Journal of Information Systems Engineering and Management
2025, 10(15s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Analysis of Bigdata Representation and Storage Performance in a Smart Healthcare Environment

T. Kalai Selvi[1], S. Sasirekha[2]
[1]*Assistant Professor (SLG-I), Department of Computer Science and Engineering, Erode Sengunthar Engineering College, Perundurai, Erode*
[2]*Assosciate Professor, Department of Computer Science and Engineering, National Institute of Technical Teachers' Training and Research, Chennai*
*Corresponding Author: tkalaiselvi1281@esec.ac.in; sasirekha@nitttrc.edu.in

ARTICLE INFO

Received: 30 Nov 2024
Revised: 20 Jan 2025
Accepted: 02 Feb 2025

ABSTRACT

Internet of Things (IoT) is crucial for improving human health, safety, and treatment. Due to its traits, big data has grown more rapidly over time. Data centres and high-performance data processing dramatically advance with big data applications. Healthcare analytics are used to gather and analyze data from the healthcare sector to provide insight and inform decisions. An individual who uses data analytics to improve healthcare outcomes is known as a health care data analyst. To make it easier to access the data on the server, numerous storage techniques were developed. However, current storing methods did not reduce the complexity of the space or raise the accuracy level. Different data storage techniques in big data are discussed to address these issues.

Keywords: Internet of Things, big data, healthcare analytics, data storage, data analytics.

INTRODUCTION

Big data is essential one for learning, manipulation, as well as forecasting of information intelligence. Data are collected and stored in the storage systems. The application servers fetch the data from storage devices to minimize computational cost. Big data storage is storage communications used to store, handle and repossess large quantity of data. It allowed the storage as well as sorting of big data to get accessed, employed and processed. Storage infrastructure is linked to the estimating server nodes to enable fast processing as well as retrieval of big data. IoT is an interconnected device with an ability to monitor and transfer the data without any human intervention. Health is an essential require of living being. Healthcare scheme is an essential metric to reveal its developmental enlargement. Smart Healthcare Systems (SHS) are employed through people for personal healthcare annotations by different smart tools. It performs the data collection and transmission process with help of smart devices.

Healthcare personnel needed to be very mobile, and they also needed to communicate with patients and other people in the system. IoT is a type of ecosystem where each item includes other relevant devices in a setting to automate chores in the home and business. The Internet of Things (IoT) contains physical things with sensors that can connect to other devices and exchange data with them across Internet communication networks. The planned architecture is shown in Figure.1 handles data storage, data transfer, and data collection. The architecture included sensors (or IoT devices) that were installed on various cloud users, or patients.

Data collection is a technique for obtaining and analyzing data on specific variables in a working system so which pertinent questions asked as well as outcomes can assessed. Data collection server receives collected data. The gathered information is saved as an input matrix. For convenient data access in a cloud setting, the unified data is kept on the data management server. MHEALTH (Mobile HEALTH) dataset consists of recordings of 10 distinct volunteers' body motion as well as vital signs taken while were occupied in various physical behaviors.

Manuscript is structured as: Section 2 reviews drawbacks on conventional bigdata representation and storage techniques. Section 3 explains study as well as investigation of existing big data representation and storage methods. Section 4 performs probable comparison among them. Section 5 describes limitations of existing big data representation and storage methods. Section 6 provided the conclusion.

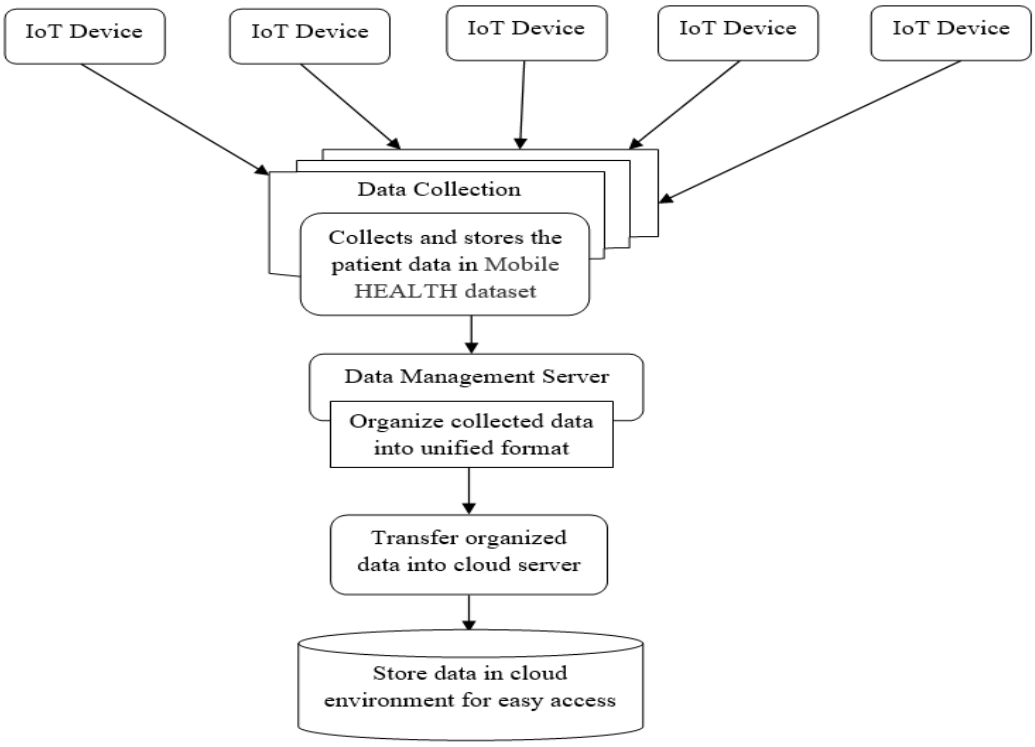

FIGURE 1. Architecture Diagram of Cloud Data Representation and Storage in Smart Environment

RELATED WORKS

A multi-model data representation framework was designed in [1] with category theory for transformation. However, accuracy level was not enhanced.

The hybrid linear feature extraction scheme was designed in [2] to mention supervised multi-class classification problems. Robust sparse linear discriminate analysis was designed to address unifying criterion to retain the classification merits. But, computational cost was not reduced.

A self-organizing fuzzy inference ensemble approach was introduced in [3].The designed framework was employed to vary decision boundaries depending on distance between prototypes for attaining higher classification results. However, failed to minimize time complexity.

Provenance-Based Label Propagation Algorithm was designed in [4] to normalize and cluster number of unbalanced provenance. The separate physical storage medium was used to save hot as well as cold information independently. Hot/cold scheduling system was employed to modernize as well as schedule information mechanically. However, the space complexity was not minimized.

For optimizing sub-dataset study over distributed storage schemes, Storage distribution aware technique was designed [5]. An efficient algorithm was employed to attain meta-data distributions. But, computational cost was not minimized by storage distribution aware method.

Rarity-aware data recovery method was designed in [6] to address data recovery issues. The main objective was to establish infrequency indicator to perform replica allocation as well as service necessity. However, space complexity was not minimized.

For assisting patients in evade the hospital visit during viral epidemics. E-health technology as well as remote patient monitoring was carried out in [7]. The designed technology allowed patients to identify and predict the sickness in remote disease discovery. But, the data confidentiality level was not improved by designed technology.

Integrated wearable as well as minimum-cost smart RMS scheme was designed in [8]. The designed system was employed for respiratory abnormality diagnosis through identifying human respiration parameters. But, computational complexity was not decreased.

Multi-Objective Automated Negotiation based Online Feature Selection was introduced in [9] to determine the advancement of online machine learning for enhancing the classification performance. However, it failed to enhance accuracy level.

Intelligent sampling technique was designed in [10]. The designed technique addressed the size dimension issues efficiently. However, time complexity was not minimized.
For efficient feature selection, Shark Smell Integrated Cat Swarm Algorithm (SSI-CSA) method was designed in [11]. The designed scheme enhanced the accuracy performance. Though accuracy level was improved by SSI-CSA model, error rate was not minimized.

BIGDATA STORAGE IN SMART HEALTHCARE ENVIRONMENT

Big data is a developing and well-liked one among various users for storing, analyzing, and managing the huge amount of data. Many machine learning (ML) methods are introduced to extract the information from Big Data. Hybrid storage addresses the real-time processing needs by big data. Internet of Things (IoT) is an interconnected device with an ability to monitor and transfer the data without any human intervention. Health is an essential require of living being. Healthcare scheme is essential metric to reveal its developmental expansion. India has huge population as well as ranked second place in world. SHS are employed through people for individual healthcare examination by different smart devices. It performed the data collection and transmission process with help of smart devices.

1. Unified representation and transformation of multi-model data

The multi-model data representation framework was carried out using category theory for transformation. Mapping among multi-model data as well as categorical representation was carried out for mutual transformation. The mapping was done using entrance way which tolerate the data about categorical depiction of any object. The transformation algorithm was used to convert input information to categorical demonstration. Existing works failed to improve the accuracy. In order to overcome these problems multi-model data representation framework is designed. The designed algorithm was generic to wrap every methods as well as their combinations. General approach was used for united modeling as well as multi-model data management. The category theory was right device for representing different data methods through defined as well as adequately general theory. The designed algorithm modifies produced types to append integrity limitations for guaranteeing respective recognizers as well as references. The designed approach was sufficient to wrap every known cases. Intra-model references were proliferated to particular DBMS. In inter-model references, propagation varied based on system combination. A polystore and multi-model DBMS was taken as separate system with single wrapper. In polystore system combination, DBMS not managed the references. The categorical framework maintained the information and the integrity constraints were checked externally.

2. A supervised discriminant data representation

The hybrid linear feature extraction system was used for addressing the supervised multi-class classification issues. RSLDA and ICS_DLSR were designed to address unifying criterion to retain the classification merits. Conventional works space complexity was not reduced. To overcome this problem, hybrid linear feature extraction scheme was introduced into transformation to choose features. The designed scheme accurately represented information to conserve row-sparsity consistency assets of samples from similar class. Linear transformation as well as orthogonal matrix were determined through iterative alternating reduction system depending on steepest descent gradient technique as well as initialization scheme. Designed approach was generic in sense which allowed permutation as well as tuning of linear discriminant embedding techniques.

3. Self-organizing fuzzy inference ensemble Scheme

A novel self-organizing fuzzy inference ensemble (SOFEnsemble) framework was introduced to reduce the deficiency. It was able of self-learning transparent prognostic method as of streaming data on chunk-by-chunk basis during human-interpretable procedure. Existing works, time complexity was not minimized; hence to overcome these issues SOFEnsemble framework is designed. Base learner was used to take self-adjust decision margins depending on inter-class as well as intra-class distance among prototypes recognized from consecutive data chunks for enhanced classification precision. An ensemble framework was used to attain higher classification precision and computational efficiency on large-scale issues. SOFEnsemble was introduced with streaming data on chunk-by-chunk basis. The designed framework recognized the representative prototypes from every data chunk and combined the newly identified prototypes from current data chunk with the previously identified ones from historical chunks. SOFEnsemble framework was used depend on inter-class and intra-class distances in accurate boundaries for decision-making. The training samples were arbitrarily dispersed to diverse

base learners to increase diversity and predictive accuracy. The designed framework minimized the computational encumber of every base learner. Respiratory sensor attached under nose to gather different human breathing parameters. The respiration signal was precisely identified below human breathing states as well as ambient factors upon different environmental situations.

4. Efficient Provenance Management via Clustering and Hybrid Storage

Provenance was sort of metadata which record formation as well as transformation of data objects. It was used in different areas like security, search, and experimental documentation. Provenance-Based Label Propagation method was introduced to normalize and cluster uneven provenance. Existing works, the accuracy was not improved and failed to reduce the time complexity. To overcome these problems, Provenance-Based Label Propagation method is designed. The divide physical storage mediums like SSD and HDD were used to save the hot as well as cold information individually. Hot/cold scheduling system was employed to modernize as well as schedule the data among them mechanically. Feedback method was used to situate as well as to compress the cold information consistent with query request. The designed system enhanced the provenance query performance with minimum runtime overhead. Feedback approach was employed to compress provenance data for long time to decrease storage overhead. Designed strategy was employed with different provenance workloads. The provenance clustering as well as hybrid storage enhanced query performance by minimum runtime overhead and space overhead.

PERFORMANCE ANALYSIS OF DATA REPRESENTATION AND STORAGE IN SMAT HEALTHCARE ENVIRONMENT

Experimental evaluation of existing data representation methods is implemented in Java platform on online available MHEALTH (Mobile HEALTH) dataset from Kaggle. During experimental study, the amount of healthcare data packets is considered as an input. MHEALTH dataset details are described in section 1. Sensors are located on subject's (i.e., patient) chest, right wrist and so on that are utilized in measuring the motion practiced through different body parts, such as, acceleration, rate of turn and so on. Result analysis performed through existing techniques with parameters are,

- Accuracy,
- Time Complexity and
- Space complexity

1. Accuracy

It is referred as ratio of number of healthcare data which are properly classified based on features to total number of data points. It is formulated as,

$$\text{Accuracy} = \frac{\text{Number of healthcare data points correctly classified based on features}}{\text{Total number of healthcare data points}} * 100 \quad (1)$$

It is measured in percentage (%).When accuracy is improved, method is more effective. Table 1 describes the accuracy comparison for four different existing methods.

Table 1. Tabulation of Accuracy

Number of healthcare data (Number)	Accuracy (%)			
	Multi-model data representation framework	Hybrid linear feature extraction scheme	Self-organizing fuzzy inference ensemble framework	Provenance-Based Label Propagation Algorithm
100	75	84	68	72
200	77	88	70	74
300	74	86	67	71
400	72	83	65	68
500	70	81	63	65
600	73	84	60	63
700	76	87	62	66
800	78	89	64	68

900	80	92	67	70
1000	83	94	70	72

Table 1. explains accuracy with number of healthcare data. Accuracy compared for conventional multi-model data representation framework, hybrid linear feature extraction scheme, self-organizing fuzzy inference ensemble framework and provenance-based label propagation algorithm. Let us assume that number of healthcare data as 700, accuracy of multi-model data representation framework, hybrid linear feature extraction scheme, self-organizing fuzzy inference ensemble framework and provenance-based label propagation algorithm is 76%, 87%, 62%, and 66%.

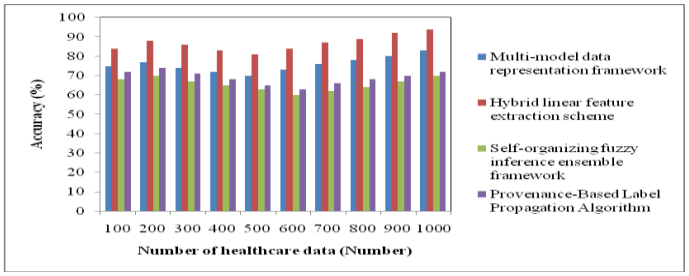

FIGURE 2. Measurement of Accuracy

Figure 2. explains accuracy for different healthcare data. The accuracy using hybrid linear feature extraction scheme is higher when compared to multi-model data representation framework, self-organizing fuzzy inference ensemble framework, hybrid linear feature extraction scheme, and provenance-based label propagation algorithm. This is due to the application of linear transformation and orthogonal matrix through iterative alternating minimization system with steepest descent gradient technique as well as initialization scheme. As a result, the accuracy level gets increased. Consequently, accuracy level of hybrid linear feature extraction scheme is increased by 15%, 32% and 26% when compared to multi-model data representation framework, self-organizing fuzzy inference ensemble framework and provenance-based label propagation algorithm respectively.

2. Time Complexity

It is described as time consumed for healthcare data classification based on features. It is defined as variation of starting time and ending time of healthcare data. It is calculated as,

$$T_{Com} = N * \text{time consumed for healthcare data point classification} \quad (2)$$

From (2), time complexity is determined. It is computed in milliseconds (ms).'N' denotes number of healthcare data points. When time complexity is minimum, the method is more effective.

Table 2. Tabulation of Time Complexity

Number of healthcare data (Number)	Time Complexity (ms)			
	Multi-model data representation framework	Hybrid linear feature extraction scheme	Self-organizing fuzzy inference ensemble framework	Provenance-Based Label Propagation Algorithm
100	25	32	35	40
200	27	35	37	43
300	29	38	40	46
400	31	40	42	49
500	33	43	46	51
600	35	45	49	54
700	38	48	52	57
800	40	50	54	59
900	43	52	57	62
1000	47	55	60	65

Table 2. describes time complexity with respect to number of healthcare data. Let us assume that number of healthcare data as 300, time complexity of multi-model data representation framework, hybrid linear feature extraction scheme, self-organizing fuzzy inference ensemble framework and provenance-based label propagation algorithm is 29ms, 38ms, 40ms, and 46ms.

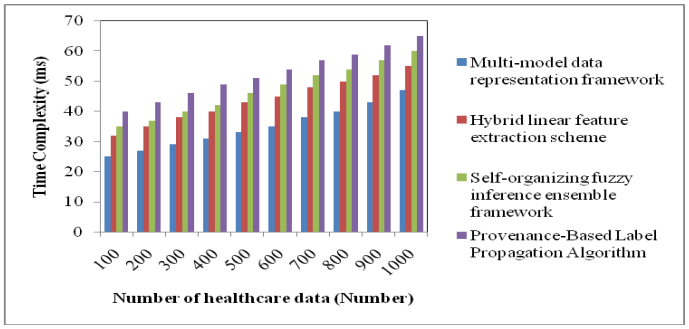

Figure 3. Measurement of Time Complexity

Figure 3. explains time complexity for different healthcare data. The time complexity using multi-model data representation framework is lesser when compared to self-organizing fuzzy inference ensemble framework, hybrid linear feature extraction scheme, and provenance-based label propagation algorithm. This is because of applying the transformation algorithm to convert input information to the categorical demonstration. Designed algorithm was standard to wrap every method and their amalgamations for unified modeling and multi-method data management. Therefore, the time complexity gets reduced. Consequently, time complexity of multi-model data representation framework is reduced by 21%, 27% and 34% when compared to hybrid linear feature extraction scheme, self-organizing fuzzy inference ensemble framework and provenance-based label propagation algorithm respectively.

3. Analysis on Space Complexity

It is defined as product of number of data packets and space consumed by one data packets for data storage in cloud. It is measured in megabytes (MB). It is computed as,

$$S_{Com} = N * \text{Space consumed by one healthcare data packets} \quad (3)$$

From (3), S_{com} represent the space complexity. When space complexity is minimum, technique is more effective.

Table 3. Tabulation of Space Complexity

Number of healthcare data (Number)	Space Complexity (ms)			
	Multi-model data representation framework	Hybrid linear feature extraction scheme	Self-organizing fuzzy inference ensemble framework	Provenance-Based Label Propagation Algorithm
100	45	34	21	36
200	47	37	23	39
300	49	40	26	42
400	52	42	29	44
500	54	45	32	48
600	58	48	35	50
700	60	52	38	54
800	62	55	40	57
900	65	58	43	60
1000	68	60	45	63

Table 3. explains the space complexity with number of healthcare data. Let us assume that number of healthcare data as 900, space complexity of multi-model data representation framework, hybrid linear feature extraction

scheme, self-organizing fuzzy inference ensemble framework and provenance-based label propagation algorithm is 65MB, 58MB, 43MB, and 60MB.

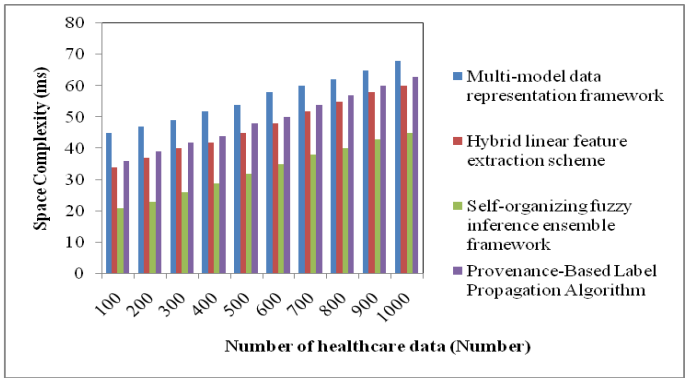

Figure 4. Measurement of Space Complexity

Figure 4. explains space complexity for different number of healthcare data. The space complexity using self-organizing fuzzy inference ensemble approach is minimum when compared to multi-model data representation framework, hybrid linear feature extraction scheme, and provenance-based label propagation algorithm. This is due to the application of self-powered as well as wearable RMS scheme to observe human breath. Top triboelectric film move up as well as down owing to breathing air flow that induce electrical respiration signals. It attained higher sensitivity under airflow speed range. Consequently, space complexity of self-organizing fuzzy inference ensemble framework is reduced by 42%, 30% and 33% when compared to multi-model data representation framework, hybrid linear feature extraction scheme, and provenance-based label propagation algorithm respectively.

Pseudocode for comparison of Accuracy, Time and Space Complexity

```
function compareAlgorithms(algorithms, dataset):
  results = []  // To store results for each algorithm
  for algorithm in algorithms:
    // Step 1: Measure Time Complexity
    startTime = getCurrentTime()  // Record the start time
    output = algorithm.run(dataset.input)  // Run the
algorithm
    endTime = getCurrentTime()  // Record the end time
    executionTime = endTime - startTime
    // Step 2: Measure Space Complexity
    memoryUsed = measureMemory(algorithm,
dataset.input)  // Memory usage during execution

    // Step 3: Measure Accuracy
    accuracy = calculateAccuracy(output,
dataset.expectedOutput)  // Compare output with ground
truth
    // Step 4: Record Results
    results.append({
      "algorithm": algorithm.name,
      "timeComplexity": executionTime,
      "spaceComplexity": memoryUsed,
      "accuracy": accuracy
    })
  // Step 5: Compare and Print Results
  printResults(results)
  return results

function measureMemory(algorithm, input):
  // Initialize memory tracking
  startMemory = getCurrentMemoryUsage()
  algorithm.run(input)  // Run the algorithm
  endMemory = getCurrentMemoryUsage()
  return endMemory - startMemory

function calculateAccuracy(output,
expectedOutput):
  // Implement a suitable metric based on the
task (e.g., precision, recall, F1-score)
  return compare(output, expectedOutput)

function printResults(results):
  print("Comparison of Algorithms:")
  for result in results:
    print("Algorithm:", result.algorithm)
    print("Time Complexity:",
result.timeComplexity, "ms")
    print("Space Complexity:",
result.spaceComplexity, "MB")
    print("Accuracy:", result.accuracy * 100, "%")
```

DISCUSSION AND LIMITATION ON DATA REPRESENTATION AND STORAGE IN SMAT HEALTHCARE ENVIRONMENT

Multi-model data representation framework used unified representation of multi-model information. The mapping among input information as well as categorical representation was carried out with entrance path which bears data about categorical demonstration of any object. The designed algorithm was generic to cover all their combinations. However, accuracy level was not enhanced.

Hybrid linear feature extraction system addressed the supervised multi-class classification issues. A unifying criterion was used with transformation to select the features accurately. Linear transformation as well as orthogonal matrix were determined depending on steepest descent gradient method. But computational cost was not minimized through hybrid linear feature extraction scheme.

A self-organizing fuzzy inference ensemble approach was used for self-learning the transparent predictive method as of streaming information on chunk-by-chunk manner by human-interpretable procedure. Base learner varied their decision boundaries depending on the distances among prototypes recognized from data chunks for enhanced classification accuracy. But time consumption was not reduced by designed framework.

Provenance-Based Label Propagation method was used to regularize as well as group number of uneven provenances. The divide physical storage mediums were used to save hot as well as cold information that update as well as schedule data automatically. The feedback mechanism located and compressed infrequently employed cold data consistent with query request. However, space complexity was not reduced by designed algorithm.

Future Work

Future work of study is to carry out effective data representation and storage methods through enhanced accuracy and minimum space complexity by machine learning (ML)methods.

CONCLUSION

Comparison of dissimilar data representation and storage techniques is studied. From the analysis, it is obvious which the computational cost was not minimized through hybrid linear feature extraction scheme. The space complexity and time complexity was not decreased by designed algorithm. In addition, accuracy level was not enhanced. Broad range of experiment on conventional techniques determines performance of data storage methods through its limitations. Lastly, research work performed by ML methods for improving data storage performance with higher accuracy and lesser space complexity.

REFERENCES

[1] Pavel Koupil and Irena Holubová, “A unified representation and transformation of multi-model data using category theory”, Journal of Big Data, Springer, Volume 9, Issue 61, 2022, Pages 1-49

[2] F. Dornaika, A. Khoder, A. Moujahid and W. Khoder, “A supervised discriminant data representation: application to pattern Classification”, Neural Computing and Applications, Springer, Volume 34, 2022, Pages 16879–16895

[3] Xiaowei Gu, Plamen Angelov and Zhijin Zhao, “Self-organizing fuzzy inference ensemble system for big streaming data classification”, Knowledge-Based Systems, Elsevier, Volume 218, 22 April 2021, Pages 1-15

[4] Die Hu, Dan Feng, Yulai Xie, Gongming Xu, Xinrui Gu and Darrell Long, “Efficient Provenance Management via Clustering and Hybrid Storage in Big Data Environments”, IEEE Transactions on Big Data, Volume 6, Issue 4, 01 December 2020, Pages 792 - 803

[5] Jun Wang, Xuhong Zhang, Jiangling Yin, Ruijun Wang, Huafeng Wu and Dezhi Han, “Speed up Big Data Analytics by Unveiling the Storage Distribution of Sub-Datasets”, IEEE Transactions on Big Data, Volume 4, Issue 2, 01 June 2018, Pages 231 – 244

[6] Songyun Wang, Jiabin Yuan, Xin Li, Zhuzhong Qian, Fabio Arena and Ilsun You, “Active Data Replica Recovery for Quality-Assurance Big Data Analysis in IC-IoT”, IEEE Access, Volume 7, July 2019, Pages 106997 - 107005

[7] Nagendra Singh, S.P. Sasirekha, Amol Dhakne, B.V. Sai Thrinath, D. Ramya and R.Thiagarajan, “IOT enabled hybrid model with learning ability for E-health care systems”, Measurement: Sensors, Elsevier, 2022, Pages 1-10

[8] Yingzhe Li, Chaoran Liu, Haiyang Zou, Lufeng Che, Peng Sun, Jiaming Yan, Wenzhu Liu, Zhenlong Xu, Weihuang Yang, Linxi Dong, Libo Zhao, Xucong Wang Gaofeng Wang and Zhong Lin Wang, “Integrated wearable smart sensor system for real-time multi-parameter respiration health monitoring”, Cell Reports Physical Science, Elsevier, Volume 4, Issue 1, January 2023, Pages 1-15

[9] Fatma BenSaid and Adel M. Alimi “Online feature selection system for big data classification based on multi-objective automated negotiation”, Pattern Recognition, Elsevier, Volume 110, February 2021, Pages 1-15

[10] Kheyreddine Djouzi, Kadda Beghdad-Bey and Abdenour Amamra, “A new adaptive sampling algorithm for big data classification”, Journal of Computational Science, Elsevier, Volume 61, May 2022, Pages 1-15

[11] J.C. Miraclin Joyce Pamila, R. Senthamil Selvi, P. Santhi and T.M. Nithya, “Ensemble classifier based big data classification with hybrid optimal feature selection”, Advances in Engineering Software, Elsevier, Volume 173, November 2022, Pages 1-15.