Research Article

A Proposed Integration Architecture for University Research Data Repository to Support University and University Hospital on Medical Digital Image Management and Analytics using Hadoop

Iskandar Ishak[1], Fatimah Sidi[2], Rusli Abdullah[3], Yusmadi Yah[4], Azizi Sabron[5], Shahril Iskandar Amir[6], Saiful Ramadzan Hairani[7], Ahmad Sobri Muda[8], Anas Tharek[9]

[1,2]Department of Computer Science, Faculty of Computer Science and Information Technology,
Universiti Putra Malaysia
[3,4]Department of Software Engineering and Information System,
Faculty of Computer Science and Information Technology, Universiti Putra Malaysia
[5,6,7]Infocomm Development Centre, Universiti Putra Malaysia
[8,9]Department of Radiology, Teaching Hospital, Universiti Putra Malaysia
iskandar_i@upm.edu.my[1], fatimah@upm.edu.my*[2], rusli@upm.edu.my[3], yusmadi@upm.edu.my[4], azizisabron@upm.edu.my[5], shahril.amir@upm.edu.my[6], ramadzan@upm.edu.my[7], asobri@upm.edu.my[8], anastharek@upm.edu.my[9]

Article History: Received: 10 November 2020; Revised: 12 January 2021; Accepted: 27 January 2021; Published online: 05 April 2021

Abstract: Big Data has been used in university and hospital due to its enormous potential in managing large volume and many types of data. However, university that also has hospitals may need to integrate their data repository to have a single site access for easier system administration and management. The needs of image analytics for both researchers in the university and physicians in the university hospital demand the need of Big Data platform such as Hadoop framework. Based on the literatures, there are no papers that describe in detail the integration of big data for university, which include its own teaching hospital. Therefore, this paper focuses on the proposed research data architecture for university and university hospital to support data repository for both with capability of image analytics using Hadoop technology.

Keywords: Big Data, data repository, analytics, integration architecture, medical images

1. Introduction

In recent years, the use of Big Data technology has been touted as the significant component to improve the management of organization. One of the sectors that include the use of Big Data technology is higher education sector. Higher education sector, which falls under the education umbrella, has a slight but obvious difference as it also includes research and innovation. Managing research and innovation provide greater challenges as it involves funds, expertise, research teams, research outputs and also research funder that may include government agencies as well as private agencies. Some universities also have its own hospitals and research and practices of research findings are directly applied to the hospitals. This increases the complexity of the university management as it include the critical services from the university hospital and thus increase the complexity of the IT services such as data analytic and large amount of health data that include digital images.

The inclusion of Big Data capability have improved many aspects in managing the complexity of university management through data analytics by providing better understanding towards the organizational processes through data-driven decision-making (Mjoohl et al, 2019). For university, Big Data is very important to support smart campus initiative by providing powerful computation infrastructure for Learning Management System analysis such as finding correlations across multiple data sources, predicting an entity behavior, or analyzing social networks (Banica et al., 2014). In healthcare domain, Big Data is used as compared to traditional data storage due to the fact that traditional data storage for patients is not scalable enough for the increasing number of patients and applications and Big Data approach have since taken over this role and implemented in hospitals (Belle et al., 2015; Sobhy, El-Sonbaty and Abou Elnasr, 2012). Big data is defined as massive collection of shareable data originating from any kind of private or public digital sources, which represents on its own a source for ongoing discovery, analysis, and Business Intelligence and Forecasting (Banica et al., 2014; Hussain et al., 2020).

Hadoop is a component in Big Data computing that provides a powerful solution in managing and processing large amount of data coming from multiple data sources and in different formats. It is a distributed architectural platform that comprises a name node and many data nodes (Farhan. Z, et al, 2020). In recent years, Hadoop has frequently been used in the field of health services (i) to develop the framework, (ii) to develop medical large data processing systems, and (iii) to analyze large-scale medical images (Erguzen and Erdal, 2018).

However, there are some issues regarding the implementation of Big Data in university and hospitals due to its technically demanding of experts as well as demanding cost in terms of the equipment. Big Data implementation demands high cost in acquiring the infrastructures and software as well as its maintenance. Also, Big Data physicians needs to be able to acquire new knowledge such as data science to allow them to blend in as a person who can use the Big Data system in processing the data and involved in developing or applying machine learning algorithm in creating prediction models and data visualization. This requires cost and time to prepare them to be expert in data science for healthcare and can work with Big Data application.

With the inclusion of teaching hospital within a university in developing countries such as in Malaysia, (Universiti Putra Malaysia and Universiti Kebangsaan Malaysia) the harnessing of Big Data is very important to serve two purposes; to serve the hospital in terms of data analytics to support decision making in patient treatment and to support research activity of the university through research data repository that allowing collaborative research through sharing of research datasets and image analytics. In order to implement Big Data approach to the two medical expertise needs to be equipped with Big Data-related knowledge to build related data models and to interpret all the results produced using Big Data platform.

There are a number of literatures that have implemented hadoop-based approach to support hospital information systems (Yu and Wang, 2012, Hom, 2014; Jung, Kim, Han, and Jeong, 2014; Sahoo, 2014; Zahra et al., 2019; Chen and Fu, 2015). IBM InfoSphere Guardium provides database activity monitoring and auditing capabilities that enable user to integrate Hadoop data protection into existing enterprise data security strategy **(**Hom, 2014). User can configure the system and use InfoSphere Guardium security policies and reports for Hadoop environments. It does not involve wireless sensor network security communication. HDSM is a Hadoop-based distributed sensor node management system, which uses Hadoop MapReduce framework and distributed file system (Jung, Kim, Han, and Jeong, 2014). Each sensor node imitates DVR (digital video recorder) for sensing video data. All sensor nodes are connected to HDSM manager via gigabit ethernet. So HDSM is not suitable to lightweight sensor node and application. Cloudwave platform is proposed to access and query large volumes of electrophysiological signal data using the Hadoop Distributed File System (HDFS) storage module. Cloudwave allows users to search for clinical events using ontology and semantics reasoning (Sahoo, 2014). However, it does not involve biomedical data security communication. Erguzen and Erdal (2018) proposed a Hadoop-based system for healthcare digital imaging. The system managed to improve the medical image compression method that we have been developed before to create a middle layer platform that performs data compression and archiving operations. With this study, a platform using MapReduce programming model on Hadoop has been developed that can be scalable. Based on the literatures, it is imperative that Big Data has already played an important role for hospitals to support healthcare services.

2. Materials and Methods

In this section, the proposed architecture for university data repository that supports both university and university hospital is presented. In order to integrate, a number of steps need to be taken in designing the proposed integration architecture for data repository. In order to design the integrated architecture, requirements of the proposed architecture are developed based on the approach by Eri et al. (2012), Jabar et al. (2014), Al-Kahtani (2016), Hussain et al. (2016); Gheni et al. (2016) and Kirmse et al. (2018). The requirements gathered are on data type, network as well as business functions of both university and university hospital. In terms of data, medical images play an important role to determine patients' condition for example through x-ray images. In such university, these data are used both by as research as well as for health service.

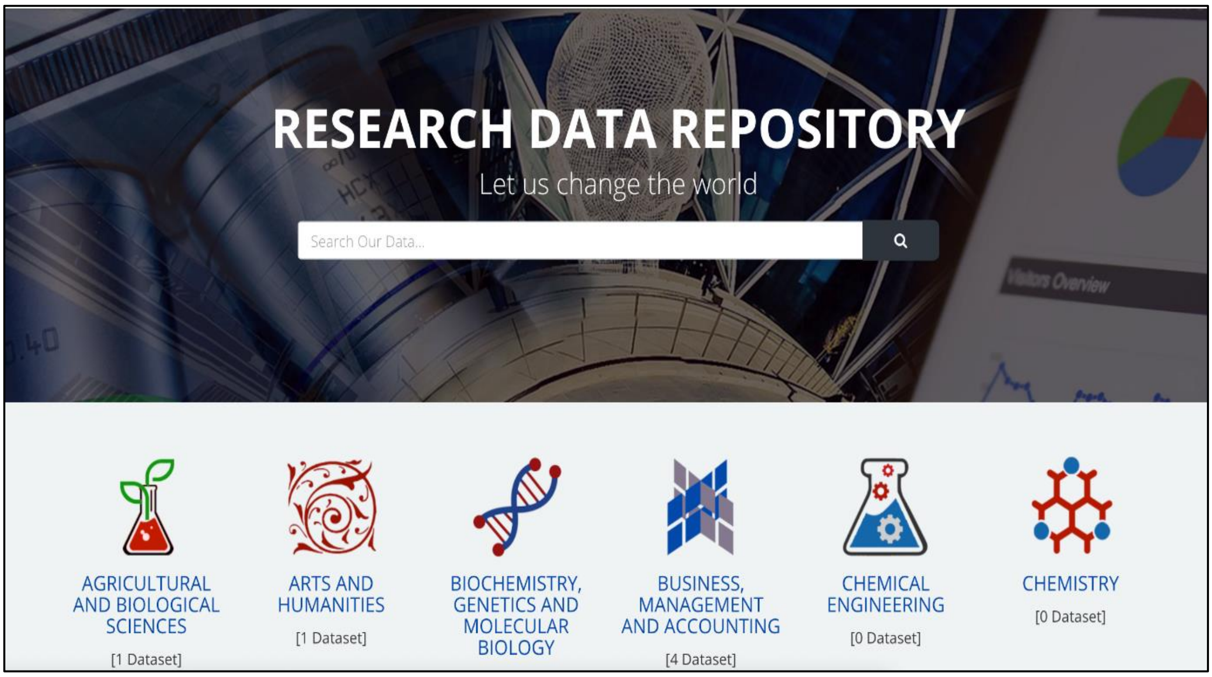
One of the most common image formats for medical health is Digital Imaging and Communications in Medicine (DICOM) format (Genereaux et al., 2018). A number of x-rays and MRIs (Magnetic Resonance Imaging) use DICOM format images. Each DICOM instance is fully usable since it is fully self-describing. Each instance contains a complete set of meta-data, including:

i. Study-level attributes, such as the study unique identifier, the study description, and the patient demographics.
ii. Series-level attributes, such as the series unique identifier, the modality, and the body part imaged.
iii. Instance-level attributes, such as the instance unique identifier, the image resolution, and the table position where the image was acquired.

Each attribute of the metadata has a defined data type and cardinality and is identified by a 32-bit tag—usually shown as two sets of four hexadecimal digits. For example, the tag (0010, 0010) identifies the data attribute containing the patient's name, which has the data type of a person name, and can occur at least once.

Each DICOM instance is identified using a global unique identifier (meaning, no two instances ever use the same identifier, worldwide) (Genereaux et al., 2018).

Universities and university hospitals use DICOM format images to perform analysis as well as research in medical domain. These DICOM images are the main focus of the integration part in which it provides challenges in terms of the formats and volumes. Therefore, ample storage needs to be prepared to store these images to support the data repository integration. In terms of network, multiple considerations are given such as on network resilience, congestion mitigation, performance, scalability and partitioning. Since the proposed integration of data repository serve both for university and hospital university, network latency may not affect big data processing over the Hadoop platform, but any large decreases in network performance may trigger failures in the outcome, since different jobs of these applications need to be executed in parallel in order to assist in accurate analysis.



RESEARCH DATA REPOSITORY

Let us change the world

Search Our Data...

AGRICULTURAL AND BIOLOGICAL SCIENCES [1 Dataset]

ARTS AND HUMANITIES [1 Dataset]

BIOCHEMISTRY, GENETICS AND MOLECULAR BIOLOGY

BUSINESS, MANAGEMENT AND ACCOUNTING [4 Dataset]

CHEMICAL ENGINEERING [0 Dataset]

CHEMISTRY [0 Dataset]

Figure 1. Existing University data repository system interface example

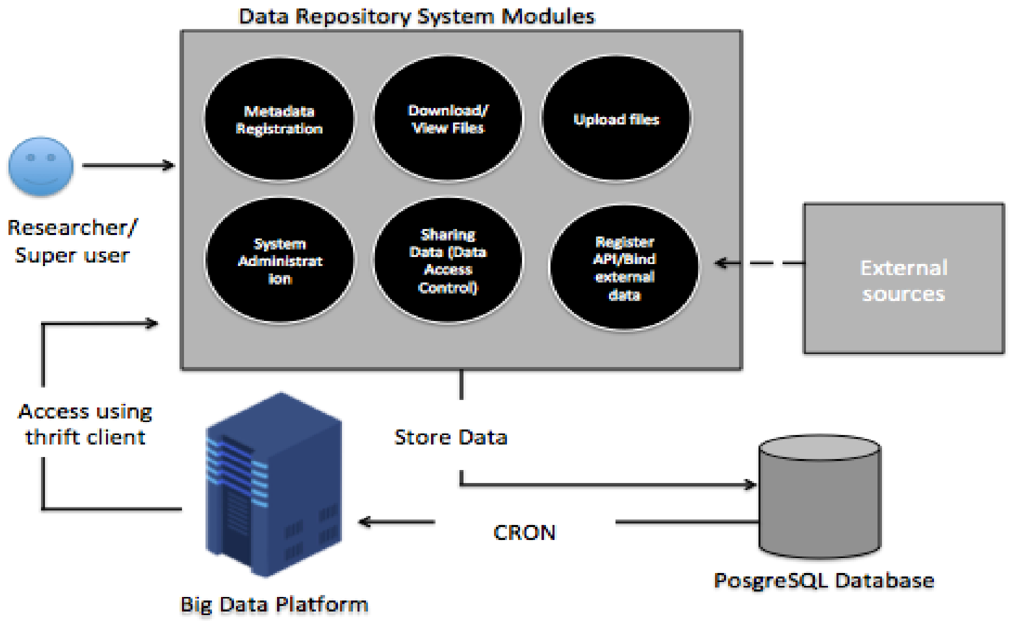


Figure 2. University data repository system overview

3. Results and Discussion

In this section, the integration architecture of the university hospital and university research data repository is discussed. Since the research data repository for university is designed in a way to cater management of research data amongst university researchers, the inclusion of university hospital in a university needs data from this hospital to be on a same platform for easy management of data and to save cost by having single point of access and single site of data repository.

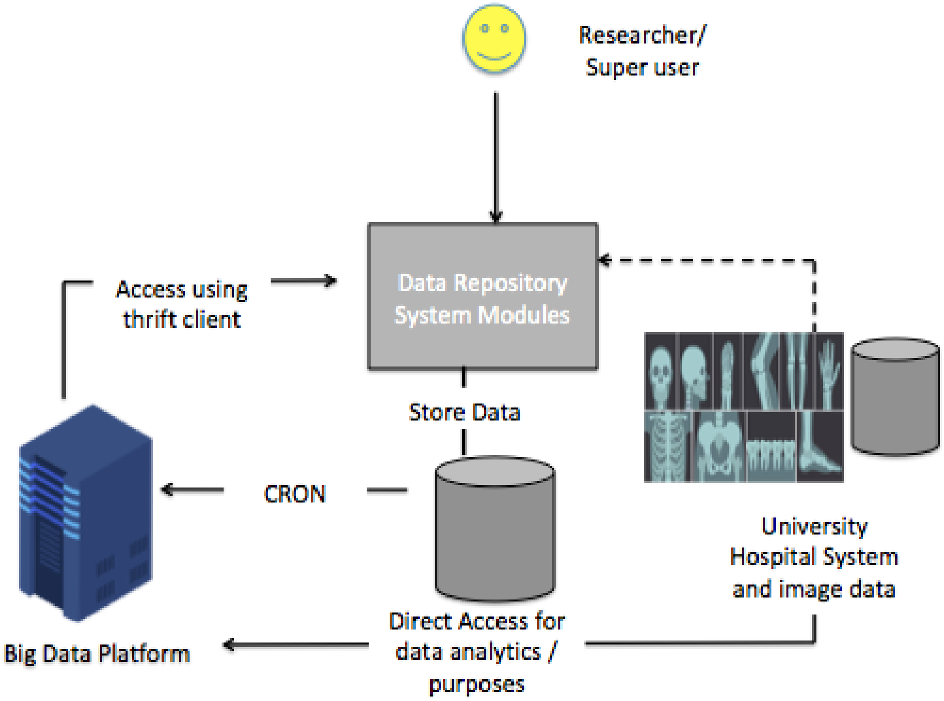


Figure 3. Proposed architecture for integrating university hospital image data source and information system with university data repository system

Figure 3 shows the proposed architecture of integrating university hospital data images system to be integrated with the university data repository system with the support of Hadoop. In order to integrate both systems, the imaging system from the university hospital will be provided an access to the data repository system as an external system or external data source through a specific API. Through this integration, validated image data from the university hospitals devices can be stored in the data repository database as research datasets. The integration will involve two sources of data from the university hospital that is image files from devices in the hospital, and image metadata taken from related patients through the university hospital information system. This will ensure that the images to be stored are supplied with relevant information so that it can be analysed and processed accordingly by other researchers. An efficient approach of processing skyline queries will be adopted in this architecture (Alwan et al., 2016; 2017; Saad el al., 2014; 2016). Allowing users to access the data repository using their analytical application platform direct to the Big Data platform will allow analytical processing to be performed over the images. This is to ensure the usage of Hadoop for data analytic operation over the data images stored in Hadoop. Some data images will also be kept in the data repository system through the data repository system modules with semantic schema matching technique (Hossain et al., 2014), classification technique (Ektefa et al., 2011a), threshold-based similarity (Ektefa et al., 2011b) and classification of knowledge transformation (Sidi et al., 2018) in which modules in Figure 1 will be used to store these data on the data repository.

4. Conclusion

In this paper, integration architecture to integrate university research data repository with university hospital image data is proposed. This is done to support research activities for health data image for university through the university's data repository and to support the university hospital health image diagnose and analysis especially on the machine learning capability. The proposed architecture will provide seamless integration of data repository for health image in terms of data management with image analytics capability data to be used both by university and university hospital. Security is one aspect that is needs to be put into focus as access of the data may involve various access levels. Openness of research data, which is supported by a number of platforms may increase security risk of the system with proposed layout architecture. Cost is another issue of setting up hadoop-based data repository for university and its own teaching hospital as cloud-based platform incur cost for the long run.

5. Acknowledgment

This research is supported by Universiti Putra Malaysia under the Putra Grant Scheme (UPM/700-2/1/GP-IPN/2017/9558200 & UPM/700-2/1/GP- IPN/2017/9558600).

References

1. Al-Kahtani, M. S. 2016. Big Data Networking: Requirements, Architecture and Issues. International Journal of Wireless & Mobile Networks. 8(6), 35-49. 10.5121/ijwmn.2016.8604.
2. Alwan, A. A., Ibrahim, H., Udzir, N. I., & Sidi, F. 2016. An efficient approach for processing skyline queries in incomplete multidimensional database. Arabian Journal for Science and Engineering, 41(8), 2927-2943. doi:10.1007/s13369-016-2048-z.
3. Alwan, A. A., Ibrahim, H., Udzir, N. I., & Sidi, F. 2017. Processing skyline queries in incomplete distributed databases. Journal of Intelligent Information Systems, 48(2), 399-420. doi:10.1007/s10844-016-0419-2.
4. Banica Logica and Radulescu Magdalena, 2015. Using Big Data in the Academic Environment, Procedia Economics and Finance, Volume 33.
5. Belle, Ashwin & Thiagarajan, Raghuram & Soroushmehr, S.M.Reza and Navidi, Fatemeh & Beard, Daniel & Najarian, Kayvan. 2015. Big Data Analytics in Healthcare. BioMed Research International.
6. Chen, H. and Fu, Z. 2015. Hadoop-Based Healthcare Information System Design and
7. Ektefa, M., Jabar, M. A., Sidi, F., Memar, S., Ibrahim, H., & Ramli, A. 2011a. A threshold-based similarity measure for duplicate detection. IEEE Conference on Open Systems, ICOS 2011, 37-41. doi:10.1109/ICOS.2011.6079233.
8. Ektefa, M., Sidi, F., Ibrahim, H., Jabar, M. A., & Memar, S. 2011b. A comparative study in classification techniques for unsupervised record linkage model. Journal of Computer Science, 7(3), 341-347. doi:10.3844/jcssp.2011.341.347.
9. Ergüzen, Atilla and Erdal, Erdal. 2018, An Efficient Middle Layer Platform for Medical Imaging Archives, Journal of Healthcare Engineering, 2018:1-12, https://doi.org/10.1155/2018/3984061.
10. Eri, Z. D., Abdullah, R., Jabar, M. A., & Murad, M. A. A. 2012. Virtual communities model using ontology of group classification for research communities. International Conference on Information Retrieval and Knowledge Management, CAMP'12, 126-130. doi:10.1109/InfRKM.2012.6205019
11. Farhan. Z, Kavipriya.A, Abinaya.C, & M.Ezhilarasan. 2020. Enhanced Image Segmentation Using Convolutional Recurrent Neural Networks. *IIRJET*, V-5, I-3, IT-78 - IT-83.
12. Genereaux, B. W., Dennison, D. K., Kinson, H., Horn, R., Silver, E. L., O'Donnell, K. and Kahn Jr, C. E. 2018. DICOMweb™: Background and Application of the Web Standard for Medical Imaging. Journal of Digit Imaging. 31(3):321-326. doi:10.1007/s10278-018-0073-z.
13. Gheni, A. Y., Jusoh, Y. Y., Jabar, M. A., & Ali, N. M. 2016. Factors affecting global virtual teams' performance in software projects. Journal of Theoretical and Applied Information Technology, 92(1), 90-97.
14. Hom, K. J. 2014. Big Data Security and Auditing with IBM InfoSphere Guardium, https://developer.ibm.com/hadoop/docs/integration/guardi-um/big-data-security-auditing-ibm-infosphere-guardium/.
15. Hossain, J., Sani, N. F. M., Affendey, L. S., Ishak, I., & Kasmiran, K. A. 2014. Semantic schema matching approaches: A review. Journal of Theoretical and Applied Information Technology, 62(1), 139-147.
16. Hussain, A., Manikanthan, S.V., Padmapriya, T., Nagalingam, M. (2020). Genetic algorithm based adaptive offloading for improving IoT device communication efficiency. Wireless Networks, 26 (4), pp. 2329-2338.
17. Hussain, A., Mkpojiogu, E.O.C., Yusof, M.M. (2016). The effect of proposed software products' features on the satisfaction and dissatisfaction of potential customers. AIP Conference Proceedings, 1761, art. no. 020052.
18. Jabar, M. A., Khalefa, M. S., Abdullah, R. H., & Abdullah, S. 2014. Meta-analysis of ontology software development process. International Review on Computers and Software, 9(1), 29-37.

19. Jung, I.-Y., Kim, K.-H., Han, B.-J. and Jeong, C.-S. 2014. Hadoop- based distributed sensor node management system, Interna- tional Journal of Distributed Sensor Networks, vol. 2014, Article ID 601868, 7 pages.
20. Kirmse, A., Kraus, V., Hoffmann, M., and Meisen, T. 2018. An Architecture for Efficient Integration and Harmonization of Heterogeneous, Distributed Data Sources Enabling Big Data Analytics. *ICEIS*.
21. Mjhool, A. Y., Alhilali, A. H., and Al-augby, S. H. 2019. A proposed architecture of big educational data using hadoop at the University of Kufa, International Journal of Electrical and Computer Engineering (IJECE), 9(6), 4970~4978, doi: 10.11591/ijece.v9i6.pp4970-4978.
22. Saad, N. H. M., Ibrahim, H., Alwan, A. A., Sidi, F., & Yaakob, R. 2014. A framework for evaluating skyline query over uncertain autonomous databases. Procedia Computer Science, 29 1546-1556. doi:10.1016/j.procs.2014.05.140.
23. Saad, N. H. M., Ibrahim, H., Sidi, F., Yaakob, R., & Alwan, A. A. 2016. Computing range skyline query on uncertain dimension, doi:10.1007/978-3-319-44406-2_31.
24. Sahoo, S. S. 2014. Biomedical big data for clinical research and patient care: role of semantic computing, in Proceedings of the IEEE International Conference on Semantic Computing (ICSC ’14), pp. 3–5, June.
25. Sidi, F., Ishak, I., & Jabar, M. A. 2018. Malayik: An ontological approach to knowledge transformation in malay unstructured documents. International Journal of Electrical and Computer Engineering, 8(1), 1-10. doi:10.11591/ijece.v8i1.pp1-10.
26. Sobhy, D., El-Sonbaty, Y. and Abou Elnasr, M. 2012. MedCloud: healthcare cloud computing system, in Proceedings of the 7th International Conference for Internet Technology and Secured Transactions (ICITST ’12), pp. 161–166, IEEE, London, UK, December.
27. Wireless Security Communication Implementation. , Mobile Information Systems, Volume 2015, Article ID 852173, 9 pages, http://dx.doi.org/10.1155/2015/852173.
28. Yu, H and Wang, D. 2012. Research and implementation of massive health care data management and analysis based on hadoop, in Proceedings of the 4th International Conference on Computational and Information Sciences (ICCIS ’12), pp. 514–517, IEEE, August.
29. Zahra, F., Hussain, A., bt Mohd, H.(2019). Verification process of metric based usability evaluation model for chronic disease management mobile applications. International Journal of Innovative Technology and Exploring Engineering, 8 (8 S), pp. 475-482.