Computer Communications 160 (2020) 718–729

Contents lists available at ScienceDirect

Computer Communications

journal homepage: www.elsevier.com/locate/comcom

Information processing in Internet of Things using big data analytics

Chaomin Li

Moral Culture Research Center of Hunan Normal University, Changsha, 410081, China

ARTICLE INFO

Keywords:
Internet of Things (IoT)
Fog computing
Cloud computing
Naïve Bayes classifier
Smart and real-time healthcare information processing (SRHIP)

ABSTRACT

With innovation in persistent technologies, such as wearable sensor gadgets, sensor devices, and wireless ad-hoc communication networks connect everyday life things to the Internet, normally referred to as Internet of Things (IoT). IoT is observed as an active entity for design and development of smart and context awareness services and applications in the area of business, science and engineering discipline. These applications and services could vigorously respond to the surroundings transformation and users' preference. Developing a scalable system for data analysis, processing and mining of enormous real world based datasets has turned into one of the demanding problems that faces both system research scholars and data management research scholars. Employing big data analytics with IoT technologies is one of the ways for handling the timely analyzing information (i.e., data, events) streams. In this paper, we propose an integrated approach that coalesce IoT systems with big data tools into a holistic platform for real-time and continuous data monitoring and processing. We propose Fog assisted IoT based Smart and real time healthcare information processing (SRHIP) system in which large amounts of data generated by IoT sensor devices are offloaded at Fog cloud form data analytics and processing with minimum delay. The processed data is then transferred to a centralized cloud system for further analysis and storage. In this work, we introduce a Fog-assisted model with big data environment for data analytic of real time data with remote monitoring and discuss our plan for evaluating its efficacy in terms of several performance metrics such as transmission cost, storage cost, accuracy, specificity, sensitivity and F-measure. The proposed SRHIP system needs less transmission cost of 40.10% in comparison to SPPDA, 100% fewer bytes are compromised in comparison to GCEDA. Our proposed system data size reduction of 60% reduction due to proposed compression scheme in comparison to other benchmark strategies that offer 40% of reduction.

1. Introduction

In the modern world of big data, atmost all the companies worldwide use various techniques to process their data to get useful and practical information. These techniques include various algorithms, statistical models, prediction and classification models, decision making processes, and many more, which we simply term it as big data analytics. Social media plays a vital role in creating massive amounts of data. This is because most of the industries irrespective of the size of their organization (it may be a corporate company or a small level industry) uses social media to get in connection with their employees and customers, promote new products or services. In turn customers also rely on social media to know about new products and services [1]. Other devices such as GPS, sensors, smart phones, IOT devices, telemetry also produces massive amounts of data.

This massive increase in data initiated the development of big data analytics for real time problems. This on other hand throws a big challenges for machine learning approaches like highly distributed data, eliminating noisy data, limitation of labeled data, and much more. Data analytic techniques like indexing, storage, information processing and tools associated with these techniques also face lot of challenges due to increase in data [2]. Besides having high performing solutions, it is not possible to perform the techniques in the traditional batch processing devices. Large scale performing techniques trigger the experts to use distributed processing devices. The use of highly complex models to analyze big data, initiates the use of distributed computing devices [3]. In general, massive increase in data has initiated two important aspect of analysis. one is the increase of the technology and the tools associated with the analysis and the other one is associated with the corporation and the organization which possess big data called big data analytic capabilities. As the big data circumstances increases, techniques and tools associated with it also increases.

Yet, due to the large number of alternative techniques and tools it is often very difficult for the researchers and others to analyze and choose the better choice for every problem. They find it difficult in choosing the appropriate models for analysis. Thus, the number of researches on this area is still growing and is not likely to change in future. More research to solve the scalability and fault tolerance has been carried out to meet the challenges [4,5]. Organizations should always be updated

E-mail address: sd20200424@163.com.

https://doi.org/10.1016/j.comcom.2020.06.020
Received 24 April 2020; Received in revised form 10 June 2020; Accepted 20 June 2020
Available online 12 July 2020

in their digital technology aspects in order to compete with other organizations and to withstand the digital world. Organization's capacity is judged in the way they are aware of newly emerging digital technology and response to those technologies. Decision making methods is a key factor for information processing which needs thorough market study. The decisions have to be taken instantly based on the current situation and it keeps varying all the time. In this scenario, organizations should adopt the latest technology and methods for an effective processing of information. So, the organization should update their decision making technologies often to meet their challenges [6].

With advance in technology, it is now possible to communicate between a device and internet called Internet of Things which gets data from the devices and converts it to useful information. IOT is emerging as an essential part of big data and artificial intelligence. Most of the data incorporated in big data analysis and artificial intelligence systems are received from IOT devices. In turn the information processing and the decision making models of big data analytics makes the IOT system more efficient. Thus, the IOT systems, big data analytics, and machine learning algorithms work together for an efficient system. This efficient system helps the man kind in many ways and by reducing the labor cost. Processing and analyzing the data from a distributed environment is the most important factor of IoT which can be applied on business and research based applications. Huge data would be stored and processed which may be more challenging, so classification techniques can be used [7,8].

IoT in healthcare is constantly increasing. Remote monitoring of patients and elder citizens, home care are the recent trends of healthcare using IoT. Using RFID, tracking, monitoring are other fields in health care. These technologies can be adopted by medical people, common people and hospitals. However more people come up with a need for these applications. People are concerned about their health factors and so use IoT devices to track their health personally. Some of the people need IoT applications to monitor their elders or sick people at home. More efficient IoT applications are needed to solve this. The inventions in the health care system are just at the starting stages while comparing the needs of the medical field. The Consummation of the Health Service Research will be extensively achieved by overlapping the interdisciplinary personnel and services into a single paradigm: Comprising of Medical, Government, Insurances, Statistics, Social, and Research. The system should analytically balance between the cost, efficiency, quality, attainable and delivery [9,10].

The successful outcome of any Health service Research is to provide thorough and satisfactory services to individuals and the public. The system apart from medical services should comprise of insurance services, research activities, government support and much more interdisciplinary services to enhance the system. The health service should enclose right from booking an appointment of the doctors till getting an insurance policy if needed [11]. The advanced research may include possible ways for payment of hospital bills, including policy innovations, nursing home, home care for elderly people, involving robotics in hospitals. Health Service Research investigates how other domains such as social, financial, economy; technology affects the quality and cost of the healthcare system. The health care research would be beneficial for the individuals, society, organization, government and researchers [12].

Many more advanced research in this field needs to be invented where IoT will play a vital role. Smart pills, home care, personal health monitoring, robotics for care taking, real-time health monitoring system are the upcoming researches in health care using IoT. A new technology called the Internet of healthcare things (IoHT) is in the evolution stage which would get mostly implemented before 2020. The survey shows that from 2017 to 2022 there will be a lot of research taking place in the field of healthcare using IoT [13].

Devices of IoT are used, controlled and monitored by the users from remote places and can access data from it. The users communicate with the IoT devices through specially designed programs or applications. The end point at the user end and the devices allows them to communicate with each other. Service providers facilitate the communication between the devices and the users through various technologies [14]. Communication between many end users and many numbers of devices is also possible through LAN or WIFI technology. The nature of devices connected to the network may be the same or different types. The information shared by IoT devices may not always be small, thus requires network collaboration. The data need to be transferred to other networks for analysis, processing or storage purposes. Here cloud computing comes into the picture for storing data produced by IoT devices. The cloud technology solves the problem of scalability and vitality faced by IoT devices [15].

Cloud computing allows industries to manage the huge data efficiently and retrieve useful information from the collected data which are the basic challenges faced by big data technologies. The cloud provides huge space for storage and processing of data to ensure the quality of the applications. But cloud computing is affected by number of violations and leakages. Framework should be designed to preserve privacy of data. Encryption techniques can be more useful to reduce the theft of data by unauthorized users [16]. The distributed approach should provide separate partition for data and analysis to make the system more independent and efficient. This increase the accuracy of analysis which ensures a secured cloud platform for data. Security is the major challenge in cloud based environment and new technologies should evolve to ensure security and privacy of data in cloud environment [17].

Fog computing is a technology used in between the IoT devices and cloud storage. It is similar to edge computing on bringing the technologies involved in processing closer to data. Fog computing is used for security purposes. In the Fog environment, the processing technology is placed at LAN and the data is exchanged from the endpoints to the gateway of fog. In edge computing, the technology is placed at the end point or gateway. Edge computing allows only one device at a time to share the data. While fog computing is more scalable and allows multiple devices to pass data at a time. Since edge computing cannot be applicable for all the applications, fog computing is used. Fog computing reduces the amount of data to be sent to the cloud and reduces bandwidth. Using the Fog system in the real world will connect all the IoT devices used by the people for monitoring their personal health [18].

Spark streaming is used to get live and real time from devices and process into batches of data. Spark engine then processes every batch separately, then the result is again stored in any storage application like database system or cloud system. Spark streaming gets continuous data at a regular interval from real time devices automatically. Apache spark applications along with IoT devices are emerging as a new technology for processing the data. Apache spark could be applied for real time applications. The data are collected from IoT devices, edge computing or fog computing approaches are applied, and then the data is transferred to spark streaming. Data is processed and triggered to real time events [19].

Hadoop framework is modeled to measure up a unique machine to several thousands of machines in a group of cluster, where every node in a cluster allows to have local computation and storage. This Hadoop framework consists of below components:

- Hadoop common: this component consists of utilities by different Hadoop components.
- Hadoop Distributed File System (HDFS): this is scattered file model that is known as component which grants storage capability in Hadoop platform. It gives higher throughput access to various applications of data by breaking larger file into small blocks and preserving those blocks in various nodes across clusters.
- Hadoop MapReduce: this component is the application of MapReduce model where accessing of larger file takes place then forming smaller pieces.

2. Literature survey

The authors Nada Elgendy and Ahmed Elragal [20] in the information generation, more data is available in our hand to make decisions. The big data is denoted as data sets which are very big and also high in variety as well as velocity. Big data analytics is very difficult by using traditional methods and tools. Studies have to be done due to the fast growth of data and solutions to be provided so that to extract knowledge and value from datasets. It is very useful in data storage and management since the advanced method should be developed. It has the advantage of fraud detection.

In the past decade, remote sensing data has developed rapidly. Since the remote sensing application method developed moderately owing to the dispute on huge imagery data storage as well as processing. To access the current remote sensing images is very expensive. Then the cost of learning curve which is used for image processing tools as well as remote sensing images is high. Yan Huang et al. [21] proposed the solution to analyze huge imagery data using cloud computing. It is based on the ArcGIS method that is used for processing the remote sensing imagery as well as storage and management it leads to next-generation remote sensing application model. It has a drawback that it is not that easy to construct a remote sensing application model. To analyze the big imagery data for remote sensing is still very hard. The challenge in the remote sensing application model is together of computation and data-intensive.

The effective smart devices like mobile phones and sensors have led to the rapid growth of data streaming applications to a greater extent; they are interactive gaming, event monitoring and augmented reality. The huge streams of data created by these applications made an Internet of things which is a large origin of big data. Shusen Yang [22] the development in fog platform leads to attention in industry as well as academia because it provides less cost-efficient and latency. The design space of fog streaming is analyzed with the factors of four important dimensions such as data, human, system, and optimization. The information streams can be processed by adopting fog architecture so that it can provide rich computing resources and high-quality communications. Fog data streaming is developed in the future by anticipating the combinations of stream processing and network edge so that it leads to growth in the field of industry and academia.

The Internet of Things (IoT) generates a big amount of data that has been growing exponentially dependent on eternal operational states. Muhammad Rizwan Anawar et al. [23] proposed that the IoT devices generate massive information which has been a trouble for expected data analytics and processing that is controlled by cloud since the explosion development of IoT. However, the Fog structure challenges those troubles with the strong perfect functionality of cloud framework found on the formation of micro clouds at the closest edge of data sources. Since IoT based big data analytics is done by the fog computing method, it is still developing and it needs research to make smart decisions. It plays a vital role in many applications such as smart cities, smart grid and health care monitoring. It has a drawback of cloud computing works only at the edge of IoT. Senthil Murugan and Usha Devi [24–27] proposed hybrid model for analyzing large amount data using some optimization techniques.

A huge amount of data is generated by IoT. It is very challenging to analyze and process the real-time data and an effective solution is to provide. To address this issue in data analytics we need to know where the data is generated and stored. Farhad Mehdi pour et al. [28] proposed a solution known as Fog engine which is combined with IoT's close to the ground as well as promotes data analytics earlier migrating massive amounts of data to a central location. The data analytics can be accomplished very close to where the data is produced; it can reduce data processing time as well as data communication overhead. The Performance is analyzed by several parameters like network bandwidth, speed of processing data, and data transfer size. It provides multiple advantages such as higher throughput, less usage of network bandwidth and lower latency. It has the drawbacks of highly expensive and less security to data.

The smartness of the city is achieved by analyzing the Internet of Things sensor data. Jianhua He et al. [29] proposed a multitier fog computing model to analyze large data for smart city applications. The multi-tier fog composed of dedicated and ad-hoc fogs with dedicated and opportunistic computing resources. This proposed model with functional modules that is to check the issues of dedicated computing framework as well as the sluggish reaction in cloud computing. Hence fog-based data analytics can largely improve the performance of smart city data analytics services when compared to cloud computing-based data analytics in terms of service utility and blocking probability. It has a drawback of analyzing the data with and without fog computing architectures for the QoS management. The authors Taiwo Kolajo and Olawande Daramola et al. [30] the big stream of data is present everywhere due to the use of a large number of applications so that it can produce a massive amount of data. Due to the inherent dynamic feature of big data, it is very difficult to apply the data mining tools and methods on the big stream of data. So, it provides a methodical and rigorous approach in analyzing big data streams. It has a drawback of accuracy, privacy problem, heterogeneity, scalability and fault tolerance.

The authors Fezile Matsebula and Ernest Mnkandla [31] proposed big data analytics in higher education. The big data in higher education comes from various aspects: social media, blogs, learning management systems, student information systems, research-based information as well as data generated by machines. Then the data is analyzed and it provides good placement to students, early warning systems as well as accurate enrollment prediction which helps to assist and identify students from the risk of dropping out or failing. Big data plays a vital role in producing competitive benefits in higher education. The big data challenge is still not satisfied by traditional processing of data methods and warehousing of data as well as evaluation of structured and unstructured data using the RDBMS. It uses educational mining techniques to make the students understand in a better way. Then it uses a learning analytic technique that is used to improve learning and education. The architecture should be designed in an efficient way that the learning analytic as well as learning management system should be integrated and it should not be designed separately.

Spano et al. [32] developed a wireless wearable ECG monitoring system which particularly focus on minimal power and cost. Chang et al. [33] developed a healthcare system which monitors diabetic patients utilizing smartphone. Cheng et al. [34] worked on developing tracking patients affected with Alzheimer's disease by locating the movement and pattern. Milici et al. [35] developed a real time monitoring system to observe the quality of sleep by investigating the rate of respiration in patients. Pasluosta et al. [36] surveyed the various types of technologies utilized for monitoring the patient's affected with Parkinson's disease.

3. Proposed work

The proposed architecture for smart and real time healthcare information processing (SRHIP) system composed of three layer of processing: (1) IoT body sensor network layer that perform data accumulation and aggregation task, (2) FoG processing and computation layer perform the information processing, analyzing and classification task using naïve bayes (NB) classifier, (3) Cloud computation layer perform data analysis, storage, decision making classification. A precise overview of the proposed SRHIP system architecture is represented in Fig. 1.

Fog computing provides various services which are handed over to IoT model by various layers. In the domain of healthcare it differentiates it from various other applications in IoT by utilizing it most significant feature such as remote observing which need higher degree of dependability. Fog computing can offer closer preprocessing to the data source, which helps in ensuring cleaner data forwarding

Fig. 1. Proposed smart and real-time healthcare information processing.

to the cloud server for in-depth analysis of data which becomes more innovative, enhanced workflow and enhance patient's care.

We have proposed work with data accumulation, aggregation, compression and encryption at aggregator node of IoT body sensor network layer; information extraction, data normalization, rule engine, data filtration, data processing using spark and Hadoop ecosystem and data classification using naïve Bayes classifier is performed at FoG processing and computation layer, further data analytics, classification and storage is performed at cloud computation layer. Detailed view of proposed system flow is shown in Fig. 2.

Various sensor are utilized in gathering information of users. For instance, ECG sensor is utilized for monitoring the heartbeat of the individual, breathing sensor is utilized for knowing the function of breathing in user, glucose sensor helps in finding the blood glucose level of a particular person, sensor measuring the body temperature also plays vital role in observing the current condition of individual. These sensor are utilized for accessing the information related to heath of individual. At various scenario the body condition of human changes, this can be monitored using real time health monitoring system for preventing the abnormal changes caused in user's and can provide necessary treating in preventing any further damage to health of individual.

3.1. IoT Body sensor network layer

The IoT Body sensor network layer consists of a data accumulation model, data encryption and compression model and data aggregation model.

3.1.1. Data accumulation model

This model accumulates information about the physical condition of a person and a variety of actions from the adjoining surroundings about every user. The collected data consists of data relevant to health, environment and locality. This data-set information is gathered from the wireless IoT sensor devices implanted into the body of the user as well as the environmental surrounding and locality of the user. These IoT sensors are proficient in sensing information and transmitting sensing information constantly using wireless network technology.

3.1.2. Data encryption and compression model

In the protected data accumulation model, accumulation and deliverance of confidential data is a decisive task. Data protection needs to be considered as the prerequisite in healthcare IoT enable application for data sharing using wireless network. Consider a situation where data from source to destination is transmitted successfully without any security then any unauthorized attackers can admittance and modify the sensitive information. Our proposed SRHIP system solves the problem of confidentiality by hierarchical symmetric key data encryption and data compression scheme.

Each sensor Node Nd_i generates an encrypted and compressed message which is sent to the aggregator node for aggregation. The algorithm for message encryption and compression explained in Algorithm 1. All sensor devices start sensing and sharing data information with secret key encryption scheme where secret key SK_{ib} is shared among sensing device nodes and BS of Fog server. Importantly, only those sensing node devices that fulfill the query condition (like wearable sensing nodes with temperature values of body greater than 37 °C or

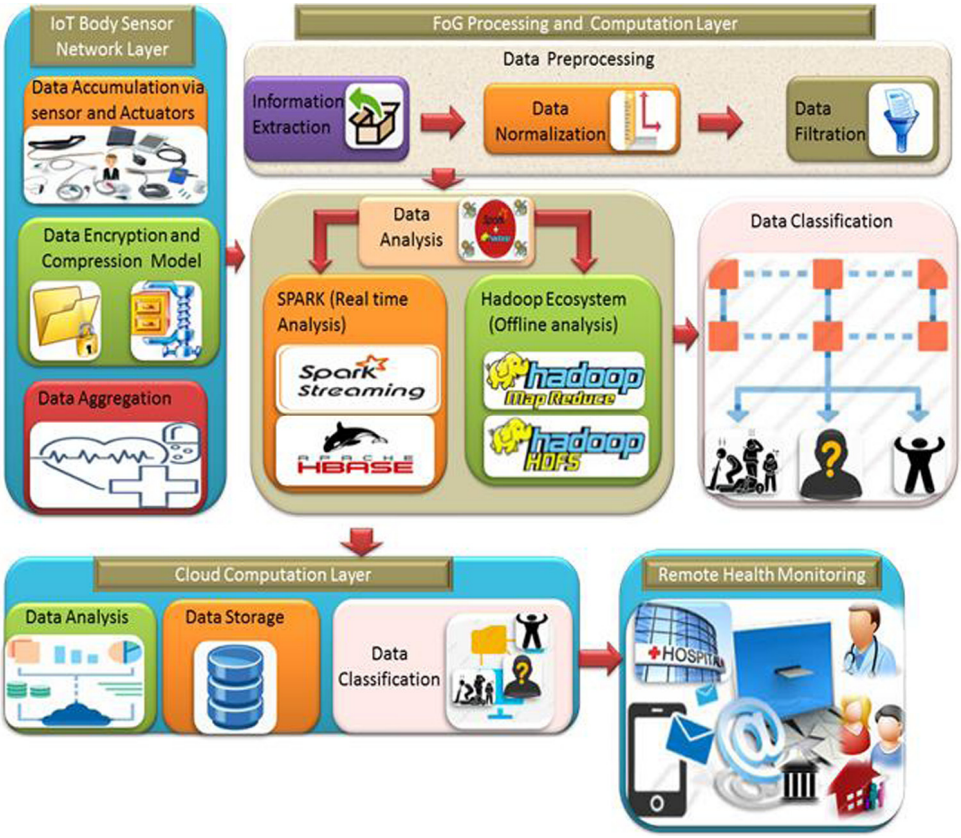


Fig. 2. Proposed system flow.

98.6 °F) can allow data sharing. At node i the sensing information pm_i is encrypted by XORing pm_i and SK_{ib}, this encrypted message is referred to as ciphertext ct_i and calculate hash function f_i by concatenating ciphertext ct_i, secret key SK_{ib} and timestamp T_s. For minimum communication cost and transmission ratio this ciphertext ct_i concatenated with hash function f_i is compressed using data compression technique i.e. D_c. Encrypted and compressed ciphertext message ctm_i generated at Node i by concatenating compressed data D_c, node ID Nid_i and timestamp T_s. Additionally, calculate hash functions $Hf(ctm_i)$ of generated ctm_i and transmitted $Hf\left(ctm_i\right)$, ctm_i and timestamp T_s to aggregator node AgN for aggregating messages from various sensing device nodes.

3.1.3. Data aggregation model

Sensing node devices start the sharing of data by data encryption using shared secret key with FoG server BS and then send out this encrypted and compressed message to AgN. If the AgN is only one hop away from BS Fog server then it can share data directly with BS otherwise AgN will share data with its nearby AgN as a mediator to swap over data with BS. Mediator AgN can aggregate this data with its own data by applying a delimiter to differentiate between the data. Algorithm 2 represents the detailed procedure of message aggregation.

3.1.4. Data pre-processing

Data preprocessing plays a vital role while dealing with different parameters with different scales and units. Data Preprocessing consists of information extraction, data normalization, data filtration, and data analysis and classification.

3.1.5. Information extraction

The base station (BS) at FoG layer received encrypted, compressed and aggregated message $CAgN_j$ from all AgN respectively. Algorithm 3 represents the detailed description of information extraction from received $CAgN_j$.

3.1.6. Data normalization

To standardize the aggregated data values, we use the Min–Max normalization technique. In Min–max normalization, transformation of particular P value to another Q value that fits in the range [x–y], which is given as follows:

$$Q = \frac{P - P_{min}}{P_{max} - P_{min}} \times (y - x) + x \tag{1}$$

where, P need to be standardize, P_{min} represent minimum value, P_{max} represent maximum value, y is the upper bound limit, and x is the lower bound limit.

3.1.7. Data filtration

During data filtration step noise removal and filtration of data is performed using Kalman filter (KF) that increases data processing speed and separates out important and irrelevant noisy data. The five step process of KF data filtration are as follows:

- Initialize transition parameter, noise covariance parameter, observed noise covariance parameter and control input parameter.
- Using previous state parameter information, new state parameter information is computed.
- Prediction of Current state based on previous state.
- Join current observation with current prediction
- Covariance updates of predicted parameter and observed parameter.

3.1.8. Data analysis

After receiving the filtered data, core processing is performed and data is performed during the data analysis process. In a Smart healthcare system with IoT enabled WSN, we need to manage the high-volume as well as high speed data (referred to as Big Data). Therefore,

Algorithm 1: **Data Encryption and compression at Sensing device node Nid_i**

Input: $pm_i \leftarrow sensing\ values\ from\ sensing\ device\ node$; $Nid_i \leftarrow Sensing\ device\ node\ ID$; $SK_{i\,b} \leftarrow Secret\ Key\ between\ Sensing\ device\ node\ and\ Fog\ server\ BS$; $ct_i \leftarrow Ciphertext\ at\ node\ i$; $T_s \leftarrow Timestamp$; $D_c \leftarrow Compressed\ data$; $hf_i \leftarrow hash\ function\ at\ sensing\ device\ node\ i$; $ctm_i \leftarrow generated\ ciphertext\ message\ of\ plaintext$; $thv \leftarrow threshold\ value$;$\Delta t \leftarrow time\ window$

Procedure $Generate\ Msg(\ ctm_i)$

begin

1. ***If*** $(pm_i > thv)$ ***then***
2. $Ciphertext\ ct_i = pm_i \oplus SK_{i^b}$
3. $hash\ function\ hf_i = hf_i(ct_i \| SK_{i^b} \| T_s)$
4. $Compressed\ data\ D_c = Compression\ (ct_i \| hf_i)$
5. $Ciphertext\ message\ of\ plaintext\ ctm_i = (D_c \| T_s \| Nid_i)$
6. $Hash\ function\ Hf = Hf(ctm_i)$
7. ***Else***
8. $Discard\ the\ message\ errorneous\ value$
9. ***End If***
10. ***End Procedure***

Algorithm 2: **Data aggregation at aggregator node**

Initialzation:

$Actm = null$; // $aggregated\ and\ compressed\ message\ by\ AgN(Aggregator\ Node)$

$T_s' \leftarrow timesamps\ calculate\ at\ AgN$;

$CAgN_j \leftarrow Encrypted, compressed\ andaggregated\ message\ from\ all\ AgN$;

$SKA_{j^b} \leftarrow Secret\ Key\ is\ shared\ between\ aggregator\ node\ and\ BS$; ;$\Delta t \leftarrow time\ window$

Recieved: $Hash\ function\ Hf = Hf(ctm_i)$

$Ciphertext\ message\ of\ plaintext\ ctm_i = (D_c \| T_s \| Nid_i)$

$T_s \leftarrow Received\ timestamp$

Compute: $Hash\ function\ Hf' = Hf'(ctm_i)$

Procedure $Aggragate\ Msg(ctm_i)$ ***begin***

1. ***If*** $\boldsymbol{T_s' - T_s} < \Delta t\ then$
2. ***If*** $(Hf'(ctm_i) = Hf(ctm_i))then$
3. $Actm = Actm \| ctm_i$
4. ***Else***
5. $Message\ discarded\ due\ to\ data\ integrity\ checking$
6. ***End If***
7. ***Else***
8. $Message\ Discarded\ due\ to\ latest\ timestamp\ freshness$
9. ***End If***
10. $CAgN_j = Actm_j \oplus SKA_{j^b}$
11. $Hf = Hf(CAgN_j)$
12. ***End Procedure***

to manage this Big Data, we need highly efficient parallel processing tools and techniques that efficiently process this enormous and large amount of high-speed data. The Hadoop ecosystem is the biggest tool for processing this high speed Big Data, it takes data from each individual unit of different datasets (e.g.,DU1, DU2 ... DU_k) and generate final data outcome for the complete data (e.g., FD1, FD2 ... FD_k) which is expressed as:

$$\sum_{i=1}^{k} DU_k => \sum_{i=1}^{k} FD_i \tag{2}$$

So in order to attain this goal, MapReduce technique is deployed at Hadoop Ecosystem in which there is at least one master node and various slave nodes. Master node receives high speed data to process.

Algorithm 3 : Information extarction at Fog layer

Receive: $CAgN_j = Actm_j \oplus SKA_{j^b}$

Compute: $Hf' = Hf'(CAgN_j)$

Procedure: $Info\_Extract\_Msg\ (CAgN_j)$ ***Begin***

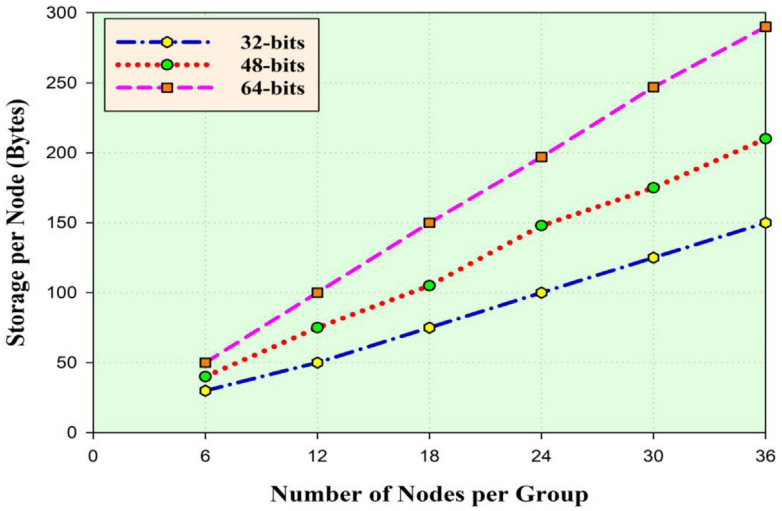
1. ***If*** $(Hf'(CAgN_j) = Hf(CAgN_j))$ *then*
2. $Actm_j = CAgN_j \oplus SKA_{j^b}$ *//Decryption*
3. ***For*** *count* 1 *to* n
4. ***Extract*** $\boldsymbol{ctm_i} = (D_c \| T_s \| Nid_i)$ *from* AgN
5. ***If*** $\boldsymbol{T_s' - T_s} < \Delta t$ *then*
6. $Decompressed\ data\ DD_c = Decompression\ (ct_i \| hf_i)$
7. ***Compute*** $hf_i = hf_i(ct_i \| SK_{i^b} \| T_s)$
8. ***If*** $(hf_i = hf_i')$*then*
9. $pm_i = ctm_i \oplus SK_{i^b}$
10. *Stored the recieved message* pm_i *at FoG layer local storage*
11. ***Else***
12. *Message discarded due to data integrity checking*
13. *End If*
14. ***Else***
15. *Message Discarded due to latest timestamp freshness*
16. *End If*
17. ***Else***
18. *Message discarded due to data integrity checking in Actm*
19. *End If*
20. *Transmit " Successful reply Msg to AgN*
21. *End Procedure*

Algorithm 4: ***Health state classification and monitoring at Fog Layer***

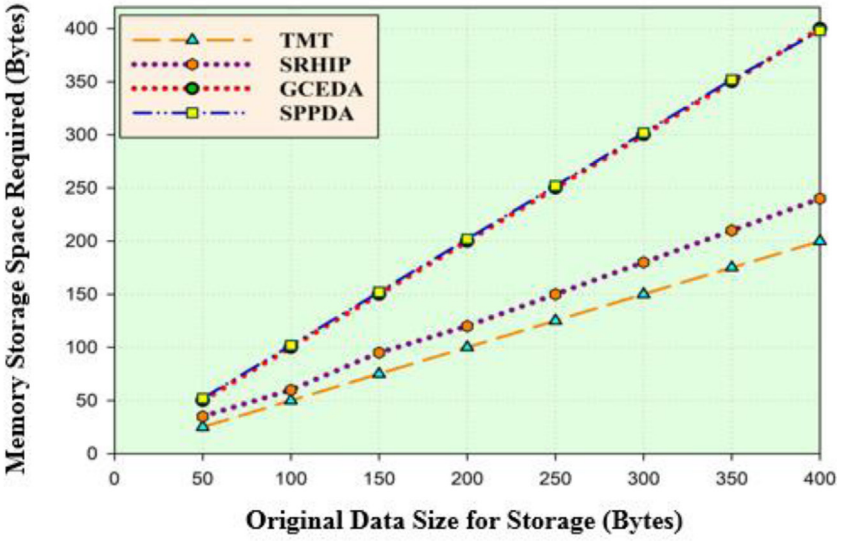
1. *Start*:
2. $\Delta t \leftarrow$ *time window*;
3. $SnA_i \leftarrow$ *sensing attribute at sensing device node like temperature, BP etc*;
4. $SnV_i \leftarrow$ *sensing value of attribute at sensing devive node like* 37℃ *etc*;
5. $thv \leftarrow$ *set threshold value*
6. ***For*** *count* $i = 1$ *to* p
7. $SnA_{i=}SnV_i$
8. ***If*** $(SnV_i > thv)$ ***then***
9. *Person is infected*
10. $Infected\ Class = SnV_i$
11. ***Else***
12. $Healthy\ Class = SnV_i$
13. $Compute: Probability\ of\ Infected\ Healthy\ Index\ (IHI)$ *using eqn.* 3
14. ***For*** $SnA_i = 1$ *to* q
15. ***If*** $(pro(IHI_{SnA_i}) > thv)$ *// thv is Predefined threshold value by Naive Bayes Classifier*
16. $Health\ State = Infected$
17. $Goto: Algorithm\ 5$
18. ***Else***
19. $Health\ State = Healthy$
20. $Goto: Start$
21. ***Exit***

At master Node this High speed data is split into various fixed size data packets which process simultaneously at different slave nodes. In order to distribute this fixed size data packet to different slave nodes Hadoop employs Hadoop Distributed File System (HDFS).

Each data packet is replicated over more than one slave node and each slave node processed stored data packet simultaneously using Map-function. The master node is responsible for concatenating results obtained from various slave nodes using Reduce-function. The main



(a)



(b)

Fig. 3. Storage with (a) variable secret key size and (b) compression scheme.

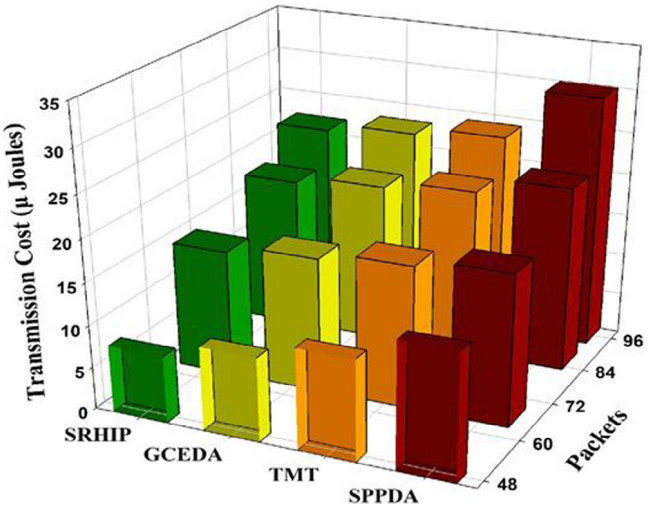


Fig. 4. Transmission cost based on proposed compression scheme.

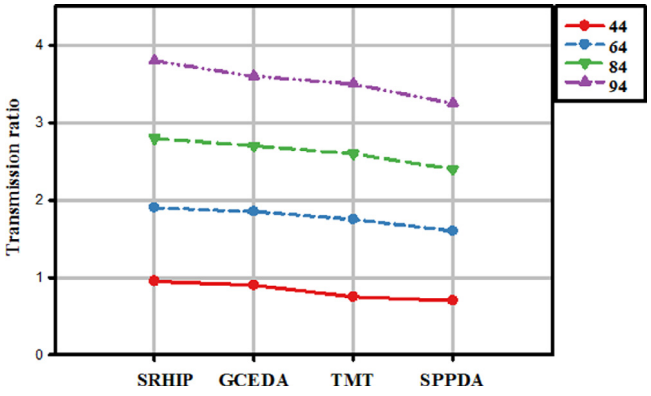


Fig. 5. Transmission ratio based on proposed system.

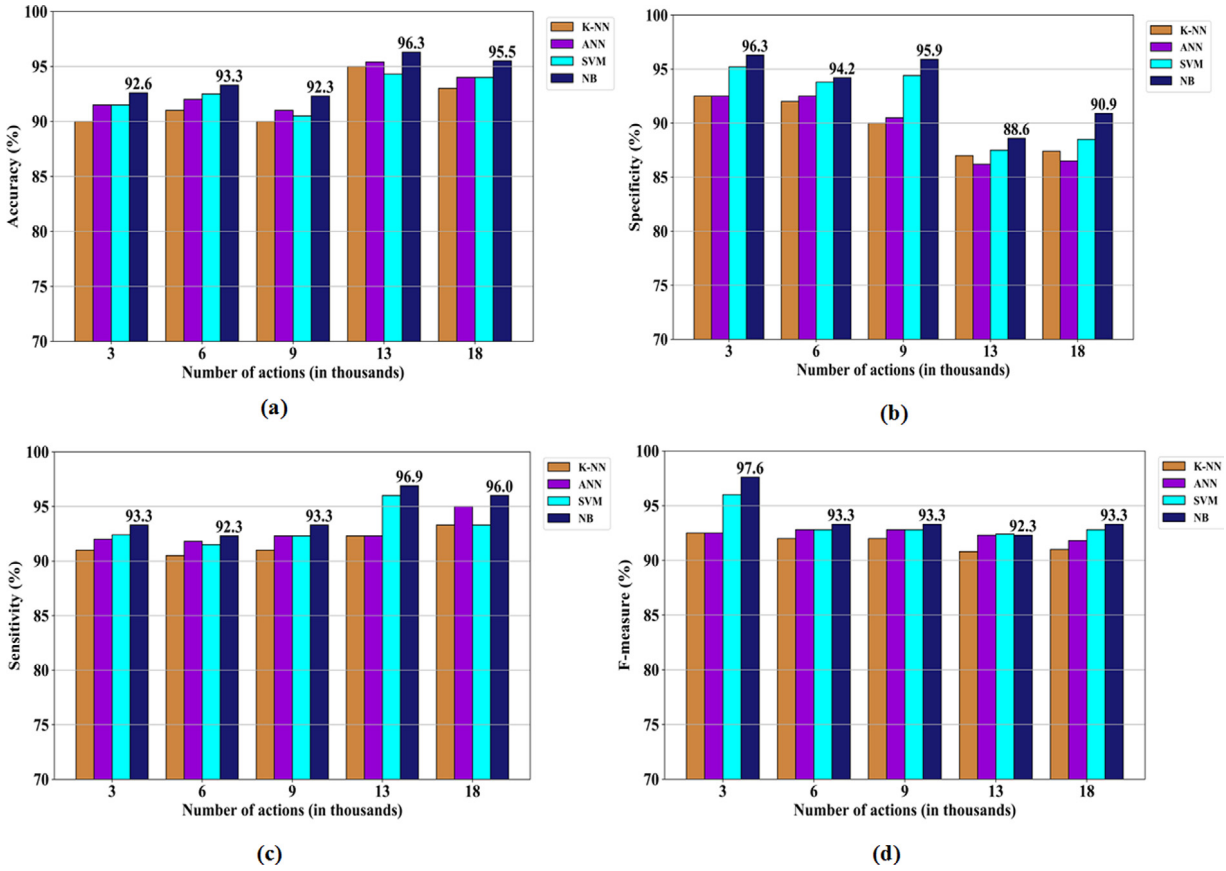


Fig. 6. Classification efficiency of healthcare dataset (a) accuracy (b) specificity (c) sensitivity (d) F-measure.

disadvantage of Hadoop standalone system is that it is suitable for batch processing only but in SRHIP we are dealing with real time information also. So, we use Apache Spark for real-time data processing along with Hadoop Ecosystem. So, we perform parallel and simultaneous processing of data using Hadoop Ecosystem while real-time data stream processing is performed using Apache Spark for extracting and computing parameters, and producing results.

Data classification phase classifies users into infected and healthy categories. Infection analysis for different sensing attribute events SnA_i (temperature, blood pressure, heart rate etc.) is carried out using probability parameter, referred to as Infected Health Index (IHI) and it can be expressed using following conditional probability equation:

$$IHI_{SnA_i} = Pro\left(\frac{S_h}{SnA_1 \cup SnA_2 \cup \ldots SnA_N}\right) \tag{3}$$

where, S_h denotes health state (infected or healthy), SnA_i denotes infected health index values for different sensing attribute events. Healthy state condition of particular person using Naïve bayes classifier is defined by following equation:

$$Pro(IHI^{\Delta t}_{SnA_i}) = Max \prod_{t=1}^{N} Pro(SnA_i^t | IHI_{SnA_i}) \tag{4}$$

where, N denotes time window length for analysis.

3.1.9. Cloud computation layer

At the FoG layer we perform data analysis efficiently still it has some limitations in terms of resources so further analysis and storage of data that is not managed and processed by fog computation layer is performed at Cloud Layer. Here also for big data analysis we use Hadoop Ecosystem for simultaneous processing of batch processes with HDFS distributed file structure and Apache SparK for real time data stream processing with HBASE real-time in-memory and lookups caching.

3.1.10. Remote health monitoring

Proposed remote health monitoring is composed of two phases. In the phase one, health analysis is performed using infection severity on the basis of naïve bayes classification model at FoG Layer explained in Algorithm 4. In the second phase, notification is sent to the caretaker and doctor along with the data analysis report so that instant action can be taken in order to cure the infection. The second phase analysis is performed at Cloud layer explained in Algorithm 5.

4. Simulation results

In order to validate our proposed work, we perform simulation by deploying health and environmental sensors over an area of 2000 × 2000 m. It consists of sink and AgN nodes that perform various functions like encryption, compression, hashing, sending, receiving, decompression and decryption using associated algorithms. Performance evaluation of the proposed SRHIP system is measured in terms of storage, transmission cost, communication ratio, classification efficiency, and throughput. Comparison evaluation of proposed work is performed with existing benchmark schemes including TMT compression method, GCEDA [37] and SPPDA [38] data aggregation method.

4.1. Compression based storage

We consider a cluster of sensor nodes with 6 to 36 nodes per cluster, single node storage with variable size of keys for different cluster sizes is demonstrated in Fig. 3(a) and from the figure it was observed that as key size increases, storage per node is increased. For clusters with size 18, storage per node required is 70 byte, 110 bytes and 144 bytes with the key size 32, 48 and 64 bits. Compressed form of storage is demonstrated in Fig. 3(b) Using compression ratio α and original data

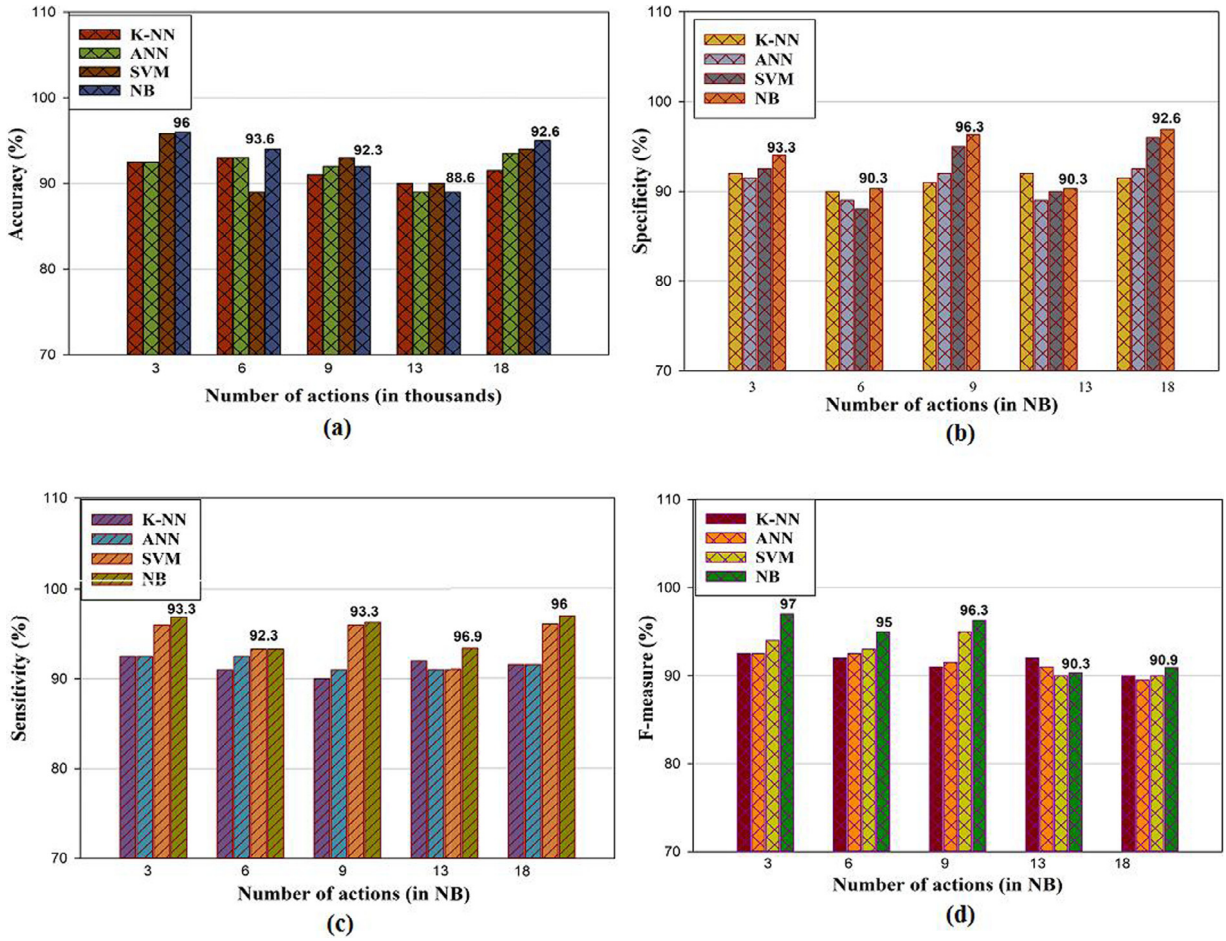


Fig. 7. Classification efficiency of environmental dataset (a) accuracy, (b) specificity, (c) sensitivity, and (d) F-measure.

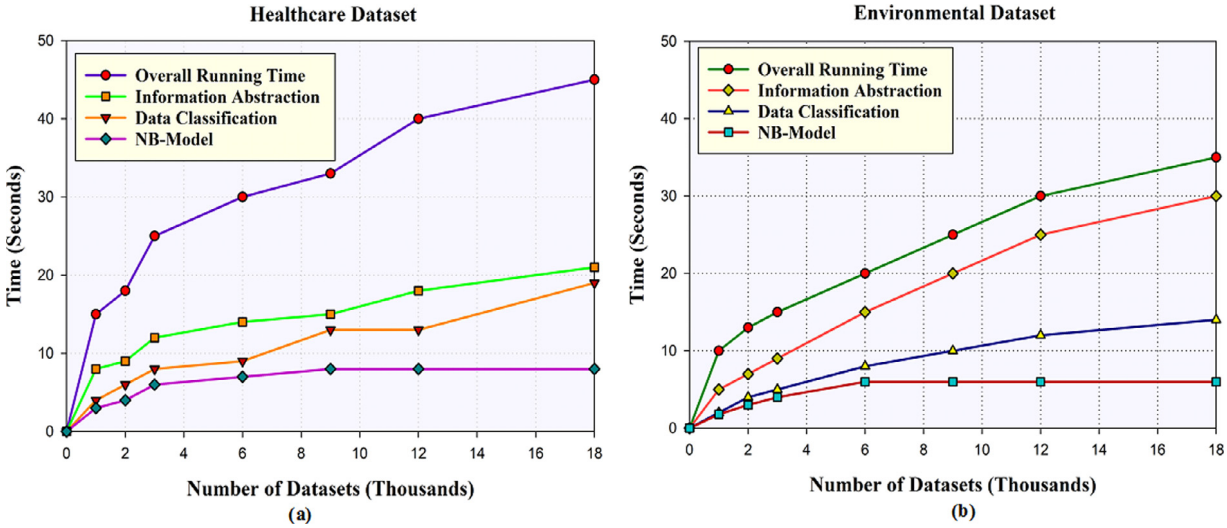


Fig. 8. Feasibility and efficiency measure of proposed NB based SRHIP system (a) Healthcare Dataset (b) Environmental surrounding dataset.

size OD_s, compressed data size CD_s can be obtained by following equation:

$$CD_s = OD_s - \frac{\alpha OD_s}{100} \tag{5}$$

Based on the compression scheme used in the proposed system we attain better storage of 4.80% than TMT, 9.63% than GCEDA and 15.66% than SPPDA.

4.2. Compression based communication-cost

Communication or transmission cost T_c in term of data aggregation message exchange over communication network is computed using following:

$$T_c = (\varepsilon_s \times N_s) + (N_r \times \varepsilon_r) + (D_r \times \varepsilon_s) \tag{6}$$

where, ε_s denotes energy consumption for transmitting a particular message, ε_r represents energy consumption in receipt of a particular message, N_s denotes Total number of sent messages, N_r denotes total number of received messages and D_r represents the number of dropped packet rate. Based on the proposed data compression scheme proposed SRHIP system achieves less transmission cost of about 2.62%, than TMT, 5.24% than GCEDA and 40.10% SPPDA. Fig. 4 demonstrate transmission cost base on proposed compression scheme.

4.2.1. Transmission ratio:

Transmission ratio is ratio of number of received packet to total number of sent packet which is expressed by following equation:

$$T_r = \frac{N_r}{N_s} \tag{7}$$

Fig. 5 demonstrates transmission ratio for different size messages. With exchange of 84 messages, ratios of transmission are 0.813, 0.872, 0.918, and 0.965 respectively for SPPDA aggregation method, GCEDA scheme, TMT approach and proposed SRHIP, respectively. From the result it was seen that proposed SRHIP attains 4.870%, 9.630% and 15.750% improved ratio of transmission in comparison to TMT approach, GCEDA scheme and SPPDA aggregation method respectively. Fig. 5 demonstrates transmission ratio based on the proposed system.

4.3. Efficiency of data classification

For evaluating classification efficiency and performance measure of the proposed SRHIP system, we use Healthcare dataset of 19,908 sample instances with attributes like breath rate, blood pressure, heart rate etc. for manifold time frames from UCI Dataset repository and Environmental surrounding dataset of more than 18,000 data instances with attributes such as pressure, temperature, humidity etc. from data repository of US EPA comprising of more than 18,000 datasets with included attributes like room temperature, pressure, humidity. Classification efficiency Naïve Bayes classifier of proposed system over both the dataset is measured in terms of accuracy, sensitivity, specificity and F-measure. Obtained results of proposed SRHIP for different dataset are then compared with existing classification techniques including Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Artificial Neural Network (ANN).

Subsequently, we did comparative study of proposed SRHIP for different dataset with existing classification techniques including Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Artificial Neural Network (ANN) based on performance metric. Accuracy, specificity, sensitivity, and F-measure of the proposed NB based SRHIP system are 96.5%, 96.7%, 93.6%, and 94.3% respectively. While the classifiers SVM, ANN and KNN achieve the accuracy of 90.1%, 91.3%, and 94.4%; specificity of 91.1%, 92.3%, and 91.3%; f-measure of 91.5%, 91.2% and 90.3% respectively. From the result it is observed that Naïve Bayes (NB) classifier is extremely efficient in health datasets analysis. Fig. 6(a)–(d) demonstrates comparison results in 2D-Graph representation for Healthcare datasets.

From the result it is observed that Naïve Bayes (NB) classifier is extremely efficient in environmental surrounding datasets analysis with accuracy, specificity, sensitivity and F-measure of 96.9%, 91.9%, 95.90% and 92.90% respectively. With two different dataset Fig. 7(a) and (b) represents simulation results for various activities including overall running time, information abstraction time, data classification time and infection severity index prediction time using Naïve Bayes Model for healthcare dataset and environmental surrounding dataset. Fig. 7(a) represents a running time of 19.20 s for information abstraction and infection severity index prediction of healthcare dataset and 5.3 From Fig. 8 it was seen that the proposed SRHIP system is extremely proficient in real time analysis for evaluating various mention activities. Henceforth, based on these evaluations, it can be concluded that the proposed model is temporally efficient and highly feasible in assessing large numbers of heterogeneous datasets.

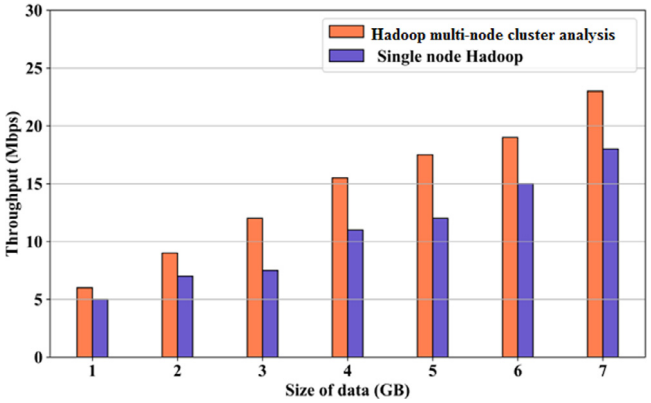


Fig. 9. Throughput of proposed SRHIP system.

Throughput: Fig. 9 shows efficiency of proposed system in terms of throughput using Hadoop and Apache Spark ecosystem with single node and multi-node cluster analysis. From Fig. 9 it is seen that with the increase in data size processing speed decreases in single node environments in comparison to multi-node environments. So the proposed SRHIP system with multi-node cluster is highly efficient in comparison to the single node Hadoop and Apache Spark ecosystem system.

5. Conclusion

In the field of healthcare, IoT sensors enable WSN to generate large amounts of high-speed and high-volume data that need an efficient approach for analyzing such a big volume of data with secure, confidential and accurate infection severity prediction mechanisms. With this aim we propose FoG assisted smart and real time healthcare information processing (SRHIP) systems that process and analyze data using Hadoop and Apache Spark systems. The collected data information from the data preprocessing step is further processed and analyzed at real-time using Hadoop and Apache Spark ecosystem. For infection severity prediction we use naïve bayes classifiers and compare their performance with existing benchmark classifiers including KNN, ANN, and SVM. Based on the comparative analysis of different classifiers, proposed NB based SRHIP outperforms other classifiers in terms of accuracy, specificity, sensitivity and F- measure. We also evaluate our proposed system in term data storage, transmission cost and transmission ratio. Proposed SRHIP system outperforms other existing benchmark schemes including SPPDA and GCEDA aggregation method and TMT approach. Simulation Results prove the dominance of our proposed EHDA scheme as compared to TMT, GCEDA and SPPDA schemes. The proposed SRHIP system needs less transmission cost of 40.10% in comparison to SPPDA, 100% fewer bytes are compromised in comparison to GCEDA. Our proposed system data size reduction of 60% reduction due to proposed compression scheme in comparison to other benchmark strategies that offer 40% of reduction. In future, this work can be extended to the deployment of sensing devices to different mobility situations where FoG servers will be shifting based on the mobility in different areas.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] S. Ramírez-Gallego, A. Fernández, S. García, M. Chen, F. Herrera, Big data: Tutorial and guidelines on information and process fusion for analytics algorithms with MapReduce, Inf. Fusion 42 (2018) 51–61.

[2] B.A. Hammou, A.A. Lahcen, S. Mouline, Towards a real-time processing framework based on improved distributed recurrent neural network variants with fasttext for social big data analytics, Inf. Process. Manage. 57 (1) (2020) 102122.
[3] N. Lozada, J. Arias-Pérez, G. Perdomo-Charry, Big data analytics capability and co-innovation: An empirical study, Heliyon 5 (10) (2019) e02541.
[4] Y. Chen, IoT, cloud, big data and AI in interdisciplinary domains, 2020.
[5] A. Alabdulatif, I. Khalil, X. Yi, Towards secure big data analytic for cloud-enabled applications with fully homomorphic encryption, J. Parallel Distrib. Comput. 137 (2020) 192–204.
[6] H. Li, Y. Wu, D. Cao, Y. Wang, Organizational mindfulness towards digital transformation as a prerequisite of information processing capability to achieve market agility, J. Bus. Res. (2019).
[7] M. Albergaria, C.J.C. Jabbour, The role of big data analytics capabilities (BDAC) in understanding the challenges of service information and operations management in the sharing economy: Evidence of peer effects in libraries, Int. J. Inf. Manage. 51 (2020) 102023.
[8] D. Jiang, The construction of smart city information system based on the Internet of Things and cloud computing, Comput. Commun. 150 (2020) 158–166.
[9] X. Xiao, X. Hou, X. Chen, C. Liu, Y. Li, Quantitative analysis for capabilities of vehicular fog computing, Inform. Sci. 501 (2019) 742–760.
[10] S. Memon, M. Maheswaran, Using machine learning for handover optimization in vehicular fog computing, in: Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing, April, 2019, pp. 182–190.
[11] S. Zhou, Y. Sun, Z. Jiang, Z. Niu, Exploiting moving intelligence: Delay-optimized computation offloading in vehicular fog networks, IEEE Commun. Mag. 57 (5) (2019) 49–55.
[12] I. Sorkhoh, D. Ebrahimi, S. Sharafeddine, C. Assi, On leveraging the computational potential of fog-enabled vehicular networks, in: Proceedings of the 9th ACM Symposium on Design and Analysis of Intelligent Vehicular Networks and Applications, November, 2019, pp. 9–16.
[13] W. Sun, Y. Zhao, L. Sun, Big data analytics for venture capital application: Towards innovation performance improvement, Int. J. Inf. Manage. 50 (2020) 557–565.
[14] K.N. Kaipa, D. Ghose, Applications to ubiquitous computing environments, in: Glowworm Swarm Optimization, Springer, Cham, 2017, pp. 157–181.
[15] M. Ghasemaghaei, G. Calic, Does big data enhance firm innovation competency? The mediating role of data-driven insights, J. Bus. Res. 104 (2019) 69–84.
[16] P. Akhtar, Z. Khan, R. Rao-Nicholson, M. Zhang, Building relationship innovation in global collaborative partnerships: big data analytics and traditional organizational powers, R & D Manage. 49 (1) (2019) 7–20.
[17] M.M. Chan, R.B. Plata, J.A. Medina, C. Alario-Hoyos, R.H. Rizzardini, M. de la Roca, Analysis of behavioral intention to use cloud-based tools in a MOOC: A technology acceptance model approach, J. Univers. Comput. Sci. 24 (8) (2018) 1072–1089.
[18] M. Garmaki, I. Boughzala, S.F. Wamba, The effect of big data analytics capability on firm performance, in: PACIS, June, 2016, p. 301.
[19] T. Rantala, K. Palomäki, K. Valkokari, Challenges of creating new B2B business through big data utilization, in: ISPIM Conference Proceedings, The International Society for Professional Innovation Management (ISPIM), 2018, pp. 1–15.
[20] N. Elgendy, A. Elragal, Big data analytics: a literature review paper, in: Industrial Conference on Data Mining, Springer, Cham, 2014, pp. 214–227.
[21] Y. Huang, P. Gao, Y. Zhang, J. Zhang, A cloud computing solution for big imagery data analytics, in: 2018 International Workshop on Big Geospatial Data and Data Science (BGDDS), IEEE, 2018, pp. 1–4.
[22] S. Yang, IoT stream processing and analytics in the fog, IEEE Commun. Mag. 55 (8) (2017) 21–27.
[23] M.R. Anawar, S. Wang, M. Azam Zia, A.K. Jadoon, U. Akram, S. Raza, Fog computing: An overview of big IoT data analytics, Wirel. Commun. Mobile Comput. 2018 (2018).
[24] S.M. Nagarajan, U.D. Gandhi, Classifying streaming of Twitter data based on sentiment analysis using hybridization, Neural Comput. Appl. 31 (5) (2019) 1425–1433.
[25] N.S. Murugan, G.U. Devi, Detecting streaming of Twitter spam using hybrid method, Wirel. Pers. Commun. 103 (2) (2018) 1353–1374.
[26] N.S. Murugan, G.U. Devi, Feature extraction using LR-PCA hybridization on twitter data and classification accuracy using machine learning algorithms, Cluster Comput. 22 (6) (2019) 13965–13974.
[27] N.S. Murugan, G.U. Devi, Detecting spams in social networks using ML algorithms-a review, Int. J. Env. Waste Manage. 21 (1) (2018) 22–36.
[28] F. Mehdipour, B. Javadi, A. Mahanti, FOG-engine: Towards big data analytics in the fog, in: 2016 IEEE 14th Intl Conf on Dependable, Autonomic and Secure Computing, 14th Intl Conf on Pervasive Intelligence and Computing, 2nd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress, DASC/PiCom/DataCom/CyberSciTech, IEEE, 2016, pp. 640–646.
[29] J. He, J. Wei, K. Chen, Z. Tang, Y. Zhou, Y. Zhang, Multitier fog computing with large-scale IoT data analytics for smart cities, IEEE Internet Things J. 5 (2) (2017) 677–686.
[30] T. Kolajo, O. Daramola, A. Adebiyi, Big data stream analysis: a systematic literature review, J. Big Data 6 (1) (2019) 47.
[31] F. Matsebula, E. Mnkandla, A big data architecture for learning analytics in higher education, in: 2017 IEEE AFRICON, IEEE, 2017, pp. 951–956.
[32] E. Spanò, S. Di Pascoli, G. Iannaccone, Low-power wearable ECG monitoring system for multiple-patient remote monitoring, IEEE Sens. J. 16 (13) (2016) 5452–5462.
[33] S.H. Chang, R.D. Chiang, S.J. Wu, W.T. Chang, A context-aware, interactive M-health system for diabetics, IT Prof. 18 (3) (2016) 14–22.
[34] H.T. Cheng, W. Zhuang, Bluetooth-enabled in-home patient monitoring system: Early detection of alzheimer's disease, IEEE Wirel. Commun. 17 (2010) 74–79.
[35] S. Milici, A. Lázaro, R. Villarino, D. Girbau, M. Magnarosa, Wireless wearable magnetometer-based sensor for sleep quality monitoring, IEEE Sens. J. 18 (5) (2018) 2145–2152.
[36] C.F. Pasluosta, H. Gassner, J. Winkler, J. Klucken, B.M. Eskofier, An emerging era in the management of Parkinson's disease: wearable technologies and the internet of things, IEEE J. Biomed. Health Inf. 19 (6) (2015) 1873–1881.
[37] D. Mantri, N.R. Prasad, R. Prasad, Grouping of clusters for efficient data aggregation (GCEDA) in wireless sensor network, in: 2013 3rd IEEE International Advance Computing Conference, IACC, IEEE, 2013, pp. 132–137.
[38] C. Zhang, C. Li, J. Zhang, A secure privacy-preserving data aggregation model in wearable wireless sensor networks, J. Electr. Comput. Eng. 2015 (2015).