML/DL in recent years and the increasing availability of OD.

Since visualization is crucial and aids in simplifying the conduct of SLR (Stefanovic et al., 2022), the RQ1, RQ2, and RQ4 have been answered in a tabular format.

To obtain an answer to RQ1, Table 1 was created, which roughly describes the methods and techniques mentioned in the primary papers on open data. In certain papers, multiple methods and techniques were employed, allowing for an experimental insight into which ones exhibit greater efficiency.

The majority of selected papers in their research utilize CNN, precisely 35,3% of them. The papers provide detailed explanations of the architectures used, with the highest number of works employing variations of U-Net (Garcia-Pedrero et al., 2019; Mainali et al., 2023; Rutherford et al., 2022; Stewart et al., 2020). Subsequently, researchers predominantly employ Random Forests for regression and classification. As can be seen in the results, many researchers use different machine learning methods in the context of regression and classification to test which ones perform better on the data and yield better results. 14,71% were used for clustering.

Table 1. Utilized methods and techniques in primary studies

Subfield of AI	Methods/techniques		Primary studies	N	%
ML	Regression models	Random Forest (RF)	(Boeke et al., 2019; Meneghetti et al., 2023; Nai et al., 2023; Pareeth et al., 2019; Pargent et al., 2023; Shulajkovska et al., 2023)	15	22,06%
		Support Vector Regressor (SVR)	(Boeke et al., 2019; Shulajkovska et al., 2023)		
		Logistic Regression (LR)	(Ma & Faye, 2022; Shulajkovska et al., 2023)		
		Linear Regression (LNR)	(Cocca et al., 2020; Shulajkovska et al., 2023)		
		Elastic Net	(Shulajkovska et al., 2023)		
		Gradient Boosting Regressor	(Shulajkovska et al., 2023)		
		Bayesian Ridge (BR)	(Shulajkovska et al., 2023)		
	Classification models	Random Forest	(Boeke et al., 2019; Ma & Faye, 2022)	11	16,18%
		Support Vector Machine (SVM)	(Boeke et al., 2019; Ma & Faye, 2022; Mizuno et al., 2022)		
		Support Vector Regression (SVR)	(Shulajkovska et al., 2023)		
		K-nearest neighbor (k-NN)	(Boeke et al., 2019; Shulajkovska et al., 2023)		
		Adaboost	(Ma & Faye, 2022)		
		Extreme tree classifier (ETC)	(Verhegghen et al., 2022)		
		Decision Tree (DT)	(Shulajkovska et al., 2023)		
	Clustering models	K-means	(Mizuno et al., 2022; Trento Oliveira et al., 2023)	10	14,71%
		DBSCAN	(Carter et al., 2021; Mazurek & Hachaj, 2021)		
		OPTICS	(Mazurek & Hachaj, 2021)		
		Gradient Boosting Classifier (GBM)	(Ruiz-Rizzo et al., 2022; Veigel et al., 2023)		
		LightGBM	(H. Liu & Ma, 2023; Ruiz-Rizzo et al., 2022)		
		XGBoost	(Verhegghen et al., 2022)		
	Visual pattern mining (VPM)		(Balaniuk et al., 2020)	1	1,47%
	Extremely Randomized Tree (ERT)		(Jin et al., 2022)	1	1,47%
	Stochastic Gradient Descent (SGD)		(Shulajkovska et al., 2023)	1	1,47%
DL	Deep learning-based side-channel analysis		(Imafuku et al., 2023)	1	1,47%
	GNN		(Holmberg et al., 2023)	1	1,47%
	ANN		(Sulavko, 2022)	1	1,47%
	RNN		(Rajan et al., 2020)	2	2,94%
		LSTM	(Ma & Faye, 2022)		
	CNN	DeepLavV3+	(Edpuganti et al., 2023; Touzani & Granderson, 2021)	24	35,3%
		Faster R-CNN	(Chen et al., 2021)		
		YOLO	(Chen et al., 2021; Mazurek & Hachaj, 2021)		
		DarkCovidNet	(Harkness et al., 2022)		
		COVID-Net	(Harkness et al., 2022)		
		CoroNet	(Harkness et al., 2022)		
		U-Net	(Garcia-Pedrero et al., 2019; Mainali et al., 2023; Rutherford et al., 2022; Stewart et al., 2020)		
		Mask R-CNN	(Carter et al., 2021)		
		RS-NET	(Jeppesen et al., 2019)		
		Twin CNN	(P. Liu et al., 2019)		
		Multilayer CNN	(Sulavko, 2022)		
		LUN-BiSeNetV2	(Zhang & Zhang, 2023)		
		MobileNet	(Mazurek & Hachaj, 2021)		
		RetinaNet	(Mazurek & Hachaj, 2021)		
		MAML-CNN-LSTM-Attention-based model	(Xu et al., 2023)		
		Fully CNN (FCN)	(Balaniuk et al., 2020)		
		No concrete type specified	(Berlanga et al., 2022; Fibæk et al., 2021; Rajan et al., 2020)		

In summary, multiple ML models and techniques are used, highlighting that many researchers use a variety of them in their work for comparisons and similar purposes. As mentioned earlier, DL methods, are predominantly used, often combined with some other ML methods alongside convolutional neural networks. RQ2 is intended to provide an answer to the most common data source for researchers within the selected papers. Table 2 provides an insight into the answer to this question, and the papers were selected based on the origin of the data.

If the datasets are from official portals of cities, countries, or any institutions related to the state and are state-owned, they are presented as such. All individual datasets are presented separately. Each referenced paper indicates where the data was collected from. The answer to RQ2 can be seen in Table 2, indicating that researchers most commonly find relevant datasets on official open data portals belonging to public and state institutions. These data can then be partly classified as open government data. Indeed, in practice, it is observed that governments worldwide are creating many initiatives for data openness. In addition to these portals, Kaggle has significantly assisted researchers in finding data in the selected papers, and other sources are also listed. This insight into sources is significant in aiding researchers in easily locating data that is in an open format and available for sharing, distribution, and reuse.

It is important to emphasize that, based on the inclusion and exclusion criteria, all reviewed articles are publicly available, and each one provides the exact data source, which can be useful for researchers.

To address RQ3, focusing on the strengths and limitations of AI applied to open data, researchers have consistently highlighted several critical aspects. The majority have mentioned challenges related to data quality (Cocca et al., 2020; Meneghetti et al., 2023; Mizuno et al., 2022; Veigel et al., 2023), followed by data inhomogeneity (Moher et al., 2009) and limitations in the available data (Rutherford et al., 2022; Stewart et al., 2020). Additionally, the authors emphasize the necessity for statistical analysis to assess the accuracy and reliability of models (Shulajkovska et al., 2023). Concerns regarding model obsolescence (Cashman et al., 2016; Hurbean et al., 2021) and issues related to generalization (Mizuno et al., 2022).

Table 2. Source of datasets

Source of dataset	Primary studies	%
Government or public institutions	(Carter et al., 2021; Chen et al., 2021; Cocca et al., 2020; Ma & Faye, 2022; Meneghetti et al., 2023; Mizuno et al., 2022; Ruiz-Rizzo et al., 2022; Veigel et al., 2023; Zhang & Zhang, 2023)	25,2%
Airconstructor	(Sulavko, 2022)	2,8%
AmesHousing; Titanic	(Pargent et al., 2023)	2,8%
ASCAD	(Imafuku et al., 2023)	2,8%
ANAC	(Nai et al., 2023)	2,8%
CERN OD Portal	(Holmberg et al., 2023)	2,8%
Copernicus	(Balaniuk et al., 2020; Verhegghen et al., 2022)	5,6%
GDSC	(P. Liu et al., 2019)	2,8%
GIS	(Touzani & Granderson, 2021)	2,8%
Kaggle	(Edpuganti et al., 2023; Rajan et al., 2020; Stewart et al., 2020; Xu et al., 2023)	11,2%
Landstat 8	(Jeppesen et al., 2019; Pareeth et al., 2019)	5,6%
Kitti	(Mazurek & Hachaj, 2021)	2,8%
LPIS	(Garcia-Pedrero et al., 2019)	2,8%
ODR	(Berlanga et al., 2022)	2,8%
OPENAQ	(Jin et al., 2022)	2,8%
OpenNeuro	(Rutherford et al., 2022)	2,8%
OSF	(H. Liu & Ma, 2023; Mainali et al., 2023)	5,6%
RICORD	(Harkness et al., 2022)	2,8%
Sentinel 1/2	(Fibæk et al., 2021)	2,8%
Urbanite	(Shulajkovska et al., 2023)	2,8%
WorldPop	(Trento Oliveira et al., 2023)	2,8%
ICEAS	(Boeke et al., 2019)	2,8%

The identified key strengths of implementing AI on OD include the enhancement of public administration systems through machine learning (Nai et al., 2023), machine learning models providing information about factors predicting outcomes while offering specific predictions for each case (Ruiz-Rizzo et al., 2022), and the use of open data that additionally supports research in this field (Jin et al., 2022).

To answer RQ4, papers and their purposes were observed, i.e., the purposes for which they used different methods and techniques on open data were examined. Table 3 provides an insight into the focus of the selected papers. All papers are grouped into several categories, depending on whether they propose a new solution (method, model, approach), are of a research nature, or involve analysis, comparison, or enhancement of existing systems. Some papers deal with the development of a new approach or method.

The majority of papers focus on proposing a new approach, model, or method, which is certainly since AI/ML/DL methods and techniques are quite promising and applicable in practice. Considering the huge potential of open data, the application of modern methods and techniques with open data can significantly influence the development and proposal of new methods, models, and approaches to problem-solving. Moreover, many research papers use open data, as well as analyses and comparisons of existing methods and techniques.

Additionally, it is important to emphasize that most of the papers are published in journals with a good impact factor (IF). Some of the papers are published in journals with excellent impact factors, such as (Mizuno et al., 2022) with 12,4 IF, and (Fibæk et al., 2021) with 7,6 IF.

Table 3. Research goal/purpose

Purpose		Primary studies	%
Proposing	Method	(Chen et al., 2021; Jin et al., 2022; Mizuno et al., 2022; Stewart et al., 2020; Zhang & Zhang, 2023)	14%
	Model	(Sulavko, 2022; Xu et al., 2023)	5,6%
	Approach	(Fibæk et al., 2021; Ma & Faye, 2022; Mazurek & Hachaj, 2021; Touzani & Granderson, 2021)	11,2%
Exploratory		(Balaniuk et al., 2020; Carter et al., 2021; Edpuganti et al., 2023; Garcia-Pedrero et al., 2019; Holmberg et al., 2023; Nai et al., 2023; Ruiz-Rizzo et al., 2022; Veigel et al., 2023)	22,4%
Developing approach/method		(Berlanga et al., 2022; Pareeth et al., 2019; Shulajkovska et al., 2023)	8,4%
Analysis		(Harkness et al., 2022; Mainali et al., 2023; Meneghetti et al., 2023; Rajan et al., 2020; Trento Oliveira et al., 2023; Verhegghen et al., 2022)	16,8%
Comparison		(Boeke et al., 2019; Cocca et al., 2020)	5,6%
Improvement of already existing solutions		(Imafuku et al., 2023; H. Liu & Ma, 2023; P. Liu et al., 2019; Rutherford et al., 2022)	11,2%
Other		(Jeppesen et al., 2019; Pargent et al., 2023)	5,6%

6 Discussion

The results obtained after the SLR, which were presented in the previous chapter, will be briefly discussed in this section.

It can be concluded that there has been a trend in the development and application of AI/ML/DL methods and techniques on open data in recent years. This is inevitably influenced by the development of information technologies and the increasing availability of OD. Primarily, institutions open portals for open data, but the research indicates that there are also other relevant sources assisting researchers in using, distributing, and downloading data, as shown in Table 2. Additionally, some platforms have collected data, combined it from other open portals, and enabled free usage, such as Kaggle. This way, the answer to RQ2 was obtained.

Regarding RQ1, DL methods like CNN and other neural networks, along with ML models for regression, classification, and clustering (especially RF), are most commonly used.

Addressing the problems that researchers encounter, as shown through RQ3, and improving open data sets and their sharing can lead to the development and use of AI/ML/DL on OD, and such solutions provide additional value. As shown through RQ4, there is an increasing number of papers proposing new solutions addressing the problems highlighted by researchers, and this segment can be further improved.

This paper enhances our understanding of the synergy between AI and OD and shows how this synergy can be applied in industry and society to create value and innovation.

7 Conclusion

In recent years, there has been significant development and awareness of the importance of Open Data and its potential for implementing smart solutions with advanced technologies. Researchers often face the challenge of finding high-quality, relevant data. Open data not only enhances transparency but also drives progress in research and industry.

To date, a detailed literature review on the main data sources, methods, and techniques used with advanced technologies has been lacking. This paper addresses this gap by analyzing trends and aiding in the discovery of various datasets. The review highlights the increasing use of AI/ML/DL in processing and extracting knowledge from these datasets. It provides code and algorithm descriptions, offering insights for future researchers.

The findings from this literature review shed light on the current state of the field and support the further development and application of advanced technologies on open data, aiming to improve various industry processes.

Acknowledgment

This research has been supported by the Ministry of Science, Technological Development and Innovation (Contract No. 451-03-65/2024-03/200156) and the Faculty of Technical Sciences, University of Novi Sad through the project “Scientific and Artistic Research Work of Researchers in Teaching and Associate Positions at the Faculty of Technical Sciences, University of Novi Sad” (No. 01-3394/1).

References

Adler, D. A., Wang, F., Mohr, D. C., Estrin, D., Livesey, C., & Choudhury, T. (2022). A call for open data to develop mental health digital biomarkers. *BJPsych Open*, *8*(2), e58. https://doi.org/10.1192/bjo.2022.28

Ali, M., Alexopoulos, C., & Charalabidis, Y. (2022). A comprehensive review of open data platforms, prevalent technologies, and functionalities. *15th International Conference on Theory and Practice of Electronic Governance*, 203–214. https://doi.org/10.1145/3560107.3560142

Balaniuk, R., Isupova, O., & Reece, S. (2020). Mining and Tailings Dam Detection in Satellite Imagery Using Deep Learning. *Sensors*, *20*(23), 6936. https://doi.org/10.3390/s20236936

Berlanga, G., Williams, Q., & Temiquel, N. (2022). Convolutional Neural Networks as a Tool for Raman Spectral Mineral Classification Under Low Signal, Dusty Mars Conditions. *Earth and Space Science*, *9*(10), e2021EA002125. https://doi.org/10.1029/2021EA002125

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 the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *XLII-4/W19*, 63–70. https://doi.org/10.5194/isprs-archives-XLII-4-W19-63-2019

Brinkhaus, H. O., Rajan, K., Schaub, J., Zielesny, A., & Steinbeck, C. (2023). Open data and algorithms for open science in AI-driven molecular informatics. *Current Opinion in Structural Biology*, *79*, 102542. https://doi.org/10.1016/j.sbi.2023.102542

Carter, B. P., Blackadar, J. H., & Conner, W. L. A. (2021). When Computers Dream of Charcoal: Using Deep Learning, Open Tools, and Open Data to Identify Relict Charcoal Hearths in and around State Game Lands in Pennsylvania. *Advances in Archaeological Practice*, *9*(4), 257–271. https://doi.org/10.1017/aap.2021.17

Cashman, S. A., Meyer, D. E., Edelen, A. N., Ingwersen, W. W., Abraham, J. P., Barrett, W. M., Gonzalez, M. A., Randall, P. M., Ruiz-Mercado, G., & Smith, R. L. (2016). Mining Available Data from the United States Environmental Protection Agency to Support Rapid Life Cycle Inventory Modeling of Chemical Manufacturing. *Environmental Science & Technology*, *50*(17), 9013–9025. https://doi.org/10.1021/acs.est.6b02160

Chen, L., Grimstead, I., Bell, D., Karanka, J., Dimond, L., James, P., Smith, L., & Edwardes, A. (2021). Estimating Vehicle and Pedestrian Activity from Town and City Traffic Cameras. *Sensors*, *21*(13), 4564. https://doi.org/10.3390/s21134564

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

Dakic, D., Sladojevic, S., Lolic, T., & Stefanovic, D. (2019). Process Mining Possibilities and Challenges: A Case Study. *2019 IEEE 17th International Symposium on Intelligent Systems and Informatics (SISY)*, 000161–000166. https://doi.org/10.1109/SISY47553.2019.9111591

Edpuganti, A., Akshaya, P., Gouthami, J., Sajith Variyar, V. V., Sowmya, V., & Sivanpillai, R. (2023). EFFECT OF DATA QUALITY ON WATER BODY SEGMENTATION WITH DEEPLABV3+ ALGORITHM. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *XLVIII-M-3–2023*, 81–85. https://doi.org/10.5194/isprs-archives-XLVIII-M-3-2023-81-2023

Fibæk, C. S., Keßler, C., & Arsanjani, J. J. (2021). A multi-sensor approach for characterising human-made structures by estimating area, volume and population based on sentinel data and deep learning. *International Journal of Applied Earth Observation and Geoinformation*, *105*, 102628. https://doi.org/10.1016/j.jag.2021.102628

Garcia-Pedrero, A., Lillo-Saavedra, M., Rodriguez-Esparragon, D., & Gonzalo-Martin, C. (2019). Deep Learning for Automatic Outlining Agricultural Parcels: Exploiting the Land Parcel Identification System. *IEEE Access*, *7*, 158223–158236. https://doi.org/10.1109/ACCESS.2019.2950371

Harkness, R., Hall, G., Frangi, A. F., Ravikumar, N., & Zucker, K. (2022). The Pitfalls of Using Open Data to Develop Deep Learning Solutions for COVID-19 Detection in Chest X-Rays. In P. Otero, P. Scott, S. Z. Martin, & E. Huesing (Eds.), *Studies in Health Technology and Informatics*. IOS Press. https://doi.org/10.3233/SHTI220164

Holmberg, D., Golubovic, D., & Kirschenmann, H. (2023). Jet Energy Calibration with Deep Learning as a Kubeflow Pipeline. *Computing and Software for Big Science*, *7*(1), 9. https://doi.org/10.1007/s41781-023-00103-y

Hurbean, L., Danaiata, D., Militaru, F., Dodea, A.-M., & Negovan, A.-M. (2021). Open Data Based Machine Learning Applications in Smart Cities: A Systematic Literature Review. *Electronics*, *10*(23), 2997. https://doi.org/10.3390/electronics10232997

Imafuku, K., Kawamura, S., Nozaki, H., Sakamoto, J., & Osuka, S. (2023). Non-Profiled Deep Learning-Based Side-Channel Analysis With Only One Network Training. *IEEE Access*, *11*, 83221–83231. https://doi.org/10.1109/ACCESS.2023.3301178

Jeppesen, J. H., Jacobsen, R. H., Inceoglu, F., & Toftegaard, T. S. (2019). A cloud detection algorithm for satellite imagery based on deep learning. *Remote Sensing of Environment*, *229*, 247–259. https://doi.org/10.1016/j.rse.2019.03.039

Jin, C., Wang, Y., Li, T., & Yuan, Q. (2022). Global validation and hybrid calibration of CAMS and MERRA-2 PM2.5 reanalysis products based on OpenAQ platform. *Atmospheric Environment*, *274*, 118972. https://doi.org/10.1016/j.atmosenv.2022.118972

Kijanovic, S., & Komosar, A. (2023). Research on Open Data Practice Within Education. *2023 22nd International Symposium INFOTEH-JAHORINA (INFOTEH)*, 1–6. https://doi.org/10.1109/INFOTEH57020.2023.10094111

Kitchenham, B., & Charters, S. (2007). *Guidelines for performing Systematic Literature Reviews in Software Engineering*. *2*.

Komosar, A., & Kijanovic, S. (2023). Research on the Application of Open Data in the Public Finance Sector. *2023 22nd International Symposium INFOTEH-JAHORINA (INFOTEH)*, 1–6. https://doi.org/10.1109/INFOTEH57020.2023.10094103

Liu, H., & Ma, E. (2023). An Explainable Evaluation Model for Building Thermal Comfort in China. *Buildings*, *13*(12), 3107. https://doi.org/10.3390/buildings13123107

Liu, P., Li, H., Li, S., & Leung, K.-S. (2019). Improving prediction of phenotypic drug response on cancer cell lines using deep convolutional network. *BMC Bioinformatics*, *20*(1), 408. https://doi.org/10.1186/s12859-019-2910-6

Lopes, P. V., Silveira, L., Guimaraes Aquino, R. D., Ribeiro, C. H., Skoogh, A., & Verri, F. A. N. (2024). Synthetic data generation for digital twins: Enabling production systems analysis in the absence of data. *International Journal of Computer Integrated Manufacturing*, 1–18. https://doi.org/10.1080/0951192X.2024.2322981

Ma, T.-Y., & Faye, S. (2022). Multistep electric vehicle charging station occupancy prediction using hybrid LSTM neural networks. *Energy*, *244*, 123217. https://doi.org/10.1016/j.energy.2022.123217

Mainali, K., Evans, M., Saavedra, D., Mills, E., Madsen, B., & Minnemeyer, S. (2023). Convolutional neural network for high-resolution wetland mapping with open data: Variable selection and the challenges of a generalizable model. *Science of The Total Environment*, *861*, 160622. https://doi.org/10.1016/j.scitotenv.2022.160622

Mazurek, P., & Hachaj, T. (2021). SLAM-OR: Simultaneous Localization, Mapping and Object Recognition Using Video Sensors Data in Open Environments from the Sparse Points Cloud. *Sensors*, *21*(14), 4734. https://doi.org/10.3390/s21144734

Meneghetti, N., Pacini, F., Biondi Dal Monte, F., Cracchiolo, M., Rossi, E., Mazzoni, A., & Micera, S. (2023). Predicting party switching through machine learning and open data. *iScience*, *26*(7), 107098. https://doi.org/10.1016/j.isci.2023.107098

Mizuno, S., Ohba, H., & Ito, K. (2022). Machine learning-based turbulence-risk prediction method for the safe operation of aircrafts. *Journal of Big Data*, *9*(1), 29. https://doi.org/10.1186/s40537-022-00584-5

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *BMJ*, *339*, b2535. https://doi.org/10.1136/bmj.b2535

Nai, R., Meo, R., Morina, G., & Pasteris, P. (2023). Public tenders, complaints, machine learning and recommender systems: A case study in public administration. *Computer Law & Security Review*, *51*, 105887. https://doi.org/10.1016/j.clsr.2023.105887

Norori, N., Hu, Q., Aellen, F. M., Faraci, F. D., & Tzovara, A. (2021). Addressing bias in big data and AI for health care: A call for open science. *Patterns*, *2*(10), 100347. https://doi.org/10.1016/j.patter.2021.100347

Open data and AI: A symbiotic relationship for progress | data.europa.eu. (2023, June 9). https://data.europa.eu/en/publications/datastories/open-data-and-ai-symbiotic-relationship-progress

Pareeth, S., Karimi, P., Shafiei, M., & De Fraiture, C. (2019). Mapping Agricultural Landuse Patterns from Time Series of Landsat 8 Using Random Forest Based Hierarchial Approach. *Remote Sensing*, *11*(5), 601. https://doi.org/10.3390/rs11050601

Pargent, F., Schoedel, R., & Stachl, C. (2023). Best Practices in Supervised Machine Learning: A Tutorial for Psychologists. *Advances in Methods and Practices in Psychological Science*, *6*(3), 25152459231162559. https://doi.org/10.1177/25152459231162559

Rajan, K., Zielesny, A., & Steinbeck, C. (2020). DECIMER: Towards deep learning for chemical image recognition. *Journal of Cheminformatics*, *12*(1), 65. https://doi.org/10.1186/s13321-020-00469-w

Rosés, R., Kadar, C., & Malleson, N. (2021). A data-driven agent-based simulation to predict crime patterns in an urban environment. *Computers, Environment and Urban Systems*, *89*, 101660. https://doi.org/10.1016/j.compenvurbsys.2021.101660

Ruiz-Rizzo, A. L., Archila-Meléndez, M. E., & González Veloza, J. J. F. (2022). Predicting the probability of finding missing older adults based on machine learning. *Journal of Computational Social Science*, *5*(2), 1303–1321. https://doi.org/10.1007/s42001-022-00171-x

Rutherford, S., Sturmfels, P., Angstadt, M., Hect, J., Wiens, J., Van Den Heuvel, M. I., Scheinost, D., Sripada, C., & Thomason, M. (2022). Automated Brain Masking of Fetal Functional MRI with Open Data. *Neuroinformatics*, *20*(1), 173–185. https://doi.org/10.1007/s12021-021-09528-5

Shulajkovska, M., Smerkol, M., Dovgan, E., & Gams, M. (2023). A machine-learning approach to a mobility policy proposal. *Heliyon*, *9*(10), e20393. https://doi.org/10.1016/j.heliyon.2023.e20393

Stefanovic, D., Havzi, S., Spasojevic, I., Lolic, T., & Ristic, S. (2022). Software Tools to Support Visualising Systematic Literature Review. In B. Lalic, D. Gracanin, N. Tasic, & N. Simeunović (Eds.), *Proceedings on 18th International Conference on Industrial Systems – IS'20* (pp. 79–86). Springer International Publishing. https://doi.org/10.1007/978-3-030-97947-8_11

Stewart, C., Lazzarini, M., Luna, A., & Albani, S. (2020). Deep Learning with Open Data for Desert Road Mapping. *Remote Sensing*, *12*(14), 2274. https://doi.org/10.3390/rs12142274

Sulavko, A. (2022). Biometric-Based Key Generation and User Authentication Using Acoustic Characteristics of the Outer Ear and a Network of Correlation Neurons. *Sensors*, *22*(23), 9551. https://doi.org/10.3390/s22239551

Touzani, S., & Granderson, J. (2021). Open Data and Deep Semantic Segmentation for Automated Extraction of Building Footprints. *Remote Sensing*, *13*(13), 2578. https://doi.org/10.3390/rs13132578

Trento Oliveira, L., Kuffer, M., Schwarz, N., & Pedrassoli, J. C. (2023). Capturing deprived areas using unsupervised machine learning and open data: A case study in São Paulo, Brazil. *European Journal of Remote Sensing*, *56*(1), 2214690. https://doi.org/10.1080/22797254.2023.2214690

Veigel, N., Kreibich, H., & Cominola, A. (2023). Interpretable Machine Learning Reveals Potential to Overcome Reactive Flood Adaptation in the Continental US. *Earth's Future*, *11*(9), e2023EF003571. https://doi.org/10.1029/2023EF003571

Verhegghen, A., Kuzelova, K., Syrris, V., Eva, H., & Achard, F. (2022). Mapping Canopy Cover in African Dry Forests from the Combined Use of Sentinel-1 and Sentinel-2 Data: Application to Tanzania for the Year 2018. *Remote Sensing*, *14*(6), 1522. https://doi.org/10.3390/rs14061522

What is Open Data? (2024, October 8). https://opendatahandbook.org/guide/en/what-is-open-data/

Wirtz, B. W., Weyerer, J. C., Becker, M., & Müller, W. M. (2022). Open government data: A systematic literature review of empirical research. *Electronic Markets*, *32*(4), 2381–2404. https://doi.org/10.1007/s12525-022-00582-8

Xu, M., Ren, H., Chen, P., & Xin, G. (2023). On the V2G capacity of shared electric vehicles and its forecasting through MAML-CNN-LSTM-Attention algorithm. *IET Generation, Transmission & Distribution*, gtd2.12921. https://doi.org/10.1049/gtd2.12921

Zakon o elektronskoj upravi | Paragraf Lex propisi. (2018, April 6). http://www.paragraf.rs/propisi/zakon-o-elektronskoj-upravi-republika-srbija.html

Zhang, Y., & Zhang, M. (2023). LUN-BiSeNetV2: A lightweight unstructured network based on BiSeNetV2 for road scene segmentation. *Computer Science and Information Systems*, *20*(4), 1749–1770. https://doi.org/10.2298/CSIS221205029Z