International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

A Comparative Analysis of Machine Learning Algorithms in Traditional and Cloud Computing Environments Using Medical Data

Memonah Abdullah Almatrouk[1], Mohamed Abdullah Al Hagery[2], Abdulatif Abdurhman Al Abdulatif[3]

Department of Computer Science, College of Computer, Qassim University, KSA
Email: *411207241[at]qu.edu.sa*

Department of Computer Science,College of Computer, Qassim University, KSA
Email: *hajry[at]qu.edu.sa*

Department of Computer Science,College of Computer, Qassim University, KSA
Email: *ab.alabdulatif[at]qu.edu.sa*

Abstract: ***In recent years, machine learning algorithms have become increasingly popular in healthcare for disease diagnosis and prediction. In this research, we compare the performance of various machine learning algorithms in traditional and cloud computing environments using a healthcare dataset of diabetes. We evaluate the algorithms based on accuracy, precision, recall, F1 score, and execution time. Many experiments were conducted using the Microsoft Azure platform in traditional and cloud computing environments. The results show that most algorithms perform better in the cloud environment according to the execution time values. The findings of this study can help healthcare professionals to choose the appropriate machine learning algorithm and environment for their applications.***

Keywords: Machine learning, Healthcare, Performance, Cloud Computing

1. Introduction

The overwhelming growth of healthcare data requires an effective and accurate analysis to gain the benefits of early detection of chronic diseases and improve patient care. In 2012, the McKinsey Global Institute estimated that efficient use of big healthcare data could generate $300 billion annually in the United States and £600 billion for personal location data analysis of users surplus [1]. Therefore, developing analytic platforms with massive processing and storage capabilities becomes essential to take advantage of massive amounts of healthcare data.

Analyzing various aspects of the collected big healthcare data allows healthcare institutions to understand and optimize services and expense management. It can also improve healthcare companies' decision-making and data processing, resulting in higher-quality healthcare and better patient services [2-3]. In March 2012, the United States launched an initiative to approve $200 million for big data research and development, aiming to use big data for scientific discoveries and biomedical research that would allow practitioners to make timely decisions [4]. Big data analysis can also help to identify patients with chronic disease, as well as epidemic outbreaks. Furthermore, effective real-time analysis of patients' vital data provides essential information about their health conditions, such as their blood pressure, heart rate, and blood sugar levels. This information helps healthcare authorities create preventive healthcare strategies [5].

Developing a unified data analysis framework for big healthcare data can positively affect all participants in the healthcare sector, including patients, practitioners, and management of healthcare institutions. Furthermore, it improves the quality of the care provided to the patients and multiple options for choosing the appropriate care. A unified healthcare framework can also influence the healthcare sector through innovations made in the biomedical field and granting complete and mobile access to healthcare data records everywhere and anytime for practitioners and patients.

Healthcare institutions and practitioners try to create innovative solutions. However, traditional healthcare systems cannot offer real-time, on-demand access to healthcare data and services, and sufficient resources are not available to process a massive amount of healthcare data [6]. Additionally, traditional storage systems can not comprehend and store large files as once time from external sources due to their insufficiency[7]. Therefore, healthcare companies want to replace traditional systems that cannot handle increasing healthcare big data [8].

Big data and continual data increase are essential to the healthcare sector. However, the most suitable Machine Learning algorithms for analyzing this data in the cloud computing environment have not been identified, as each data type has different characteristics. Therefore, appropriate solutions that keep pace with development and provide accessible and valuable services to support decision-makers in selecting accurate and fast tools are needed to analyze big health data, research diseases, and predict them. Consequently, this study aims to experiment with cloud computing technology for big data storage to enhance Machine Learning technologies' effectiveness, performance, and accuracy. Identifying appropriate algorithms will help create a unified healthcare system based on cloud computing with Machine Learning technologies.

This study evaluates Machine Learning techniques for cloud-based healthcare applications by analyzing their performance (execution time) and accuracy. This will provide insights into the pros and cons of adopting cloud computing in healthcare data analysis.

This paper is organized as follows. Section II presents the literature review. Section III illustrates the methodology, and section IV discusses the results. Finally, section V highlights the conclusions.

2. Literature Review

The literature review presents some research efforts made by some researchers that deserve mention. Previous studies investigated cloud computing in many sectors and have proven its effectiveness. In the healthcare sector, high-quality Machine Learning techniques that have proven effective in extracting valuable information from data are needed to mine and analyze data. Therefore, many Machine Learning methods have been applied in the healthcare sector.

		metrics	KNN	RF	SVM	DT
Data size= 768 rows	Traditional	precision	0.70	0.77	0.76	0.68
		recall	0.71	0.78	0.74	0.68
		f1-score	0.70	0.77	0.71	0.68
		Accuracy	60.00%	66.00%	76.00%	64.00%
	Cloud computing	precision	0.70	0.77	0.76	0.68
		recall	0.71	0.78	0.74	0.68
		f1-score	0.70	0.77	0.71	0.68
		Accuracy	60.00%	66.00%	76.00%	64.00%
Data size= 100,000 rows	Traditional	precision	0.99	1	0.99	1
		recall	0.99	1	0.99	1
		f1-score	0.99	1	0.99	1
		Accuracy	99.99%	100%	99.97%	100%
	Cloud computing	precision	0.99	1	0.99	1
		recall	0.99	1	0.99	1
		f1-score	0.99	1	0.99	1
		Accuracy	99.99%	100%	99.97%	100%
Data size= 200,000 rows	Traditional	precision	1	1	0.99	1
		recall	1	1	0.99	1
		f1-score	1	1	0.99	1
		Accuracy	100%	100%	99.97%	100%
	Cloud computing	precision	1	1	0.99	1
		recall	1	1	0.99	1
		f1-score	1	1	0.99	1
		Accuracy	100%	100%	99.97%	100%

Cloud computing is a promising technology that can store and analyze data while providing various on-demand services. Unlike the traditional method, cloud computing uses distributed-based databases to store data. To reduce costs on the infrastructure, cloud computing supplies computing, storage, and application, among other services for Internet users. Cloud computing is used in many services and activities, the most important of which are storage and computing[9]. The massive resources of the cloud are needed to evaluate Machine learning techniques that help to manage patients suffering from various diseases, especially diseases that the naked eye cannot detect. For example, heart disease is the most deadly disease in human health. It exceeds the permissible range in heart rate, pressure, and body. Therefore, cloud computing has been a general approach for meeting present requirements and achieving time-sensitive applications in healthcare. Cloud infrastructure's flexibility and cost features make it an attractive solution, as it can reliably handle data redundancy, fault recovery, and power management. Specifically, Electronic Health Records (EHRs) utilize cloud computing to manage the massive amount of data; this reduces the cost of storage, processing, and retrieval while ensuring data availability.

The health sector is one of the most critical sectors that gives a massive amount of data. It can be used in different algorithms of Machine Learning, which contributes to the health sector in data analysis and knowledge extraction, which is very helpful for supporting decision-makers. To profit from data, the analysts used specialized tools and software based on SQL queries to extract data and predictive values [10].

Many deaths worldwide are caused by increasing chronic diseases due to lifestyle and other factors. This increase in disease has resulted in more costs to the healthcare system. Machine Learning algorithms play a vital role in effectively analyzing the large amount of data generated by the healthcare system, and using these algorithms supports decision-makers in the healthcare sector because they can diagnose and predict disease based on datasets of patient histories. This leads to innovations in medical-related tools with greater accuracy, new treatment technologies, and the ability to reduce costs, identify deficiencies, and provide high-quality healthcare [11].

Regarding the healthcare sector, Machine Learning algorithms have demonstrated essential and influential roles in obtaining critical information and knowledge that detects

patients' cases and disease types by predicting, diagnosing, and classifying certain diseases [12].

3. Methodology

The methodology of this study involves the following steps:

1) Determining a set of Machine Learning techniques such as SVM, Naïve Bayes, and k-nearest neighbors algorithms.
2) Identify the healthcare datasets that contain the patient's health data and identify cloud computing platforms.
3) Apply the datasets to the Machine Learning algorithms using the two environments.
4) Implement well-known Machine Learning techniques in the healthcare domain in a cloud-computing platform., then measure the performance and accuracy separately and compare the results of the two environments.

4. Results and Discussion

We present the evaluation results of six algorithms and the performance comparison between the traditional environment and the cloud to prove which is better.

Table 1: The Diabetes disease dataset with double size

Comparing results between algorithms in both environments shows which is the best. As we can see in Table 1, which shows the results of traditional and cloud computing environments, have the same results as the evaluation of the four models in the diabetes dataset. The result of the diabetes dataset is shown in Figure 1. It has the highest values in the F-score, where the SVM=0.75 algorithm and then the algorithms of RF and DT work in a decreasing manner, and finally, KNN =0.60, while the recall measure was the highest result in the algorithm SVM = 0.76 then RF, DT, and KNN, respectively. At the precision measure, the highest value was in the SVM algorithm =0.75, DT and RF =0.67, and KNN=0.62. We conclude from this that the best performance of the previous algorithms for diabetes data is the SVM algorithm, while the KNN algorithm got the least accuracy in this dataset.

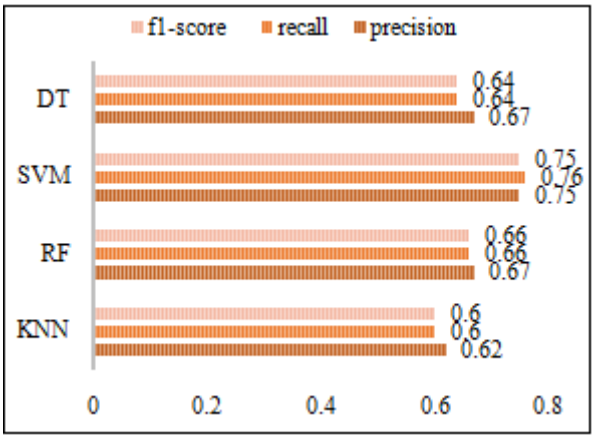

Figure 1: Algorithms results in the diabetes dataset

But what happens to these models in case we have a big dataset? Will the results change?

We did a duple data for the dataset as shown in tables 1 to 100,000 and 200,000. The results are as follows:

We noticed an increase in the accuracy of the models. The models can learn more about the underlying patterns and relationships in the data if the size increases. It gives more accurate predictions but takes into account the interest in the quality of the data and its freedom from distorted and lost data. Considering the results, the RF, DT, and KNN algorithms topped the other algorithms with an accuracy of up to 100% in the dataset.

Accordingly, in this case, we concluded from this assessment In this study, the good model that contributes to developing the healthcare sector and predicting disease patients is RF, DT algorithm, and KNN, which achieved higher performance and better accuracy when testing the models than the other algorithms. At the same time, the SVM maintained an accuracy of close to 100 in the training-testing set.

We used the Tic-Toc metric to calculate the execution time of the program in the two environments, as shown in Table 2.

Table 1: Results of performance speed in traditional and cloud environments for Diabetes dataset

Data size	Environment	BF
768	Traditional	23 sec
	Cloud computing	13 sec
100,000	Traditional	1min, 44 sec
	Cloud computing	0 min, 20 sec
200,000	Traditional	2 min, 37 sec
	Cloud computing	1 min, 46 sec

In the Diabetes dataset at a size of 768, the traditional environment gave a speed equal to 23 seconds, while in the cloud environment, the time was: 13 seconds.

The data size was doubled to 100,000 patient records in datasets to find a more significant difference. A change in performance speed was monitored as performance speed increased in the traditional environment in completing the program. The time was 1 minute and 44 seconds, while the cloud environment gives only 20 seconds. This motivates more experiments to prove the superiority of the cloud environment, as the data was doubled to 200 thousand in both environments. There was a significant change in the speed of performance, as the traditional environment took longer than the cloud environment, so the execution time of the traditional environment to finish the program was when it was data size was 200,000 records. The results equal 2 min, 37 sec, while in the cloud environment, the time was: 1 min, 46 seconds.

This proves that the cloud environment is better and faster than the traditional environment, which is the goal required to deal with big data.

5. Conclusion

Machine Learning algorithms have become increasingly popular for disease diagnosis and prediction in healthcare. In this research, we compared the performance of various Machine Learning algorithms in traditional and cloud computing environments using diabetes healthcare datasets. Our evaluation criteria included accuracy, precision, recall, F1-score, and execution time. We conducted experiments using the traditional and cloud computing environments using the Microsoft Azure platform. The results showed that some algorithms performed better in the cloud environment. We also analyzed the impact of the number and type of instances of the performance for the selected algorithms using the cloud environment. Our findings can help healthcare professionals choose the appropriate Machine Learning algorithm and application environment. This research highlights the importance of considering the trade-offs between performance, cost, and scalability when choosing a machine-learning environment for healthcare applications.

References

[1] I. Hargreaves *et al.*, *Text and Data Mining: Report from the Expert Group*. 2014.
[2] P. T. Chen, "Medical big data applications: Intertwined effects and effective resource allocation strategies identified through IRA-NRM analysis," *Technol. Forecast. Soc. Change*, vol. 130, no. June 2017, pp. 150–164, 2018, doi: 10.1016/j.techfore.2018.01.033.
[3] M. Jovanovic Milenkovic, A. Vukmirovic, and D. Milenkovic, "Big data analytics in the health sector: challenges and potentials," *Manag. Sustain. Bus. Manag. Solut. Emerg. Econ.*, vol. 24, no. 1, p. 23, 2019, doi: 10.7595/management.fon.2019.0001.
[4] Kyoungyoung, J.& Kim, G., (2013). Potentiality of Big Data in the Medical Sector: Focus on How to Reshape the Healthcare System, Health Informatics Research, 19(2), 79-85, DOI: 10.4258/hir.2013.19.2.79.
[5] S. Balaji and V. Prasathkumar, "Dynamic changes by big data in health care," *2020 Int. Conf. Comput. Commun. Informatics, ICCCI 2020*, pp. 22–25, 2020, doi: 10.1109/ICCCI48352.2020.9104168.
[6] S. Rallapalli, G. Rr, U. Pavan, and K. Ketavarapu, "Impact of Processing and Analyzing Healthcare Big Data on Cloud Computing Environment by Implementing Hadoop Cluster," *Procedia - Procedia Comput. Sci.*, vol. 85, pp. 16–22, 2016, doi: 10.1016/j.procs.2016.05.171.
[7] H. S. Ray, S. Mukherjee, and N. Mukherjee, "Performance Enhancement in Big Data handling," *2020 Int. Conf. Contemp. Comput. Appl. IC3A 2020*, pp. 17–22, 2020, doi: 10.1109/IC3A48958.2020.233261.
[8] P. T. Chen, C. L. Lin, and W. N. Wu, "Big data management in healthcare: Adoption challenges and implications," *Int. J. Inf. Manage.*, vol. 53, no. September 2018, p. 102078, 2020, doi: 10.1016/j.ijinfomgt.2020.102078.
[9] I. Odun-ayo, "An Overview of Data Storage in Cloud Computing," 2017, doi: 10.1109/ICNGCIS.2017.9.
[10] D. Gaurav, J. K. P. Singh Yadav, R. K. Kaliyar, and A. Goyal, "An Outline on Big Data and Big Data Analytics," *Proc. - IEEE 2018 Int. Conf. Adv. Comput. Commun. Control Networking, ICACCCN 2018*, pp. 74–79, 2018, doi: 10.1109/ICACCCN.2018.8748683.
[11] D. Jain, B. Kadecha, and S. Iyer, "A comparative study of machine learning techniques in healthcare," *Proc. 2019 6th Int. Conf. Comput. Sustain. Glob. Dev. INDIACom 2019*, pp. 455–460, 2019.
[12] S. Benbelkacem, O. Lio, O. Ahmed, and B. Bella, "Machine learning for Emergency Department Management," vol. 11, no. 3, pp. 19–36, 2019, doi: 10.4018/IJISSS.2019070102.