STM JOURNALS

Journal of
Artificial Intelligence
Research & Advances

ISSN: 2395-6720
Volume 11, Issue 1, 2024
January–April
DOI Journal: 10.37591/JoAIRA

https://journals.stmjournals.com/joaira

Review JoAIRA

Applications of Machine Learning Algorithms in Health Data Science (HDS) for Next Research Directions: A Survey Report

Vinay Bhatt[1,*], Mayank Kumar[2]

Abstract

At present time, data science is the big trend in computer science. The functioning of this technology is purely based on other advanced technology known as machine learning (ML). Data science and ML are subsets of artificial intelligence (AI). When a process of data science is used in healthcare systems, the new system is known as health data science (HDS). HDS is a branch of data science used to handle the large amount of data in the healthcare system. Recently, data science has been used to handle and analyze large volumes of data (structured or unstructured) with accuracy by using different techniques with algorithms of ML. This survey paper presented the ML applications in data science using different previous research. In this paper, firstly discuss the introduction of the paper with related information, secondly, discuss on review of literature on behalf of previous research, thirdly, discuss ML with its techniques and examples, fourthly, discuss on stages of data science, fifthly, discuss on weakness or research gaps of previous research works according to literature review and finally discuss on proposed work for next research directions using observations to research gaps.

Keywords: AI, ML, data science, health data science, supervised learning, unsupervised learning, reinforcement learning, deep learning, deep reinforcement learning, ANN

INTRODUCTION

Data science is a progressive technology in the present time and is handled by other advanced technology known as machine learning (ML). ML is a subpart of artificial intelligence (AI). Due to ML handling different types of data using different techniques and algorithms; this technology is used in data science technology. ML is an advanced group of awareness and recognition as a technology that can evaluate huge amounts of data and computerize the responsibilities of data scientists [1]. Via connecting involuntary compilations of common techniques that have changed predictable arithmetical advances, ML has changed the method of data extraction and analysis [2]. In designing efficient and fast algorithms, as well as data-driven models for real-time data processing, ML can provide accurate outcomes with analysis [3]. ML is an important part of data science for handling large amounts of structured and unstructured data [4, 5]. ML is divided into five categories according to the types of data handled by algorithms. Data science decides the algorithms of ML for working with data. Figure 1 shows the stages of machine learning with data science.

***Author for Correspondence**
Vinay Bhatt
E-mail: vinay10191@gmail.com

[1]Research Scholar, Department of Computer Science and Engineering Asian International University, Imphal West, Manipur, India
[2]Associate Professor, Department of Computer Science and Engineering Asian International University, Imphal West, Manipur, India

Received Date: December 29, 2023
Accepted Date: January 04, 2024
Published Date: January 17, 2024

Citation: Vinay Bhatt, Mayank Kumar Applications of Machine Learning Algorithms in Health Data Science (HDS) for Next Research Directions: a Survey Report. Journal of Artificial Intelligence Research & Advances. 2024; 11(1): 16–21p.

LITERATURE REVIEW

In this section, discuss the review of literature for the next research directions on behalf of previous research (Table 1).

Table 1. Review of literature.

References	Research Category	Research Contributions
Rao DMS, and Sridhathri DS, (2023) [6]	Diabetes prediction using ML algorithms	Proposed the research on diabetic prediction using five classification machine learning algorithms as XGBoost, decision tree, K-nearest neighbor (KNN), Naïve Bayes, and random forest under precision. The outcome of this work is XGBoost algorithm is superior then other algorithms under precision [6].
Ramachandra AC, and Murthy D (2023) [7]	The approach of ML for diabetic prediction	Proposed the diabetic prediction model based on a logistic regression-based ML approach. The outcome is higher accuracy of data prediction using feature selection and regression techniques [7].
Wee BF et al. (2023) [8]	Deep learning and ML methods for diabetic detection	Proposed the review on different algorithms of ML and deep learning algorithms for diabetic detection. The outcome of this review is both algorithms are good in different fields according to purpose [8].
Madhu B et al. (2023) [9]	Diabetes risk prediction using ML algorithms	Proposed different machine learning models such as KNN, logistic regression, decision tree, random forest, AdaBoost classifier, XGBoost, and Naïve Bayes classifier for prediction accuracy under diabetic prediction from PIMA Indian diabetes datasets. The outcome of this research is XGBoost algorithm is more accurate than other algorithms [9].
Sharma A, and Mishra PK (2022) [10]	Performance of ML for breast cancer diagnosis using optimized feature selection algorithm	Proposed the work on healthcare systems for the observation to predict diseases like breast cancer using ML algorithms like DT, LR, KNN, ANN, and RF under an optimized feature selection algorithm [10].
Zhang Q et al. (2022) [11]	Data science approaches for COVID-19	Proposed the survey of data science for the research on COVID-19 using related parameters, mental health observation, diagnosis and risk measurement, digital contact tracing, communal media analytics with resource distribution, and drug improvement [11].
Kumar S, et al. (2022) [12]	Big data analytics with ML on sustainable finance	Proposed the study of big data analytics with ML under sustainable finance research using related parameters like climate financing, social responsive financing, green financing, and carbon financing with impact investing to manage the profit and return with unifying policies [12].
Zeng Z et al. (2022) [13]	ML methods for transcriptomics data analysis	Propose the implementation of ML and statistical methods like ANN, GCN, HMRF, and SVCA for analysis of transcriptomics data with different data sets and summarizations [13].
Khan K et al. (2022) [14]	ML application for concrete research	Propose the review on concrete research with different categories like conventional concrete, fiber reinforced, geo-polymer, and recycled aggregate using different ML methods like supervised (task-driven), unsupervised (data-driven), and reinforcement (learn from error) learning [14].
Martins RM and Gresse Von Wangenheim C (2022) [1]	Machine learning application in the teaching field	Proposed the survey on ML for teaching and used the case study in high school regarding the strategy and technology of content with learning the concept of ML with algorithm and tasks for handling the project-based problems [1].
He Y et al. (2022) [2]	ML in geochemistry and cosmochemistry	Proposed the implementation of ML methods for discovering the hidden big data related to Cosmo chemistry and geochemistry using different processes like water and soil quality prediction, sediment identification, and digital mapping [2].
Gandomi AH et al. (2022) [3]	ML in big data analytics	Proposed the review on ML for big data analytics including the process of handling the data like examining, analyzing, and varying data [3].
Liu T et al. (2022) [4]	Deep learning for medical image analysis under COVID-19	Proposed the implementation of deep learning methods like CNN algorithm for medical image analysis in the form of CT scans of COVID patients under COVID-19 [4].

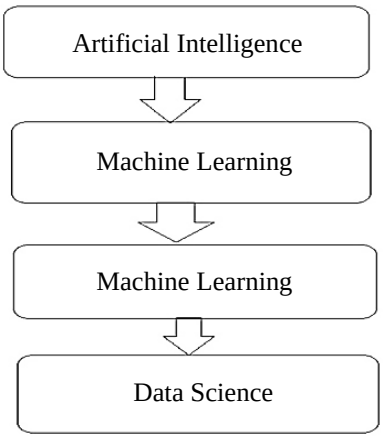

Figure 1. Stage of ML and data science under AI.

Table 2. Stages of data science.

S.N.	Data Science Stages	Description
1	Data ingestion	In this stage, collect the data in the form of structured (customer data) and unstructured data (audio, video, log files) from different data sources like social media, websites, etc.
2	Data storage and processing	In this stage, clean the data, transform, duplicate, and combine the data for storage in the data warehouse using the ETL process like extract, transform, and loads.
3	Data analysis	In this stage, analysis of the data in the form of patterns, values of distributions, and range using predictive modeling like ML or deep learning for accuracy.
4	Communicate	In this stage, present the data in the form of visualization-like reports.

DATA SCIENCE AND HEALTH DATA SCIENCE

Data science is a modern technology that combines different subjects like math, statistics, specialized programming, data analysis, AI, and ML [15]. When every process of data science is used in the healthcare system to handle the large amount of healthcare data the system is known as health data science (HDS) [16]. HDS is a branch of data science which implemented in the healthcare system to handle the bulk amount of patient data in the healthcare system. Data science technology is based on the principle of ML and an algorithm that is used for finding hidden patterns from raw data [17, 18]. Data science has four following stages (Table 2).

MACHINE LEARNING ALGORITHMS

Machine learning is an advanced field of computer science that is part of AI [19, 20]. ML is a branch of AI used to increase the growth of data science using algorithms and related data to improve human learning and data accuracy [21, 22]. There are five types of ML (Table 3).

RESEARCH GAPS OF PREVIOUS RESEARCH ON BEHALF OF THE LITERATURE REVIEW

When the review of literature is completed in this paper, discuss the limitations or research gaps for work in the next research directions on behalf of the literature review (Table 4).

PROPOSED WORK FOR NEXT RESEARCH ON BEHALF OF RESEARCH GAPS

In this section, discuss the further works for the next research directions on behalf of research gaps produced by the literature review (Table 5).

Table 3. ML types with examples.

S.N.	ML types	Description	Example
1	Supervised Learning	This technique involved training machines with a lot of training data for specific tasks.	Decision tree, logistic regression, support vector machine, K-nearest neighbor (KNN)
2	Unsupervised Learning	This technique is opposite to supervised learning means without any training machine and any training data. This technique is used for anomaly finding with clustering data.	K-means clustering,
3	Reinforcement Learning	This technique is used in research and development. In this technique, the output depends on the present input state and the next input depends on the output of the prior input.	Q-Learning, Markov decision process
4	Deep Learning	This technique is used to build a neural network whose function and structure are based on the human brain.	ANN, CNN
5	Deep Reinforcement Learning	This technique combines deep learning and reinforcement learning which is used for building robotics, smart healthcare systems, and games.	Deep Q-learning

Table 4. Research gaps for next research.

S.N.	Research gaps
1	The performance of decision tree, KNN, Naïve Bayes, and random forest is slower than XGBoost under precision for diabetic prediction [1].
2	The problem is regression and feature selection methods apply in only one approach of ML as logistic regression for diabetic prediction [2].
3	The problem is that the algorithm is based on the dataset [3].
4	The performance of comparative algorithms is not as good as XGBoost for the accuracy of diabetic prediction [4].
5	The outcome of predictive results is not 100% accurate in terms of accuracy under ML-based techniques [5].
6	The study of data science approaches mentioned in the review is not a proper fit for handling the COVID-19 pandemic [6].
7	The systematic study of big data analytics with ML under sustainable finance on a small scale includes a limited literature review [7].
8	Challenge the data analysis when increasing data complexity [8].
9	The performance of ML methods is good for small input factors according to research [9].
10	The mostly problem is how to teach ML to students [10].
11	Miss the up-to-date ML methods for handling the different types of data concerning cosmochemistry and geochemistry [11].
12	Different types of problems are faced in different types of big data analytics-based research like fraud detection, medical informatics, national intelligence, and marketing [12].
13	The main problem is handling the large imaging data sets [13].

Table 5. Proposed work for next research directions.

S.N.	Proposed work
1	Further work on improving the performance of classification-based ML algorithms such as Naïve Bayes, random forest, KNN, and decision tree for diabetic prediction.
2	Further work on comparative analysis between different algorithms of ML for diabetic prediction using feature selection and regression methods.
3	Further work is on collecting to clear dataset of diabetic patients from the healthcare system for analysis to accuracy using ML algorithms.
4	Further work on improving to performance of diabetic prediction accuracy using ML algorithms.
5	In the future, work on neuro-fuzzy with a combination of ML and deep learning for accuracy in resourceful diagnosis.

S.N.	Proposed work
6	Further work on new approaches of data science with combined ML to handle the COVID-19 pandemic infections.
7	Further study on sustainable finance using different factors like moderating, dependent, and independent variables, with different relationships like positive, negative, curvilinear, and linear by big data analytics and ML on a large scale.
8	Further work is on developing a new sequencing protocol for reducing data complexity in spatial transcriptomics data.
9	Next work on performance improvement of ML when input factors increase.
10	Further work on building a supportive programming environment on behalf of ML models.
11	Propose new approaches of ML which combine with deep learning for up-to-date ML.
12	Further work on big data analytics with advanced ML like representation learning distributed learning, active transfer learning, and parallel learning.
13	Next work on deep learning with data science for handling large amounts of data.

CONCLUSION

As we know, data science is a recent technology in computer science that is applicable in different fields. This technology is used for data accuracy in research fields using data analysis techniques by different types of related data like structured and unstructured data. When working on data science, use ML-based techniques and algorithms to handle the data. This review paper discusses the brief introduction of ML with its techniques, stages of data science, review of literature, research gaps produced by the literature survey, and proposed work for the next research directions according to research gaps by review of the literature.

Future Scope

The future scope of this paper is focused on advanced algorithms of ML; advanced tools and technology of data science for handling the data and producing good accuracy.

Further work is focused on research gaps of this review paper, selecting any datasets like diabetes datasets, COVID-19 datasets, image datasets, etc. in the healthcare system which is mentioned in the proposed work section for dataset collection and presents the new approach-based work on HDS for handling the data using ML algorithms.

REFERENCES

1. Martins RM, Gresse Von Wangenheim C. Findings on Teaching Machine Learning in High School: A Ten–Year Systematic Literature Review. Inform Educ. DOI: 10.15388/infedu.2023.18.
2. He Y, Zhou Y, Wen T, Zhang S, Huang F, Zou X, Ma X, Zhu Y. A review of machine learning in geochemistry and cosmochemistry: Method improvements and applications. Appl Geochem. 2022;140:105273. DOI: 10.1016/j.apgeochem.2022.105273.
3. Gandomi AH, Chen F, Abualigah L. Machine learning technologies for big data analytics. Electronics. 2022;11:421. DOI: 10.3390/electronics11030421.
4. Liu T, Siegel E, Shen D. Deep learning and medical image analysis for COVID-19 diagnosis and prediction. Annu Rev Biomed Eng. 2022;24:179–201. DOI: 10.1146/annurev-bioeng-110220-012203, PubMed: 35316609.
5. Kalaivaani PT, Krishnamoorthi R. Design and implementation of low power bio signal sensors for wireless body sensing network applications. Microprocess Microsyst. 2020;79:103271. DOI: 10.1016/j.micpro.2020.103271.
6. Rao DMS, Sridhathri DS. Diabetes mellitus prediction using ensemble machine learning techniques. ITM Web of Conferences. EDP Sciences. 2023;56. DOI: 10.1051/itmconf/20235605015.
7. Viswanatha V, Ramachandra AC, Dhanush Murthy, Thanishka. Diabetes Prediction Using Machine Learning Approach. Strad Res. 2023;10(8):75-82. DOI: 10.37896/sr10.8/008.

8. Wee BF, Sivakumar S, Lim KH, Wong WK, Juwono FH. Diabetes detection based on machine learning and deep learning approaches. Multimed Tools Appl. 2023;83:24153–24185. DOI: 10.1007/s11042-023-16407-5.
9. Madhu B, Aerranagula V, Mahomad R, Ravindernaik V, Madhavi K, Krishna G. Techniques of machine learning for the purpose of predicting diabetes risk in PIMA Indians. E3S Web of Conferences. EDP Sciences. 2023;430. DOI: 10.1051/e3sconf/202343001151.
10. Sharma A, Mishra PK. Performance analysis of machine learning based optimized feature selection approaches for breast cancer diagnosis. Int J Inf Technol. 2022;14:1949–1960. DOI: 10.1007/s41870-021-00671-5.
11. Zhang Q, Gao J, Wu JT, Cao Z, Dajun Zeng D. Data science approaches to confronting the COVID-19 pandemic: A narrative review. Philos Trans A Math Phys Eng Sci. 2022;380:20210127. DOI: 10.1098/rsta.2021.0127.
12. Kumar S, Sharma D, Rao S, Lim WM, Mangla SK. Past, present, and future of sustainable finance: Insights from big data analytics through machine learning of scholarly research. Ann Oper Res. 2022:1–44. DOI: 10.1007/s10479-021-04410-8.
13. Zeng Z, Li Y, Li Y, Luo Y. Statistical and machine learning methods for spatially resolved transcriptomics data analysis. Genome Biol. 2022;23:83. DOI: 10.1186/s13059-022-02653-7.
14. Khan K, Ahmad W, Amin MN, Ahmad A. A systematic review of the research development on the application of machine learning for concrete. Materials. 2022;15:4512. DOI: 10.3390/ma15134512.
15. Manley K, Nyelele C, Egoh BN. A review of machine learning and big data applications in addressing ecosystem service research gaps. Ecosyst Serv. 2022;57:101478. DOI: 10.1016/j.ecoser.2022.101478.
16. Sarker IH. Machine learning: Algorithms, real-world applications and research directions. SN Comput Sci. 2021;2:160. DOI: 10.1007/s42979-021-00592-x, PubMed: 33778771.
17. Shahraki A, Abbasi M, Taherkordi A, Jurcut AD. A comparative study on online machine learning techniques for network traffic streams analysis. Comput Networks. 2022;207:108836. DOI: 10.1016/j.comnet.2022.108836.
18. Mahesh B. Machine learning algorithms-a review. Int J Sci Res. 2020;9:381–386.
19. Jin W. Research on machine learning and its algorithms and development. J Phys Conf Ser. IOP Publishing. 2020;1544(1):012003. DOI: 10.1088/1742-6596/1544/1/012003.
20. Pandey D, Niwaria K, Chourasia B. Machine Learning Algorithms: A Review. Feb 2019;6(2): 916–922.
21. Ray S. A Quick Review of Machine Learning Algorithms. 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), Faridabad, India. 2019. pp. 35-39. DOI: 10.1109/COMITCon.2019.8862451.
22. Divya KS, Bhargavi P, Jyothi S. Machine learning algorithms in big data analytics. Int J Comput Sci Eng. 2018;6:63–70. DOI: 10.26438/ijcse/v6i1.6370.

