November-December 2019
ISSN: 0193-4120 Page No. 2410 - 2415

Medical Dataset Preparation Platform Based on a Common Data Model for Machine Learning

Min-Gi Pak[1], Seong-Min Han[2], Chung Sub Lee[3], Chang-Won Jeong[3*], Kwon-Ha Yoon[4]
[1]*Dept of Medical Science, Wonkwang University, Ik-san Jeon-buk, South Korea, pmg0612@wku.ac.kr*
[2]*Dept of Computer Software Engineering, Wonkwang University, Ik-san Jeon-buk, South Korea, zhsks528@wku.ac.kr*
[3]*Medical Convergence Research Center, Wonkwang University, Ik-san Jeon-buk, South Korea, {cslee99, mediblue}@wku.ac.kr (*corresponding author)*
[4]*Dept of Radiology, Wonkwang University School of Medical and Hospital, Ik-san Jeon-buk, South Korea, khy1646@wku.ac.kr*

Article Info
Volume 81
Page Number: 2410 - 2415
Publication Issue:
November-December 2019

Article History
Article Received: *5 March 2019*
Revised: *18 May 2019*
Accepted: *24 September 2019*
Publication: *12 December 2019*

Abstract

In this paper, we investigate the medical imaging standard model and the platform based on the model as an extension of international standard OHDSI OMOP-CDM. To this end, we present a scheme of the medical imaging standard model based on the DICOM meta tag information, focusing on linking with the existing common data model (CDM). We also present the structure and functions of the web-based dataset platform. Finally, the results of executing web service provided by the implemented platform are shown.

Keywords: *Anonymization, Common Data Model, Medical Image, OHDSI OMOP-CDM, Standardization*

1. INTRODUCTION

The medical industry is undergoing transformation due to big data and artificial intelligence (AI) technologies in the 4th industrial revolution. In particular, research on data analysis combining medical image data and artificial intelligence is being actively conducted; however, standardization of medical images (an unstructured data) is insufficient. It is difficult, therefore, to build dataset required for the research. As an alternative, a method for converting data of each institution into CDM was proposed by the Observational Health Data Science and Informatics (OHDSI) to perform various clinical studies [1]-[3]. Medical imaging information is generated in compliance with the international standard DICOM; however, the detailed data formats are different for each institution. Using the collected images directly to clinical research or AI learning research has been, therefore, a major task [4]. To address this problem, we propose a medical imaging standard linked with a standardized CDM by extracting the tag information from a DICOM file containing medical imaging information. Furthermore, a platform that can produce the standardized dataset as well as manage the data by standardizing clinical data is also required. In addition, as the data volume increases, it is difficult to construct dataset necessary for research. Therefore, the function to classify and provide large dataset is required. This will allow large dataset to be applied to research for artificial intelligence. In this regard, this paper presents a web-based platform containing various medical image dataset that can be utilized for standardizing medical image information, clinical research, and artificial intelligence and the provision of dataset according to artificial intelligence learning types.

2. RELATED WORK

2.1 AI – based Medical Image Analysis

AI-based medical image analysis method can be applied to various medical images, such as ultrasound, endoscopy, pathology images, a wide range of internal organs, X-ray, CT, and MRI. In particular, deep learning algorithms and model structures have been introduced to quickly and accurately produce results. For instance, researches on the detection, quantification, and classification of lesions have been actively conducted [5]. Moreover, new AI research is being conducted and new methods are being developed.

2.2 Common Data Model

As hospital information systems have separately distinct structures for different medical information, a single hospital can only conduct a single-center study based on its patients' data. As each patient group in each medical institution has different characteristics, evaluating conclusions for all patients is a difficult task with an individual hospital data. Therefore, clinical studies based on multi-center patient data can provide meaningful results. To implement this solution, changing each hospital's data into the same format, a common data model (CDM), would be required. Types of CDM, which depends on the purpose of the study, include Sentinel CDM (drug

surveillance), OMOP-CDM (implementation and evaluation of clinical research methodology), and PCORnet CDM (patient-centered clinical research network). Although, many kinds of CDMs have been established in Korea, the CDMs that are actively being used at university hospitals are OMOP-CDM. In 2013, it changed to OHDSI Research Network to deal with data standardization, medical device safety monitoring, and comparative effect studies. In addition, various clinical studies have been conducted with the OHDSI open source software, as shown in Table 1 [6].

2.3 Noticeable Medical Outcomes from Partnership-Common Data Model (OMOP-CDM) Big Data Research

Clinical big data research, based on Common Data Model (CDM) through distributed networks, is an effective method for medical institutions to share their disparate databases. Achilles, a platform provided by OHDSI Research Network, enables the user to visualize the CDM data by using table resources. Atlas, a web-based data analysis tool, enables easier statistical analysis, such as cohort construction, trend variable fitting, survival analysis, and relative risk calculation.

Recently researches using this distributed network method are being conducted to predict various clinical trial results. Furthermore, a variety of open source software were developed on the distributed networks [7].

Table 1: OHDSI open source software

No.	Tool Name	Description
1	ATLAS	- Analysis Platform Integrated OHDSI - Allows selection and extraction of cohorts and visualization of extracted cohorts - search for disease probability: disease incidence by cohort, number of drug users - Medication-side relationship analysis codes provided, propensity score matching analysis, regression analysis, etc.
2	ACHILLES	Standardized profiling tools for database attributes and data quality assessment
3	White Rabbit	- Scan the original data before performing ETL and identify the features - Analyze the properties of each table and column in the original data - Data distribution and frequency analysis results from each table are provided as files
4	Rabbit In A Hat	- Enables creation of mapping definitions between source data tables and CDM tables - Manually written mapping definitions define inter-table and inter-column mapping via UI before creation - Leverage analysis information generated as a result of White Rabbit performance
5	ETC	- OHDSI's open source software is available for free in the GitHub repository. - Developed by several Java-based client applications to provide support for ETL activities. - Additional tools for research activities are developed into HTML5, Web-based client applications, and Java-based Web service layers. - Developing a statistical analysis package using R.

2.4 Radiology CDM Study

In recent clinical studies, researches that have employed machine learning techniques on medical images have been actively conducted. Although medical images are stored in compliance with the DICOM international standard, different standards are used for each institution, similar to the clinical data used in the CDM standard. Selection of the optimized clinical protocol for each disease and medical information stored in key medical images should, therefore, be standardized and stored respectively [8]. In this regard, standardization is being conducted in RadLex, which is adopted by RSNA, however, its dissemination is still insufficient. Although machine learning research is being conducted based on medical images from institutions, it has faced difficulties from the stages of data collection [9]-[10]. In addition, the learning process of AI algorithms should be performed with a large amount of image data to produce highly accurate results; however, it is also very difficult to collect cases for the applicable study. We investigated, therefore, the possibility of converting medical images to CDM as we realized the necessity of standardization of medical images, which is expected to not only provide optimized medical images for research purposes but also start the implementation of a multi-center research.

3. PROPOSED SYSTEM

The system structure of the web-based platform, which is able to provide dataset for standardization of medical images, search and generation of dataset, and AI learning, is shown in Figure 1. We have designed a web client based on JavaScript – React User Interface library, the API Server, and Python – Django Rest Framework. In addition, an asynchronous distributed upload method is used in the Nginx Web Server, Message Queue, and Task Worker to handle large file uploads from each institution. By using Message Queue and Task Worker, the system was designed to reliably upload large files

November-December 2019
ISSN: 0193-4120 Page No. 2410 - 2415

collected from multiple institutions and to classify and download dataset according to the respective CDM.

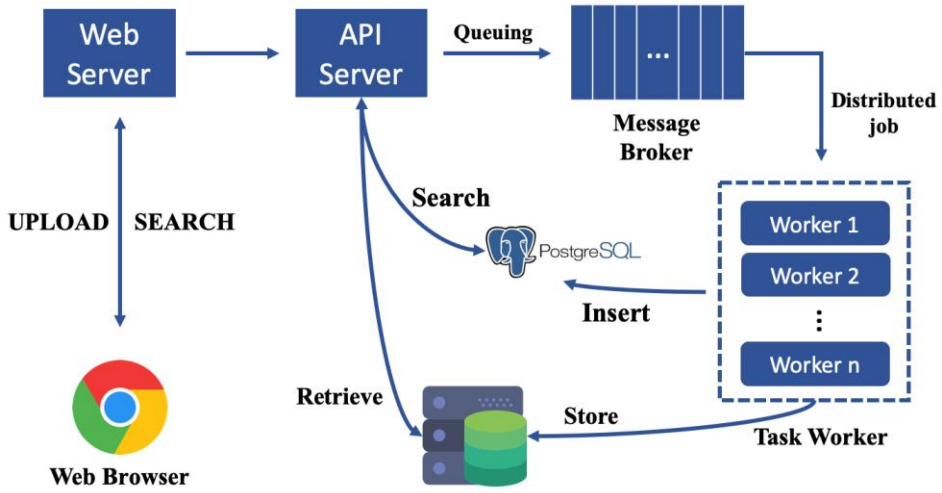

Figure 1: Web Based Medical Dataset Platform

3.1 Database Design for Medical Image Information Standardization

The design of the DB proposed in this paper is shown Figure 2. From a standardization perspective, we designed the radiology occurrence table to store the image by taking information from dataset extracted from DICOM tag information and the radiology image table to store information on the images included in each dataset. To standardize the information of each dataset, the radiology protocol, which includes the conditions of image taking for each hospital; the radiology condition in which the disease and the image are correlated; radiology person position, which evaluates the posture of the patient during the image taking; radiology modality, which determines the modality taken for the images; radiology units to display various units of medical images; radiology device, which shows the equipment to take the image; and radiology hospital information, which exhibits the hospital where the image was taken, are tables designed for storing information required for clinical research. The database will be periodically updated with the required information.

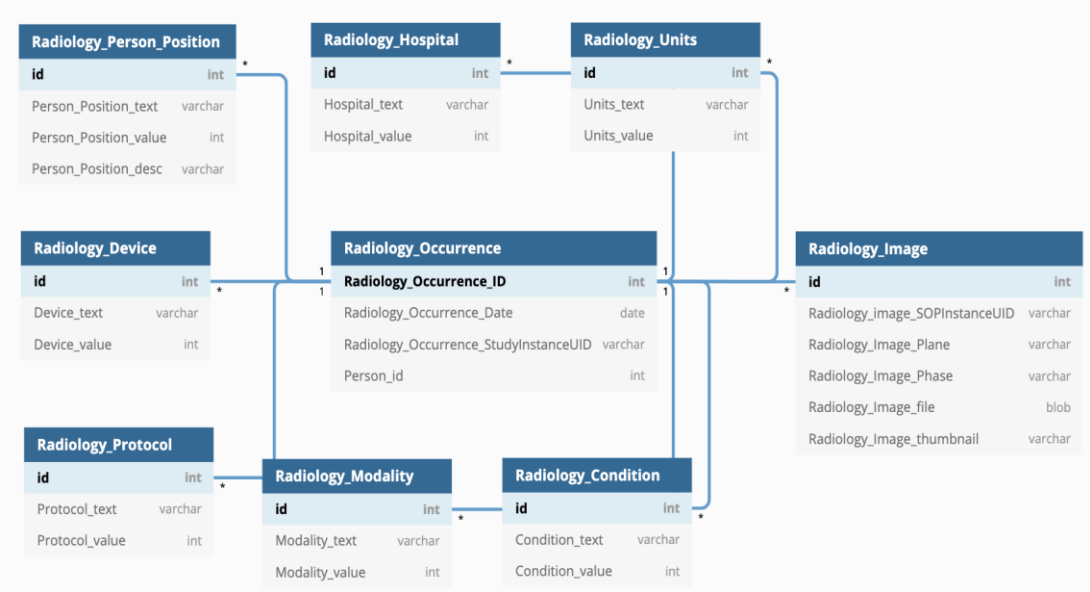

Figure 2: Database Scheme

3.2 DICOM Tag Information Extraction for Standardization

For the standardization of medical image information proposed in this paper, meta tag information (shown in Table 2 for the radiology occurrence and Table 3 for creating the radiology image object) needs to be extracted from the

November-December 2019
ISSN: 0193-4120 Page No. 2410 - 2415

DICOM file. Radiology occurrence consists of tags that distinguish dataset, such as patient information, institutions, and protocols. Meanwhile, radiology image consists of tags carrying medical image information of each DICOM file.

Table 2: DICOM Meta Tag for Radiology Occurrence

DICOM Tag Number	DICOM Tag Name
(0008, 0020)	Study Date
(0008, 0030)	Study Time
(0008, 0033)	Content Time
(0008, 0060)	Modality
(0008, 1010)	Station Name
(0008, 1030)	Protocol Name
(0010, 0010)	Patient Name
(0010, 0020)	Patient ID
(0010, 0040)	Patient Sex
(0010, 1010)	Patient Age
(0010, 4000)	Patient Comments
(0018, 0060)	KVP
(0018, 0087)	Magnetic Field Strength
(0018, 1150)	Exposure Time
(0018, 5101)	View Position
(0020, 000D)	Study Instance UID

Table 3: DICOM Meta Tag for Radiology Image

DICOM Tag Number	DICOM Tag Name
(0008, 0008)	Image Type
(0008, 0018)	SOP Instance UID
(0008, 0031)	Series Time
(0008, 0032)	Acquisition Time
(0008, 103E)	Series Description
(0018, 0050)	Slice Thickness
(0020, 0011)	Series Number
(0020, 0012)	Acquisition Number
(0020, 0013)	Instance Number
(0020, 0037)	Image Orientation (Patient)
(0028, 0010)	Rows
(0028, 0011)	Columns
(0028, 1050)	Window Center

3.3 Anonymization Policy for Privacy Protection

Various DICOM files, collected from multiple institutions, expose the patient's personal information as shown in Figure 3. As a data privacy policy, the anonymization process is conducted to delete personal information related to patient in the DICOM meta tag, as shown in Figure 4. Tags examples are the patient name (0010, 0010), patient ID (0010, 0020), patient sex (0010, 0040), and patient age (0010, 1010).

```
(0010,0010) PN PatientName = Hong Gil Dong
(0010,0020) LO PatientID = 000000
(0010,0030) DA PatientBirthDate = 20190812
(0010,0040) CS PatientSex = F
```

Figure 3: DICOM Meta Tag Information before the anonymization process

```
(0010,0010) PN PatientName =
(0010,0020) LO PatientID =
(0010,0030) DA PatientBirthDate = 20190812
(0010,1000) LO OtherPatientIDs =
```

Figure 4: DICOM Meta Tag Information after the anonymization process

3.4 Search of Dataset

Data collected from multiple institutions and standardized are implemented, as shown in Figure 5, to enable keyword search for dataset based on full text search (FTS) on a web-based platform.

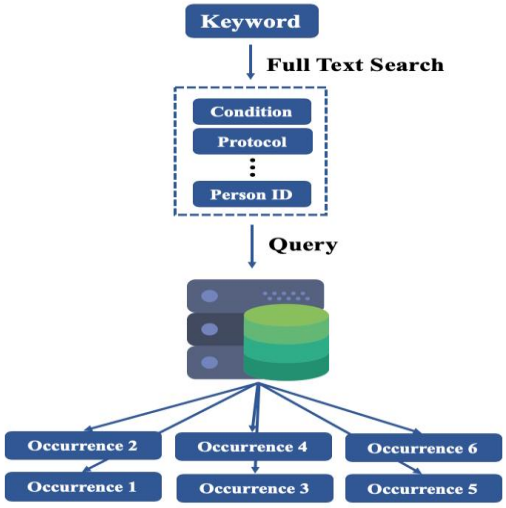

Figure 5: Full Text Search for Dataset

3.5 Download of Dataset

The lack of a system with an automated pipeline to create a dataset for AI learning hinders the implementation of the desired data format. As the importance of a function that automatically classify the dataset according to the user's needs increases with the data size, a custom dataset download functionality has been added, as shown in Figure 6. To provide custom dataset based on the user's needs, the download type can be set to basic or plane phase modes according to the user-specified configuration. In addition, the data can be downloaded in DICOM, PNG, EXCEL, and NIFTI formats, according to the AI learning data type, depending on the purpose of the researcher. By enabling the software to create

November-December 2019
ISSN: 0193-4120 Page No. 2410 - 2415

custom dataset, the cost of classifying data and dataset for AI can be greatly reduced, and the integrity of dataset for machine learning can be enhanced.

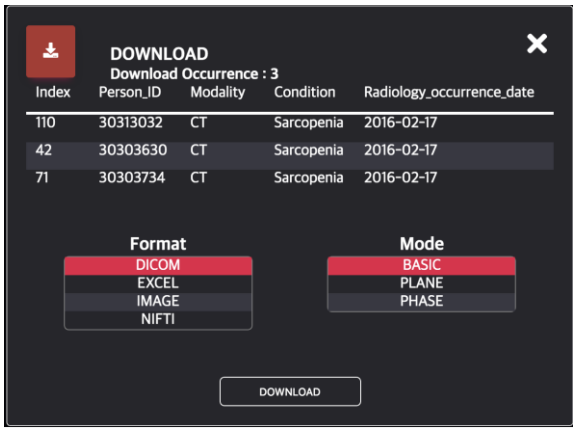

Figure 6: Download Dataset

4. RESULT

4.1 Standardization

The standardization process presented in this paper includes the DICOM tag extraction, occurrence and image object creation, and changing pixel data (7FE0, 0010) to PNG format. Consequently, from measuring the performance of the standardization, a throughput of 100 – 150 per second was obtained as shown in Figure 7. Furthermore, the time required to finish the standardization process is shown in Fig. 8 [11-13].

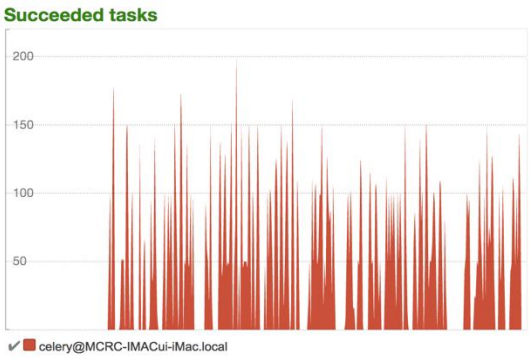

Figure 7: Standardization Conversion Success

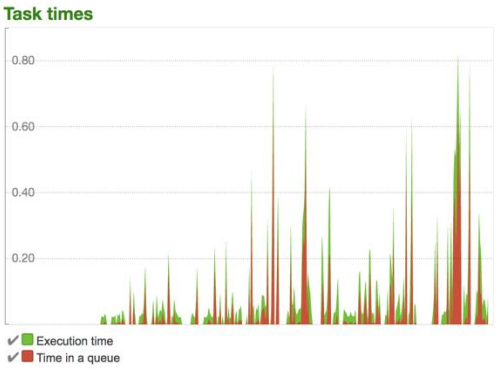

Figure 8: Time to Perform Standardization

4.2 Radiology Occurrence and Image Viewer

Web applications that can perform searches on standardized data are shown in Figure 9 and Figure 10. In the radiology occurrence viewer, searches on the user's desired dataset are enable. Meanwhile, in the image viewer, the image dataset, included in each radiology occurrence, can be visualized.

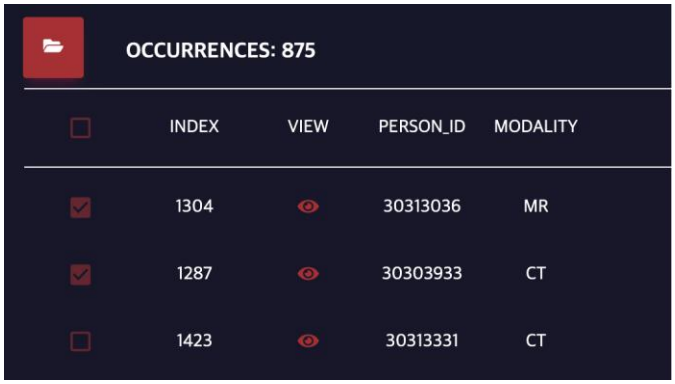

Figure 9: Radiology Occurrence Viewer

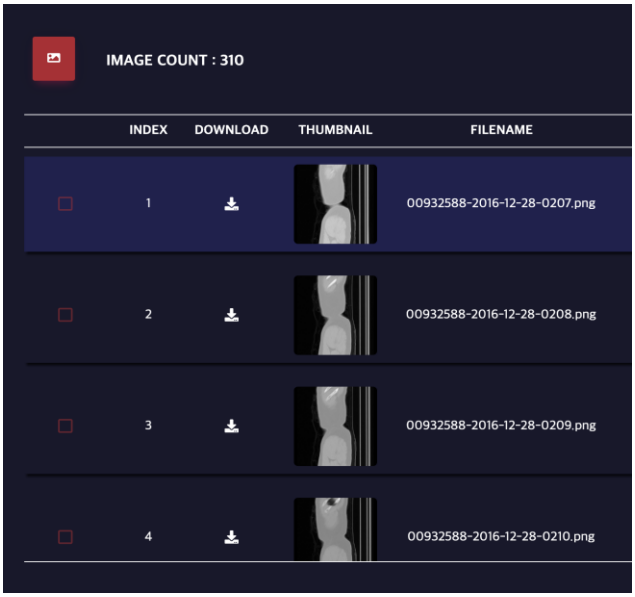

Figure 10: Radiology Image Viewer

5. CONCLUSION

In this paper, we investigated and developed the applicability and necessity of a web-based management system that provides data search and download for standardization of medical image information and clinical research. The system is expected to be applicable to clinical research, especially to the ones using artificial intelligence techniques, as the user may search the desired dataset and customized the platform according to the type of data. In addition, by collecting medical images of each institution, the reliability of rare

diseases' study or clinical research results is expected to improve. In future research, we will develop an image viewer applicable for research by implementing image tools for the images of dataset converted by the standardization process. In addition, we plan to use the accumulated data to employ various AI learning models and to further support analysis tools.

ACKNOWLEDGEMENT

This study was supported by the Korea Health Technology R&D Project through the Korea Health Industry Development Institute(KHIDI), funded by the Ministry of Health & Welfare(HI18C1216) and the Technology Innovation Program (20001234) funded By the Ministry of Trade, Industry & Energy(MOTIE, Korea).

REFERENCES

1. G. Hripcsak, et al. **Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational Researchers**, *Studies in health technology and informatics*, pp. 574-578, 2015.
2. F. FitzHenry, F. S. Resinc, S. L. Robbins, J. Denton, L. Nookala, D. Meeker, L. Ohno-Machado, M. E. Matheny. **Creating a common data model for comparative effectiveness with the observational medical outcomes partnership**, *Appl Clin Inform*, vol. 6, no. 3, pp. 536-547, Jan. 2015. https://doi.org/10.4338/ACI-2014-12-CR-0121
3. E. A. Voss, et al. **Feasibility and utility of applications of the common data model to multiple, disparate observational health databases**, *Journal of the American Medical Informatics Association*, vol. 22, no. 3, pp. 553-564, May. 2015. https://doi.org/10.1093/jamia/ocu023
4. H. Kotadiya, D. Patel. **Review of Medical Image Classification Techniques**, *In Third International Congress on Information and Communication Technology*, vol. 797, pp. 361-369, 2019. https://doi.org/10.1007/978-981-13-1165-9_33
5. G. Litjens, et al. **A Survey on Deep Learning in Medical Image Analysis**, *Medical image analysis*, vol. 42, pp. 60-88, Dec. 2017. https://doi.org/10.1016/j.media.2017.07.005
6. OHDSI, **Observational Health Data Sciences and Informatics Open Source Software**, available at https://ohdis.org/analytic-tools/
7. V. Huser, M. G. Kahn, J. S. Brown, R. Gouripeddi. **Methods for examining data quality in healthcare integrated data repositories**, *In PSB*, pp. 628-633, 2018.
8. W. D. Bidgood Jr, S. C. Horii, F. W. Prior, D. E. Van Syckle. **Understanding and Using DICOM, the data interchange standard for biomedical imaging**, *Journal of the American Medical Informatics Association*, vol. 4, no. 3., pp. 199-212, May. 1997. https://doi.org/10.1136/jamia.1997.0040199
9. A. V. Dalca, K. L. Bounman, W. T. Freeman, N. S. Rost, M. R. Sabuncu, P. Golland. **Medical Image Imputation From Image Collections**, *IEEE transactions on medical imaging*, vol. 38, no. 2, pp. 504-514, Feb. 2019. https://doi.org/10.1109/TMI.2018.2866692
10. J. Zhang, Y. Xie, Q. Wu, Y. Xia. **Medical image classification using synergic deep learning**, *Medical image analysis*, vol. 54, pp. 10-19, May. 2019. https://doi.org/10.1016/j.media.2019.02.010
11. Aymen, R. A., Alhamzah, A., & Bilal, E. (2019). A multi-level study of influence financial knowledge management small and medium enterprises. Polish Journal of Management Studies, 19 (1), 21-31.
12. Yu, D., Ebadi, A.G., Jermsittiparsert, K., Jabarullah, N., Vasiljeva, M.V., & Nojavan, S. (2019) Risk-constrained Stochastic Optimization of a Concentrating Solar Power Plant, IEEE Transactions on Sustainable Energy, https://doi.org/ 10.1109/TSTE.2019.2927735.
13. Hussain, H.I., Kamarudin, F., Thaker, H.M.T. & Salem, M.A. (2019) Artificial Neural Network to Model Managerial Timing Decision: Non-Linear Evidence of Deviation from Target Leverage, International Journal of Computational Intelligence Systems, 12 (2), 1282-1294.