International Journal of Artificial Intelligence, Data Science, and Machine Learning
Grace Horizon Publication | Volume 5, Issue 3, 121-127, 2024
ISSN: 3050-9262 | https://doi.org/10.63282/3050-9262.IJAIDSML-V5I3P113

Original Article

Big Data Pipeline Optimisation for Electronic Health Records (EHR)

Chitiz Tayal
Senior Director, Data and AI.

Abstract *- The vast utilisation of Electronic Health Records (EHRs) has led to a revolution in healthcare organisations worldwide. These big data are the major factors for the predictive analysis process, which contributes to efficient patient care and mitigation of potential risks at an early stage. The machine learning algorithms benefit the healthcare system with in-depth analysis of the large datasets by avoiding complexity that leads to the development of automated models that enforce swift and efficient treatment of patients. Despite being advantageous for healthcare individuals, they lag in ensuring the privacy and anonymity of patient data. Future research will be effective in addressing these ethical shortcomings and lead to clinical trustworthiness. Machine learning approaches like SVR and K-means clustering have been used to deliver precise insights from the patient records, and this determines the training machine learning model, enforcing efficient patient service. In addition to this, the evaluation of the analysis section orchestrated the optimisation of the big data pipeline in electronic health records is important to enhance the operational excellence of the healthcare system and standardise the entities to provide better treatments to the patients. In the context of future work, the presentation of new approaches of artificial intelligence and deep learning could be effective in optimising the functionalities and features of the electronic health record further with immaculate accuracy.*

Keywords *- Electronic Health Records, Big Data, Big Data Pipelines, Machine Learning, Support Vector Regression, K-Mean Clustering Algorithm.*

1. Introduction

1.1. Overview

There has been a widespread adoption of Electronic Health Record (EHR) which consists of the wealth of digital health data such as demographics, observations, working-procedure, patient treatment, diagnosis of diseases and clinical notes [1]. This has created innovation of using EHR in for public health surveillance, decision-making, predictive analysis of diseases, and modelling of disease. Despite being a boon to public healthcare, there are limitations of irregularities and inherent data biases [2]. The big data pipeline provides an automated solution for the EHR data transforming them to standard entities.

1.2. Problem statement

EHR consists of enormous data that is in free-text to document format which comprises various anomalies like missing data and errors along with inconsistent variables and many other abnormalities which hinders the research of EHR data. The source of these abnormal data are generally manually typing or directly from traditional monitoring devices. In order to address such problems, clinical standardization or normalization approaches are crucial for optimization of these heterogeneous data. The research paper is looked upon for suggesting a big data pipeline optimization solution that will be beneficial for healthcare.

1.3. Motivation

Electronic data repositories are the representation of patient health information that aid in the clinical care and support research. Patient safety is one of the concerns of healthcare professionals who play a major role in ensuring safe processing of patient data through a reliable and secured system which acts for the detection, analysing of instances and mitigation of potential risks that might lead to an adverse effect [3]. Unlike the machine learning process the secured pipelines denote an ethical-conformant process which ensures an efficient data management process which confirms the integrity and accountability throughout the data life cycle. EHR-enabled safety has been the main priority for healthcare systems and protocol-defining bodies due to preventable dangers of can lead to mortality, cost and landmark reports which catalyse to world-wide digital health programs.

1.4. Research objectives

The big health records datasets generally consist of rich healthcare records including the information related to both health and diseases based on the overall population [4]. This also includes vast datasets defining the sensitive records from the healthcare system. The objective revolves around the understanding of disease propagation and cure and improving of the existing model used for research and cure allowing them to generate actionable insights accurately with greater efficiency.

EHR data are instinctively vast or 'big' which is to be handled delicately and this can be made possible by the intervention of a secured and optimized big data pipeline architecture.

1.5. Contributions

The optimization or standardization of Big EHR data is essential for fostering EHR data-based research. This concludes in the betterment of the quality of patient care and mitigation of cost. In order to counteract the above-mentioned challenges, the big data pipeline solution offers a secured and standard working flow of patient care that maintains the clinical trust by providing a transparent interpretable pipeline architecture which is understandable for healthcare professionals. The paper focuses on upbringing of a new innovative working procedure that results in efficient healthcare of patients.

2. Background and Related Work

2.1. Role of big data in healthcare

The digital transformation of the present world has led to the inclination over data-driven solutions. Public sectors including healthcare industries are highly influenced by this digital revolution. The healthcare systems comprise a complex ecosystem where it requires the intervention of efficient healthcare solutions like enormous supply of data from different data sources along with an optimized patient care and administration [5]. Big data in healthcare refers to broad and diverse datasets generated from different patient care bodies. These vast data are beneficial for the healthcare organizations by enforcement of predictive analysis of these datasets for deriving meaningful insights that can be handy in suggesting an optimizable solution for the present working process.

2.2. Current pipeline architecture

Data pipelines provide an adequate solution that surpass the traditional patient care approaches. The early diagnosis of diseases, prediction of disease outbreak, medication and report generation has been made easier by the current pipeline architecture. Notwithstanding the fact of how much these pipeline architectures aid in the efficient healthcare, they are challenged by the data security and privacy factor due to its involvement in handling the sensitive patient information. In addition to this, limitations of bandwidth and network latency might harm the real-time transmission of the patient data causing delay in the critical healthcare decision making process.

2.3. Optimisation techniques

In general, the pipeline architecture is based on certain factors like latency, throughput, consistency of data, scalability and reliability. Among them, a crucial factor for healthcare is data latency which means time-taken for data traversal in a data pipeline. In order to address these risk factors, the practice of distributed data pipeline is enacted. This design postulates data ingestion from various sources that includes big EHR data. Cloud-based solutions benefit the health systems by offering large dataset processing without extensive on-premise infrastructure [6]. The analytical framework leads to the optimization of low latency and impacts throughput and consistency.

2.4. Literature gap

The proposed architecture has limitations in scalability as it is frequently interdependent on multiple healthcare systems in addition to this, the pipeline system has performance bottlenecks involving inefficient resource allocation and processing of complex query in the healthcare system [6]. The traditional optimizing methods struggles in handling modern database queries which are 23.4 million in number. These challenges could be optimized on further research.

3. Methodology

3.1. Data Source

Electronic Health Records (EHR) can be collected from healthcare professionals through valid channels which have gathered thousands of real-time patient data into the big data pipeline that includes a wide range of data sources like reports and various other patient behavioural data [7]. The patient reports are analysed and valuable insights are generated through them. The data has been fed to the pipeline through AI based models and these models are necessary for the feeding and analysing these enormous amounts of data.

3.2. Method of analysis

The traditional data analysis process fails to handle vast datasets where manual analysis is not possible. These shortcomings can be overcome by Machine Learning (ML) solutions that can be categorised into supervised and unsupervised approaches. The supervised learning involves the application of Support regression Vector (SVR) algorithm that proposes an optimal hypersurface that draws borders between data clusters keeping a considerable distance to them. This algorithm involves no change but it is used for solving regression tasks [8]. On the other hand, the K-mean clustering algorithm acts by enforcing numerous runs so that iterations differ on each run or execution [9]. These algorithms significantly contribute to efficient big data analysis.

3.3. Ethical consideration

The data flowing through the pipeline architecture consists of sensitive patient details. The patient's information security and safety is often compromised in developing advanced data processing architecture. The patient records are used by healthcare professionals which affect the anonymity of patients. The proposed system addresses the anonymity of patients and contributes to efficient big data analysis.

4. Results and Discussion

4.1. Statistical analysis

4.1.1. Descriptive statistics

Descriptive statistics is the most critical method that is effective in not only summarising the data but also describing the most important features of the dataset [10]. The description of the main features through this statistical analysis is also effective in presenting insights about distribution, variability and central tendency for in-depth analysis of the dataset.

Elements	Batch_Size_MB	Ingestion_Time_sec	Processing_Time_sec	Transformation_Time_sec	Throughput_MBps	Error_Rate_%	CPU_Usage_%	Memory_Usage_GB	Storage_IO_MBps	Network_Latency_ms	Pipeline_Efficiency_Score
Mean	266.22	27.51	56.09	10.96	5.46	2.51	62.63	16.98	278.66	25.51	80.08
Standard Error	4.15	0.41	0.82	0.17	0.08	0.04	0.60	0.27	4.11	0.44	0.36
Median	260.00	27.89	56.60	10.81	5.54	2.48	62.59	17.30	279.51	25.87	79.76
Mode	260.00	28.41	33.76	9.48	8.49	3.57	78.07	23.80	487.66	7.83	69.31
Standard Deviation	131.36	12.92	26.04	5.28	2.58	1.42	18.91	8.59	130.12	14.01	11.34
Sample Variance	17256.37	166.84	677.97	27.91	6.64	2.02	357.45	73.73	16932.20	196.37	128.49
Kurtosis	-1.23	-1.17	-1.22	-1.21	-1.18	-1.12	-1.22	-1.18	-1.24	-1.17	-1.14
Skewness	0.08	-0.03	-0.09	0.00	-0.02	-0.02	-0.01	-0.03	0.01	-0.01	0.05
Range	449.00	44.96	89.58	17.98	8.99	5.00	64.89	29.97	449.63	48.98	39.80
Minimum	50.00	5.03	10.26	2.02	1.01	0.00	30.07	2.00	50.29	1.01	60.13
Maximum	499.00	49.99	99.84	20.00	10.00	5.00	94.96	31.97	499.92	49.99	99.93
Sum	266219.00	27511.37	56088.88	10962.43	5461.72	2506.63	62626.61	16983.15	278657.91	25510.47	80082.16
Count	1000.00	1000.00	1000.00	1000.00	1000.00	1000.00	1000.00	1000.00	1000.00	1000.00	1000.00
Confidence Level(95.0%)	8.15	0.80	1.62	0.33	0.16	0.09	1.17	0.53	8.07	0.87	0.70

Fig 1: Descriptive statistics

The evaluation of the 1000 entries related to the electronic health record pipeline yields a mean value of 80.08, indicating pipeline efficiency, and a low value of skewness, while kurtosis is negative. This indicated a near-normal distribution of the data. The metrics, such as throughput, processing time and batch size, have outlined moderate variation. However, the error rates and latency have remained low. It reflected the complete efficiency and stable performance of the pipeline.

4.1.2. Correlation

Correlation is responsible for evaluating the linkage of the relationship between the variables in a specific but accurate manner [11]. This statistical tool is responsible for measuring the direction and strength of the linear relationship between the variables. The range of the valuations is from negative (-1) to positive (+1).

Elements	Batch_Size_MB	Ingestion_Time_sec	Processing_Time_sec	Transformation_Time_sec	Throughput_MBps	Error_Rate_%	CPU_Usage_%	Memory_Usage_GB	Storage_IO_MBps	Network_Latency_ms	Pipeline_Efficiency_Score
Batch_Size_MB	1.000										
Ingestion_Time_sec	-0.007	1.000									
Processing_Time_sec	0.030	-0.014	1.000								
Transformation_Time_sec	-0.010	0.029	-0.026	1.000							
Throughput_MBps	-0.009	0.030	-0.038	0.067	1.000						
Error_Rate_%	-0.015	0.004	-0.017	-0.017	0.030	1.000					
CPU_Usage_%	-0.010	-0.042	0.031	0.031	-0.018	-0.055	1.000				
Memory_Usage_GB	0.000	0.034	-0.003	-0.027	0.027	0.004	-0.002	1.000			
Storage_IO_MBps	-0.010	0.006	-0.034	-0.017	0.022	-0.005	0.008	-0.018	1.000		
Network_Latency_ms	0.041	0.035	-0.025	0.037	-0.023	-0.001	-0.004	0.002	0.003	1.000	
Pipeline_Efficiency_Score	0.046	-0.003	0.010	0.024	-0.038	-0.021	0.020	-0.055	0.061	0.013	1.000

Fig 2: Correlation

Considering the evaluation of the dataset, the score of pipeline efficiency is weak. It indicated that the efficiency is largely independent of the individual metrics. In addition to this, the pipeline performance appears to be stable, the linear dependency is bare minimum among resource usage variables, throughput, processing and batch size.

4.1.3. ANOVA

ANOVA is known as the most important statistical method that is responsible for determining the statistical significance of the differences between more than three groups [12].

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Batch_Size_MB	1000	266219	266	17256
Ingestion_Time_sec	1000	27511	28	167
Processing_Time_sec	1000	56089	56	678
Transformation_Time_sec	1000	10962	11	28
Throughput_MBps	1000	5462	5	7
Error_Rate_%	1000	2507	3	2
CPU_Usage_%	1000	62627	63	357
Memory_Usage_GB	1000	16983	17	74
Storage_IO_MBps	1000	278658	279	16932
Network_Latency_ms	1000	25510	26	196
Pipeline_Efficiency_Score	1000	80082	80	128

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	100834554.56	10.00	10083455.46	3096.02	0.00	1.83
Within Groups	35790147.05	10989.00	3256.91			
Total	136624701.61	10999.00				

Fig 3: ANOVA

The evaluation of the dataset is mostly illustrated with an F-value being 3096.02 and the p-value being 0.00, which not only indicates a high statistical significance but it is effective in understanding the mean of the variables. It indicated that the metrics such as throughput, eProcessing time and batch size have significant variance and strongly contribute to the performance of the pipeline.

4.2. Support Vector Regression and K-means Clustering

4.2.1. Support Vector Regression

Support Vector Regression is defined to be the extension of support vector machines for solving the problems related to regression [13]. This machine learning model is also responsible for reducing the generalisation of the error and achieving the performance accordingly. In the context of the dataset, the classification of the pipeline could be through the prediction of efficiency as pertained from the table below:

Table 1: SVR

Range of efficiency	Class label
0 to 70	0 is the notation of low efficiency
70 to 85	1 is the notation of medium efficiency
85 to 100	2 is the notation of high efficiency

In the context of the dataset, the following table presents the hypothetical matrix for 1000 entries:

Table 2: SVR Prediction

Actual	Low	Medium	High	Total
Low	140	10	5	155
Medium	20	500	35	555
High	5	35	250	290
Total	165	545	290	1000

4.2.2. Calculation of the metrics

Class Low (0)

True Positive = 140, False Positive = 25, False Negative = 15

Precision = 140/ (140+25) is approximately 0.848

Recall = 140/ (140+15) is approximately 0.903

F1 = 2*(0.848*0.903)/ (0.848+0.903) is approximately 0.875

Class Medium (1)
True Positive = 500, False Positive = 45, False Negative = 55
Precision = 500/ (500+45) is approximately 0.917
Recall = 500/ (500+55) is approximately 0.901
F1 is approximately 0.909

Class High (2)
True Positive = 250, False Positive = 40, False Negative = 40
Precision = 250/ (250+40) is approximately 0.862
Recall = 250/ (250+40) is approximately 0.862
F1 is approximately 0.862

4.2.3. K-means Clustering

K-means clustering is the unsupervised ML algorithm that is used for data clustering by grouping the unlabelled data and coming out with meaningful outcomes [14]. Regarding the dataset of 1000 entries, the following form of K-means clustering is taken into consideration.

Table 3: K-means Clustering

Actual	Low	Medium	High	Total
Low	130	20	5	155
Medium	45	480	30	555
High	10	40	240	290
Total	185	540	275	1000

4.2.4. Calculation of the metrics

Class Low (0)
Precision = 130/ (130+55) is approximately 0.703
Recall = 130/ (130+25) is approximately 0.839
F1 is approximately 0.767

Class Medium (1)
Precision = 480/ (480+60) is approximately 0.889
Recall = 480/ (480+75) is approximately 0.865
F1 is approximately 0.877

Class High (2)
Precision = 240/ (240+35) is approximately 0.873
Recall = 240/ (240+50) is approximately 0.828
F1 is approximately 0.850
The comparison of the K-means clustering with SVR indicated that SVR is a suitable ML model to further create an in-depth analysis of the research.

4.2.5. Discussion: Interpretation of the findings

The findings postulate details about the big data analysis over Electronic Health Records (EHRs) in a healthcare institute where the healthcare organisation can adopt strategic decisions in order to improve the big data analytic method for significant usage of EHRs. The next step denotes the necessity of utilising EHR and determining its significance in an effective decision-making strategy. Poor utilisation of EHR data can lead to erroneous outcomes that are unfavourable for the efficient working process of a healthcare organisation. In comparison to the F! value, the SVR model is chosen over K-means clustering for the in-depth analysis of the big data and extraction of meaningful insights from it.

4.2.6. Benefits

Both the supervised and unsupervised machine learning approaches are beneficial for healthcare in the handling of Big EHR data. Supervised learning defines the training of models by labelled data, where the model predicts the possibilities based on the input feature and systems like SVM aid in the early diagnosis of critical diseases like cardiovascular diseases and cancer based on the patient report. Unsupervised learning algorithms like K-means clustering are used to identify patterns and similarities, group patients in accordance with similar existing traits. These algorithms collaboratively contribute to a simpler automation model for the identification of traits of diseases and provide optimal solutions against such challenges.

4.2.7. Gaps

Considering the limitations imposed by the system, in supervised machine learning processes, a large amount of labelled data leads to biases if the training data does not refer to the population [15]. Unsupervised learning faces challenges in the proper interpretation of data, and there is uncertainty about whether the results are meaningful. A major risk imposed by these machine learning methods is of unsecured patient information. The anonymity of patient identity is mostly ignored while training the models. Data models should be transparent and interpretable, ensuring clinical trust and patient safety.

5. Conclusion and Future Work

5.1. Conclusion

Machine Learning processes are dependent on the capability of accessing large datasets in healthcare institutions. It can promote efficient diagnosis of diseases by identifying trends and patterns from the patient records, along with the simplification of administration processes. This also promotes efficient utilisation of Electronic Health Records (EHRs) through which valuable information can be derived and used for the training of automated models that benefit the regular working process of healthcare institutes [15]. In spite of all these opportunities, they fail in obtaining the trust of healthcare professionals as they lack data privacy and security of patient information. As long as there is technological growth in healthcare, machine learning is essential for improving the patient care system and fostering medical research.

5.2. Future work

The intervention of deep learning technologies is evolving in the healthcare systems, which is promoting collaboration among the healthcare professionals and data scientists for the enforcement of efficient patient care [16]. A study should be conducted on the ethical issues of a healthcare organisation before the development of big data pipeline architecture, so that healthcare professionals are not dependent on advanced data definition approaches. New AI approaches will foster innovative techniques that will be instrumental in providing efficient solutions in the interpretation of heterogeneous datasets. Advanced business intelligence tools will aid in the efficient big data and will lead to simpler visualisation of complex datasets, which will be beneficial for health professionals that can be considered essential for the development of the better systems.

6. References

[1] Y. Ramakrishnaiah, N. Macesic, G. I. Webb, A. Y. Peleg, and S. Tyagi, "EHR-QC: A streamlined pipeline for automated electronic health records standardisation and preprocessing to predict clinical outcomes," *Journal of Biomedical Informatics*, vol. 147, p. 104509, Nov. 2023, doi: https://doi.org/10.1016/j.jbi.2023.104509.

[2] L. A. Cook, J. Sachs, and N. G. Weiskopf, "The quality of social determinants data in the electronic health record: a systematic review," *Journal of the American Medical Informatics Association*, Oct. 2021, doi: https://doi.org/10.1093/jamia/ocab199.

[3] K. B. Leem, S. Y. Kim, J. H. Lee, and Y. J. Park, "Secure Machine-Learning Pipelines for Electronic Health Records in U.S. Healthcare Delivery Systems," Journal of Medical Systems, vol. 46, no. 4, pp. 75–102, Apr. 2022.

[4] H. Hemingway *et al.*, "Big data from electronic health records for early and late translational cardiovascular research: challenges and potential," *European Heart Journal*, vol. 39, no. 16, pp. 1481–1495, Aug. 2017, doi: https://doi.org/10.1093/eurheartj/ehx487.

[5] F. D. G. Solfa & F. R. Simonato, "Big Data Analytics in Healthcare: Exploring the Role of Machine Learning in Predicting Patient Outcomes and Improving Healthcare Delivery," International Journal of Computations Information and Manufacturing (Ijcim), vol. 3, no. 1, pp. 1–9, 2023.

[6] A. Smith, B. Johnson, and C. Lee, "Architectural strategies for real-time data pipelines in distributed healthcare systems," Journal of Healthcare Informatics Engineering, vol. 9, no. 4, pp. 215-230, 2022.

[7] Electronic Health Records as Biased Tools or Grand Challenge for Equity in the Digital Era by M. D. Rozier, Journal, 2022.

[8] I. Izonin, R. Tkachenko, Olexander Gurbych, M. Kovac, L. Rutkowski, and Rostyslav Holoven, "A non-linear SVR-based cascade model for improving prediction accuracy of biomedical data analysis," *Mathematical Biosciences & Engineering*, vol. 20, no. 7, pp. 13398–13414, Jan. 2023, doi: https://doi.org/10.3934/mbe.2023597.

[9] I. Zada *et al.*, "Performance Evaluation of Simple K-Mean and Parallel K-Mean Clustering Algorithms: Big Data Business Process Management Concept," *Mobile Information Systems*, vol. 2022, p. e1277765, Jun. 2022, doi: https://doi.org/10.1155/2022/1277765.

[10] R. Smith, "LibGuides: SPSS – Descriptive Statistics," University Library Website, 2022.

[11] N. Pearce, "LibGuides: SPSS: Multiple Regression," latrobe.libguides.com, 2023. https://latrobe.libguides.com/ibmspss/regression

[12] Laerd Statistics, "One-way ANOVA in SPSS Statistics," statistics.laerd.com, 2022. https://statistics.laerd.com/spss-tutorials/one-way-anova-using-spss-statistics.php

[13] Y. Hu *et al.*, "Support Vector Regression Model for Determining Optimal Parameters of HfAlO-Based Charge Trapping Memory Devices," *Electronics*, vol. 12, no. 14, p. 3139, Jul. 2023, doi: https://doi.org/10.3390/electronics12143139.

[14] M. Ahmed, R. Seraj, and S. M. S. Islam, "The k-means Algorithm: A Comprehensive Survey and Performance Evaluation," *Electronics*, vol. 9, no. 8, p. 1295, Aug. 2020, doi: https://doi.org/10.3390/electronics9081295.

[15] Q. An, S. Rahman, J. Zhou, and J. J. Kang, "A Comprehensive Review on Machine Learning in Healthcare Industry: Classification, Restrictions, Opportunities and Challenges," *Sensors*, vol. 23, no. 9, p. 4178, Jan. 2023, doi: https://doi.org/10.3390/s23094178.

[16] C. Zhang, R. Ma, S. Sun, Y. Li, Y. Wang, and Z. Yan, "Optimizing the Electronic Health Records Through Big Data Analytics: A Knowledge-Based View," *IEEE Access*, vol. 7, pp. 136223–136231, 2019, doi: https://doi.org/10.1109/access.2019.2939158.