Measurement 154 (2020) 107424

Contents lists available at ScienceDirect

Measurement

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

An approach to mitigate challenges to the Electronic Health Records storage

Rumenigue Hohemberger [a,c], Cinara Ewerling da Rosa [a], Felipe Rubin Pfeifer [b], Rodrigo Mello da Rosa [b], Paulo Silas Severo de Souza [b], Arthur Francisco Lorenzon [c], Marcelo Caggiani Luizelli [c], Fábio Diniz Rossi [a,*]

[a] Federal Institute of Education, Science and Technology Farroupilha, Alegrete, RS, Brazil
[b] Pontifical Catholic University of Rio Grande do Sul, Porto Alegre, RS, Brazil
[c] Federal University of Pampa, Alegrete, RS, Brazil

ARTICLE INFO

Article history:
Received 16 October 2019
Received in revised form 22 November 2019
Accepted 16 December 2019
Available online 7 January 2020

Keywords:
Big data
eHealth
IoT
Storage

ABSTRACT

The Internet of Things (IoT) has been embraced in our daily lives as it was in many sectors of the economy. In the health sector, many possible applications have emerged, most often in regards to the detection and prevention of chronic diseases. One of many possibilities is patient monitoring through wearable devices and mobile applications. In order to promptly identify vital signs changes and to reduce the incidence of serious diseases, collecting data in real-time and feeding it to Electronic Health Records (EHR) in hospitals or medical clinics can further be of assistance to decision making. The downside of this process is that it involves storing such a large amount of data, thus becoming almost unfeasible. This work attempts to address this problem and present possible solutions and insights.

1. Introduction

The Internet of Things (IoT) [1] is referred to as the global connection of "intelligent objects" through the network structure of the Internet. The concept also refers to the various applications and networking technologies to connect them. This communication allows, among other facilities, continuous recording of data on the state of these objects during their use. Intelligent objects from the IoT generate a lot of detailed data continuously in the environment they are in. The big data technology allows this volume of data to be stored, combined with other data sources, and then analyzed with good performance. Several of the most prominent applications of IoT are within the health area.

Electronic Health (eHealth) [2] is the efficient and effective use of information and communication technologies (ICT) in the medical field, especially in all patient-centered areas. There is a prime component responsible for accelerating the process of eHealth adoption within the health care chain: wearable devices. There has been a constant growth in the consumer market of wearable devices for end-users. One of the many categories of these widely disseminated devices consists of fitness bracelets. These bracelets serve to measure various vital signs, such as heart rate, blood pressure, hours of sleep, distances traveled, calories burned during the day.

The data obtained from wearable devices can be read by mobile applications that analyze such data in order to provide insights regarding stress, sleep, and exercise [3]. This kind of data can be beneficial when stored in an EHR [4] [5], as historical health records, can be used for patients diagnostics. Although this process seems reasonable, it might not be as simple as it appears to be due to the large amount of data that must be stored over time. In the area of computing, greater attention has always been directed at the processing capacity and speed of data communications.

However, only in the last years, the data storage capacity has become the focus of research on widely distributed computing environments. This study aims to present the problem of storing data from medical sensors due to the large amount of space required to maintain a complete and current EHR. Based on this scenario, it is possible to highlight the following contributions from this article:

- A first attempt to state and present the problem of data growth from IoT environments, and the lack of storage space for such data.

* Corresponding author.
E-mail addresses: rumenigue.hohemberger@iffarroupilha.edu.br (R. Hohemberger), cinara.rosa@iffarroupilha.edu.br (C.E. da Rosa), felipe.rubin@acad.pucrs.br (F.R. Pfeifer), rodrigo.mello@acad.pucrs.br (R.M. da Rosa), paulo.silas@acad.pucrs.br (P.S.S. de Souza), arthurlorenzon@unipampa.edu.br (A.F. Lorenzon), marceloluizelli@unipampa.edu.br (M.C. Luizelli), fabio.rossi@iffarroupilha.edu.br (F.D. Rossi).

https://doi.org/10.1016/j.measurement.2019.107424
0263-2241/

- A discussion on the exponential increase of sensor data *versus storage capacity for data collected in IoT-supported eHealth applications.*
- *Data compression methods for minimizing storage requirements without losing relevant information on patients' symptoms over time.*

The remainder of this paper is organized as follows. Section 2 presents an overview of challenges faced in terms of storage and data reduction when applied to eHealth. Section 3 presents a formulation of the storage problem, which is the focus of this work. Section 4 presents our proposed solutions followed by Section 5 where are presented some evaluation. Finally, Section 6 reiterates the work and proposed solutions while also giving an insight into future research.

2. Background and related work

A frequent recurrence in eHealth applications regards acquiring patients' data and sending it to the cloud where it can store. This section presents the background and related work for eHealth applications in eHealth, followed by challenges and approaches for reducing the significant amount of data generated by such applications.

2.1. Storage challenges in eHealth applications

In broad terms, IoT platform providers who wish to support the Healthcare industry must provide many features such as simple connectivity, secure device management, cloud data storage, big data analytics, and others [6]. Services who plan on giving users and researchers access to the stored health data must also consider the retrieval of such data in a presentable format [7] [8].

Other problems regarding data access and ownership should also be considered [9] [10]. During a patient's lifetime, EHR may become fragmented across multiple institution-centered silos due to events such as switching health insurance companies [6], and such events may appear as a problem later when a historical analysis over a patient health state is required. Nonetheless, efforts are being made to assure the security and integrity of patient's data between institutions [11] [12] [13]. (See Table 1).

While data transmission, storage, and analysis of time-series information is a frequent recurrence in eHealth systems, a lifetime patient monitoring and EHR storage will unmistakably bring about big-data aspects for any eHealth service.

An exponential increase in data generation [14] will accompany the increasing adoption of eHealth. Therefore, data reduction strategies are increasingly becoming an overwhelming need for IoT-supported eHealth applications.

Table 1
Description of parameters used by SAS for filtering vital signs coming from sensors in order to reduce the amount of stored data while preserving measurements of interest, i.e., vital signs that may represent patient's symptom variations.

Parameter	Description
$\mathbb{B}$	Set of patients' vital signs gathered from sensors
ϱ	Size limit for each batch of measurements $\in \mathbb{B}$ that will be used as input for the clustering algorithm
$\mathbb{K}$	Set of random vital signs $\in \mathbb{B}$ used as initial centroids in the distribution
$\mathbb{C}$	Set of clusters of vital signs calculated according to the similarity of data
$\mathbb{M}$	Set of cluster centroids (e.i., vital signs) to be stored. Each element in this set represents a different behavior experienced by the patient over time. This data is stored and presented by SAS to help doctors to identify patient's symptom variations

2.2. Data reduction strategies

The problem of handling large amounts of generated data in computer systems is not quite a recent challenge. In fact, since the beginning of computing, reducing the size of data was a topic of interest, considering how expensive and capacity-constrained the storage devices were [15] [16]. More recently, the advent of the Internet and the increasing popularity of electronic devices boosted, even more, the research on data compression to support data-intensive applications such as high-quality video streaming and online gaming nowadays.

Classical data compression approaches took advantage of data repetition to minimize the overall file size. One of the first techniques is dated to 1838 when Samuel Morse proposed the Morse Code [17] to reduce the size of telegraphs by using smaller sequences to represent recurrent letters. Latter in 1951, the same principle was adopted by David Huffman with the Huffman Coding [18], which assigns each character a prefix code whose size depends on the relative frequency it appears in the file [16] [19]. More recently, several investigations employed data compression algorithms to overcome several research challenges:

- Minimizing power consumption in resource-constrained wireless sensor networks [20];
- Reducing the amount of data necessary for storing and handling data generated by IoT scenarios such as smart cities [21];
- Avoiding network bottlenecks caused by data-intensive IoT applications [22]

Our work complements the previous investigations focused on using lossy data compression strategies to address the challenge of reducing the volume of stored data while preserving data richness to aid the analysis of patients' symptoms over time in IoT eHealth applications.

3. Problem description

Let us consider that each measurement of one of the wearable devices' sensors requires 1 byte of storage. If one takes into account the heart rate, one would need between 70 and 100 bytes per minute (usually). To exemplify the problem, it is considered that storing a heart rate history is performed once every second (60 beats per minute).

Taking this assumption into account, and also considering a life expectancy of 80 years, the following conclusion can be made: to store the heartbeat of a patient during his entire life would require 2.6 gigabytes of space. If one multiplies the number of vital signs and biomarkers that can currently (or in the near future) be monitored by sensors, by the number of people using wearable devices (implantable, swallowable or non-portable) daily, one would quickly reach the mark of Yottabytes [23–27].

As previously depicted, the significant increase in IoT networked devices will lead to an exponential rise in the volume of data that companies will have to manage.

While previous investigations have focused on filtering and modeling the data coming from sensors in order to reduce storage usage, they fail to address the problem that such amount of data will undeniably continue to grow, and such methods will start appearing less effective. Although filtering raw sensor data will still be required due to the necessity to ensure data reliability.

In order to address such issues and considering that the volume of data generated from Wireless Sensor Network will continue to grow, this work proposes different methods that reduce the required amount of storage while ensuring that no relevant information is lost. Further, this work also presents an analysis in terms of costs for each proposed method.

4. Proposed methods

The methods proposed in this section take into account only the metrics discussed in the previous section: the heart rate measured every second and a life expectancy of 80 years.

The first method consists of a Naive **(N)** data storage model, which stores the heart rate every second for the entire lifetime. This method presents the worst case and is unfeasible in the medium and long term due to the large storage capacity required to maintain an EHR.

The second method takes into account some assumptions: the measurement frequency and, consequently, the need for storage space is more significant in childhood/youth and old age. In adulthood, measurements may be less frequent. This is due to the greater vulnerability in these two specters of human life. Therefore, this method is Age **(A)** based, and the frequency measurement curve can be seen in Fig. 1. To do so, this work also proposes to maintain the measurement and storage of data every second between the ages of 0–18 years, and from 60 years. Between 18 and 60 years, this work assumes measuring once a day.

One way to further reduce the amount of data to be stored, improving the previous proposal, is to manage data measured between 18 and 60 years. The algorithm could calculate the standard deviation over the mean of the measurements. If there were no significant outliers, only one measurement per week could be stored (representing the whole set of days). Therefore, this new method would take into account the Age plus Standard Deviation **(ASD)**. The storage of measurements carried out with priority areas (0–18, 60–80) still occurs every second.

So far, there are discussions about EHR in a general way. Now, this work is going to propose a method that can further reduce the impact of storing measurements on data storage, and it turns to particular cases where patients have a family medical history of contracting a specific condition, such as hereditary diseases [28–30].

An example is coronary artery disease (CAD), a congenital heart disease (CHD) [31]. Fig. 2 shows the most common period of onset of the problem and its evolution over time (separated by gender). Based on this behavior, one can infer that the beginning of the issues with coronary affections such as heart attacks occur, for the most part, from the age of 30.

As mentioned before, the probability of developing such conditions increases according to family medical history. Therefore, this work proposes a method that begins storing data only from the age of 30, with one new entry per day (averaged and standard deviation of measurements every hour), yet changing to one measure

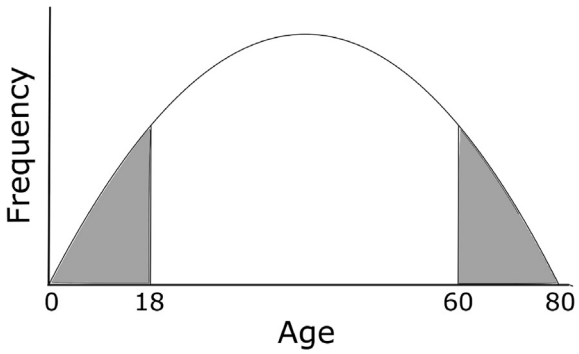

Fig. 1. Normal curve that represents the life expectancy vs. the frequency of measurement and storage of health sensor data. The most critical areas are youth and old age.

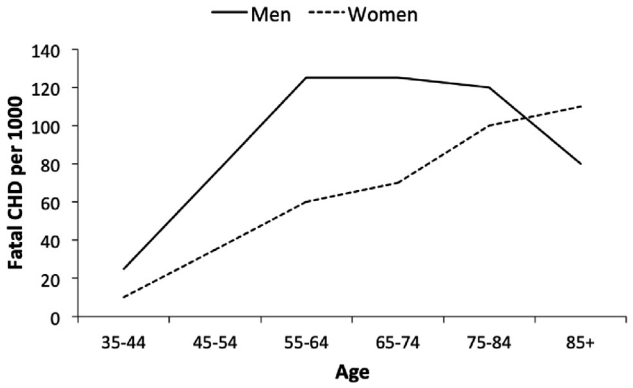

Fig. 2. This chart shows the increase in the number of heart attacks over time, for different genders.

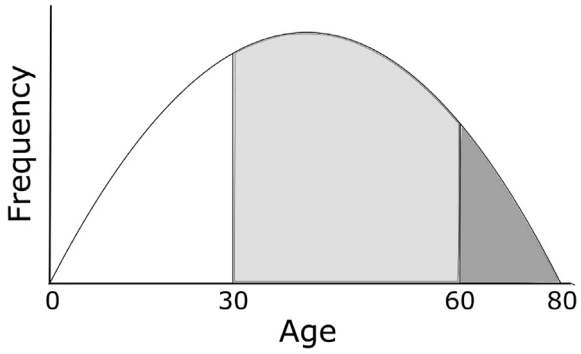

Fig. 3. Normal curve of measurements and data storage according the data shown in Fig. 2.

every second from the age of 60. Fig. 3 exemplifies this method that takes into account age coupled with a family history **(AFH)**.

Until now, this work has presented methods that focus on minimizing the saturation of storage devices by controlling the intervals between measurements, although still taking into account the patient's characteristics such as age, gender, and family medical history. Although these methods do manage to reduce the amount of data being stored, doctors may not be able to analyze specific symptoms over time.

Therefore, this work also proposes a Symptom-Aware Storage **(SAS)** method that stores measurements every second in a temporary buffer and employs a classification method hourly to label measurements according to their meaningfulness. The measurements that end up labeled as similar won't all be stored. Instead, a single measurement of each similarly labeled subset will be preserved.

Consider a patient being monitored for an hour. During the first 30 min, he/she experiences an increase in his/her heart rate, and after such a period (the remaining 30 min), his/her heart rate becomes normal again. Instead of preserving all measurements gathered during this period, SAS classifies this one-hour records and stores only a record that represents a specific behavior. In the given an example, only two measures would be of interest and thus end up recorded: one representing the period the patient might have suffered from tachycardia and others representing the average heart rate period. By storing only these two measurements, doctors would still be able to detect the two existing symptoms. A structured representation for the SAS method is presented in Algorithm 1.

Algorithm 1 Symptom-Aware Storage method.

procedure SAS
  while *true* **do**
    $\mathbb{B} \leftarrow \mathbb{B} \cup m$
    if $|\mathbb{B}| \geqslant \varrho$ **then**
      $\mathbb{K} \leftarrow n$ random measurements $\in \mathbb{B}$
      while centroids changed **do**
        for each measurement $B_k \in \mathbb{B}$ **do**
          Assign $B_k$to the nearest $C_i \in \mathbb{C}$
        end for
        for each cluster $C_i \in \mathbb{C}$ **do**
          $\mathbb{K}_i \leftarrow$ centroid of $\mathbb{C}_i$
        end for
      end while
      $\mathbb{M} \leftarrow \mathbb{M} \cup \mathbb{K}$
      $\mathbb{B} \leftarrow \{\}$
    end if
  end while
end procedure

5. Evaluation and discussion

To evaluate the required space and costs of storing data using each different proposed method, this work used the cumulative sum on Matlab. The results are shown in Fig. 4 and Table 2.

Regarding the required storage space, one can see in the Figure that although method A has advantages over N, ASD manages to reduce the needed space by half, while still guaranteeing that the necessary data will be available in the EHR. The gain (reduction) of storage space (50%) is mostly due to the economy imposed on measurements between 18 and 60 years, which in cases where there are no outliers, only one measurement per week will be stored.

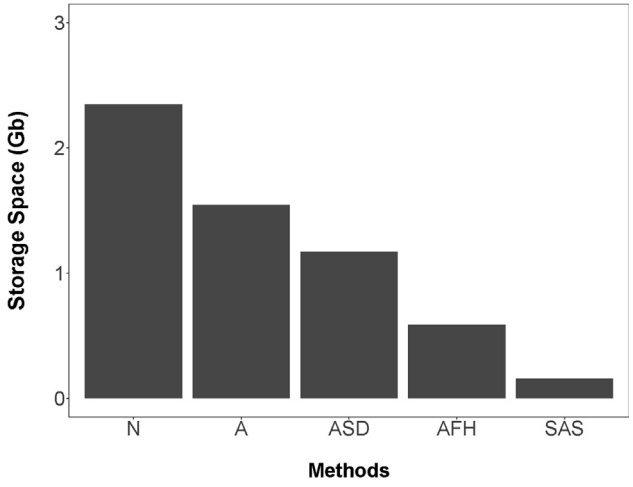

Fig. 4. Comparison of the amount of storage required between the different storage methods.

When a specific condition is monitored, data storage shows a much more substantial reduction of space (75%). This is due to not storing the first 30 years of life. Although there is no complete EHR available for visualization, and there might be other unforeseen conditions that could affect a patient's health, clinics for specialized treatments could still benefit from using this method.

Data storage is a strategic factor for a modern enterprise. Its importance is revealed when this work evaluates the excessive use of technology for communication and corporate activities that, together, lead to an exponential growth of the data to be managed. This data is a vital substratum, not only for the operation of a company but also for generating insights and guiding actions aimed at innovation as well as understanding customer needs. Technologies such as big data have revealed the potential that big data analytics has to provide a market landscape and prescribe actions that will prepare companies for the future. It is also important to note that problems related to inaccuracy or loss of data can cause damage of various kinds to companies, such as unavailability of services, loss of sales, and legal issues. To avoid these setbacks and enable your company to benefit from using data to guide you in your actions, it is essential to know which storage options are most compatible with your business. They must meet the demand for information storage, trust, and distribution.

In this context, a virtuous cycle of supply and demand is established, as the increased need for data and information drives the development of Information and Communication Technologies (ICTs) and, consequently, the evolution of the capacity and volume of technological tools has made this possible. Significant growth in data and information production. It should be noted that in the current world dynamics, this demand for information is diversifying, either due to its speed in its updating, its geographic distribution, or even in areas of knowledge that still lack information production about it.

Therefore, it will significantly depend on our efforts to improve the quality of this data offering by addressing the prerequisites for gaining value from its use, as briefly described in this work.

6. Conclusions

Data coming from IoT sensors are growing exponentially. Cloud environments have a hard time processing this massive amount of data, and infrastructures need to fragment it's transforming through edge computing and mobility, yet that is only a small fraction of a big problem.

The amount of sensors that each person carries on their IoT devices is already substantial, and the tendency is to keep growing. Although this is interesting in terms of health care, as it enables online monitoring of all possible vital signs through sensors, the barrier of physical storage for such amount of data is still an issue.

This work can be divided into three highlights: (i) address the storage problem in the face of the growth of medical data from sensors, (ii) hypothesize some possible solutions, and (iii) present an evaluation of each such possible solutions and their costs. This work consists of the first attempt that puts the storage problem and possible solutions under the light spot.

In terms of implementation, any of the data reduction strategies presented in this work can be implemented on platforms that

Table 2
Average upload price on different cloud platforms, considering a life spam of 80 years of data being stored.

Storage Service	Price in $/Gb/Month	N	A	ASD	AFH	SAS
BackBlaze	$0.005	11.28	7.41	5.63	1.76	0.75
Amazon WS S3	$0.021	47.37	31.13	23.64	7.40	3.16
Microsoft Azure	$0.018	40.60	26.69	20.26	6.34	2.71
Google Cloud	$0.020	45.11	29.65	22.51	7.05	3.01

interface between the sensor data receiving server (web servers usually perform this operation) and the data storage environment (cloud databases typically perform this operation). Therefore, the data analysis and reduction platform performs pre-processing of the received data and saves only those data already filtered and reduced in the database.

There is still a lot of research to be done, due to a large number of sensors, and the combination of vital signs to be monitored, not neglecting the generation of alarms. Perhaps only the heart rate is not enough to detect a condition. Instead, a combination of factors and other signs would be required, such as blood pressure, temperature, and so on.

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] A.H. Alavi, P. Jiao, W.G. Buttlar, N. Lajnef, Internet of things-enabled smart cities: State-of-the-art and future trends, Measurement 129 (2018) 589–606.
[2] H. Zemrane, Y. Baddi, A. Hasbi, Ehealth smart application of wsn on wwan, in: Proceedings of the 2Nd International Conference on Networking, Information Systems & Security, NISS19, ACM, New York, NY, USA, 2019, pp. 26:1–26:8.
[3] S. Schlögl, J. Buricic, M. Pycha, Wearables in the wild: advocating real-life user studies, in: Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct, MobileHCI '15, ACM, New York, NY, USA, 2015, pp. 966–969.
[4] O. Sofrygin, Z. Zhu, J.A. Schmittdiel, A.S. Adams, R.W. Grant, M.J. van der Laan, R. Neugebauer, Targeted learning with daily EHR data, Stat. Med. 38 (16) (2019) 3073–3090.
[5] K.I. Mohammed, A.A. Zaidan, B.B. Zaidan, O.S. Albahri, M.A. Alsalem, A.S. Albahri, A. Hadi, M. Hashim, Real-time remote-health monitoring systems: a review on patients prioritisation for multiple-chronic diseases, taxonomy analysis, concerns and solution procedure, J. Med. Syst. 43 (7) (2019) 223.
[6] D.V. Dimitrov, Medical internet of things and big data in healthcare, Healthcare Inf. Res. 22 (3) (2016) 156–163.
[7] J. Yan, Y. Ma, L. Wang, K.-K.R. Choo, W. Jie, A cloud-based remote sensing data production system, Future Gener. Comput. Syst. 86 (2018) 1154–1166.
[8] M.H. Rahman, T.J. Tumpa, S.M. Ali, S.K. Paul, A grey approach to predicting healthcare performance, Measurement 134 (2019) 307–325.
[9] Y. Yang, X. Zheng, W. Guo, X. Liu, V. Chang, Privacy-preserving smart IoT-based healthcare big data storage and self-adaptive access control system, Inf. Sci. 479 (2019) 567–592.
[10] T. McGhin, K.-K.R. Choo, C.Z. Liu, D. He, Blockchain in healthcare applications: Research challenges and opportunities, J. Network Comput. Appl..
[11] M. Barua, X. Liang, R. Lu, X. Shen, ESPAC: enabling security and patient-centric access control for eHealth in cloud computing, Int. J. Secure. Network 6 (2–3) (2011) 67–76.
[12] H. Zhang, J. Yu, C. Tian, P. Zhao, G. Xu, J. Lin, Cloud storage for electronic health records based on secret sharing with verifiable reconstruction outsourcing, IEEE Access 6 (2018) 40713–40722.
[13] M. Staffa, L. Sgaglione, G. Mazzeo, L. Coppolino, S. D'Antonio, L. Romano, E. Gelenbe, O. Stan, S. Carpov, E. Grivas, et al., An OpenNCP-based solution for secure eHealth data exchange, J. Network Comput. Appl. 116 (2018) 65–85.
[14] M.M. Fouad, N.E. Oweis, T. Gaber, M. Ahmed, V. Snasel, Data mining and fusion techniques for WSNS as a source of the big data, Proc. Comput. Sci. 65 (2015) 778–786.
[15] P. Wayner, Compression Algorithms for Real Programmers, Elsevier, 1999.
[16] J. Uthayakumar, T. Vengattaraman, P. Dhavachelvan, A survey on data compression techniques: from the perspective of data quality, coding schemes, data type and applications, J. King Saud Univ.-Comput. Inf. Sci..
[17] R.H. Coding, Information theory, Prentice Hall..
[18] D.A. Huffman, A method for the construction of minimum-redundancy codes, Proc. IRE 40 (9) (1952) 1098–1101.
[19] K. Sayood, Introduction to Data Compression, Morgan Kaufmann, 2017.
[20] C.M. Sadler, M. Martonosi, Data compression algorithms for energy-constrained devices in delay tolerant networks, in: Proceedings of the 4th International Conference on Embedded Networked Sensor Systems, ACM, 2006, pp. 265–278.
[21] K. Hossain, S. Roy, A data compression and storage optimization framework for IoT sensor data in cloud storage, in: 21st International Conference of Computer and Information Technology (ICCIT), IEEE, 2018, pp. 1–6.
[22] W.M. Ismael, M. Gao, A.A. Al-Shargabi, A. Zahary, An in-networking double-layered data reduction for internet of things (IoT), Sensors 19 (4) (2019) 795.
[23] W. Romine, T. Banerjee, G. Goodman, Toward sensor-based sleep monitoring with electrodermal activity measures, Sensors 19 (6) (2019) 1417.
[24] S. Aria, Y. Elfarri, M. Elvegård, A. Gottfridsson, H. Grønaas, S. Harang, A. Jansen, T. Madland, I. Martins, M. Olstad, et al., Measuring blood pulse wave velocity with bioimpedance in different age groups, Sensors 19 (4) (2019) 850.
[25] K.-H. Huang, F. Tan, T.-D. Wang, Y.-J. Yang, A highly sensitive pressure-sensing array for blood pressure estimation assisted by machine-learning techniques, Sensors 19 (4) (2019) 848.
[26] C. Steinberg, F. Philippon, M. Sanchez, P. Fortier-Poisson, G. O'Hara, F. Molin, J.-F. Sarrazin, I. Nault, L. Blier, K. Roy, et al., A novel wearable device for continuous ambulatory ecg recording: proof of concept and assessment of signal quality, Biosensors 9 (1) (2019) 17.
[27] Y. Kajiwara, T. Shimauchi, H. Kimura, Predicting emotion and engagement of workers in order picking based on behavior and pulse waves acquired by wearable devices, Sensors 19 (1) (2019) 165.
[28] A. Wahrenberg, P.K. Magnusson, A. Discacciati, L. Ljung, T. Jernberg, M. Frick, R. Linder, P. Svensson, Family history of coronary artery disease is associated with acute coronary syndrome in 28,188 chest pain patients, Eur. Heart J.: Acute Cardiovascular Care 0 (0) (0) 2048872619853521..
[29] M. Fan, J. Lv, C. Yu, Y. Guo, Z. Bian, S. Yang, L. Yang, Y. Chen, F. Li, Y. Zhai, et al., Family history, tobacco smoking, and risk of ischemic stroke, J. Stroke 21 (2) (2019) 175.
[30] N. Cohen, R.Y. Brzezinski, M. Ehrenwald, I. Shapira, D. Zeltser, S. Berliner, S. Shenhar-Tsarfaty, A. Milwidsky, O. Rogowski, Familial history of heart disease and increased risk for elevated troponin in apparently healthy individuals, Clin. Cardiol..
[31] A.C. Egbe, C.S. Rihal, A. Thomas, A. Boler, N. Mehra, K. Andersen, S. Kothapalli, N.W. Taggart, H.M. Connolly, Coronary artery disease in adults with coarctation of aorta: Incidence, risk factors, and outcomes, J. Am. Heart Assoc. 8 (12) (2019) e012056.