A system for off-line validation of medical data in DICOM archive

SH G Magomedov
MIREA – Russian Technological University, Moscow 119454, Russia
e-mail: magomedov_sh@mirea.ru

Abstract. The problem of consistency of medical data in Hospital Data Management Systems is considered in the context of correctness of medical images itself to minimize possible harm from spurious DICOM files. The approach should be considered as an addition to other securing techniques like watermarking, encrypting, testing conformance with the standard. To achieve amenable accuracy in practice two aspects are taken into account: correctness of periodicity and correctness of image data (time series) itself for considered modality. This paper proposes an architecture of an information system and network filter integrated to it to provide facilities for analysis and alert management. The architecture is designed to perform the analysis of incoming data streams within components working in a clustered manner, providing horizontal scalability and fault tolerance to be BigData-ready.

1. Introduction
Today, healthcare facilities are vast ecosystems made up of a large number of networks devices, equipment, and systems that often require connection to external systems. Medical data is very sensitive to change, it poses a real threat to the health and life of patients. One does not need to have special skills to become familiar with the potential vulnerabilities that a healthcare facility may face. Therefore, the security of medical data must be ensured at every stage of receiving, transferring, processing, storing information, to ensure the confidentiality of patient data, as well as the availability and sustainability of health services at the same time [1]. Based on this, manufacturers of medical systems, as well as organizations that organize those support, need to implement measures to ensure the necessary level of protection against cyber threats, which will increase the level of safety of patients and the infrastructure of the medical institution as a whole. Consider three directions of securing medical data.

1. Ensuring incoming data conformance with DICOM standard. This should be implemented on the server-side – DICOM Server or filtering component running as a frontend for DICOM Server. The datum of interest is IOD (Information Objects Definition). An idea to propose a formal language to express IODs is not new [2]. Modern software like dcm4che [3] supports this validation.

2. Watermarking medical images. The method proposed in [4] based on reversible watermarking technique provides authentication and self-correction by dividing an image into two regions: Region Of Interest (ROI) and Region Of Non-Interest (RONI). Then the ROI is embedded into the RONI, so any change of the image may be detected and could be self-restored back to the original image by extracting the ROI from the RONI. The work [5] proposes a security technique with patient authentication support, information...
confidentiality and integrity based on reversible watermark. To provide integrity checking MD5 hash of the image is computed. Reversibility is achieved with compressed R–S-Vector determined from the image. A watermark providing confidentiality and authentication services is constructed by aggregating the compressed R–S-Vector, the hash value and patient ID. It is encrypted with AES and embedded into medical images.

3. Encryption of DICOM files. The work [6] proposes the following algorithm for providing confidentiality, integrity, and authenticity of the header and pixel data of DICOM images: an encryption and signature creation procedure; a decryption and signature verification procedure. Singla and Singh [7] developed a framework proposing two different approaches to ensure cloud data security: the Extensible Authentication Protocol for authentication and the Rijndael Encryption Algorithm used to encrypt sensitive data. Osama et al. [8] proposes a framework to secure transfer and storage of medical images on the cloud by using hybrid (a combination of symmetric and asymmetric) encryption algorithms. Their scheme consists of separate stages of hashing the DICOM header with SHA-3 and encrypting pixel data with the result of the previous stage using XTEA algorithm.

All the techniques mentioned above are necessary to ensure security of the data transferred to and stored into the archive. Since any software may have vulnerabilities, the problem of authenticity of medical data arises: whether the file in the DICOM archive is legally added and is not fabricated. This paper is organized as follows: firstly the formalization of the problem is presented, where two aspects are discussed: correctness of periodicity and correctness of image data (time series) itself for considered modality; further section describes an architecture of an information system and proposed network filter integrated to it to provide facilities for analysis and alert management; next section contains the analysis of the modeled ECG data stream illustrating the practical value of the approach; finally, the conclusion contains discussion of the pros and cons of the approach proposed.

2. Formalization of the problem
Consider data series with periodically posted samples into an archive. They may be generated not only by immovable equipment, but also wearable devices, like biomedical ECG sensors. Hardware provides data on demand of software, which usually performs periodic requests and sends data to the cloud [9–12]. If periodicity of incoming data stream is violated, it is highly likely to have an attack, if simple loss of network connection is not the case. In the following subsection the periodicity correctness model for incoming data streams is proposed. It allows some variance of the periodicity to handle rare network connection issues.

The second aspect taken into account is legality of the medical data itself [13]. Special methods and algorithms must be developed for any considered modality to perform analysis of the files and estimate possibility of fabrication of medical data with correct timestamp. A method and an algorithm intended for detection of a sharp change in observed patient’s health should be developed for any considered modality [14–17].

3. Periodicity correctness model
Let \( X = \{X_1, X_2, \ldots, X'_1, X'_2, \ldots, X'_n, \ldots, X'_1, X'_2, \ldots, X'_m, \ldots\} \) be the time series of considered modality. It is divided into two parts: reference part (samples denoted as \( X'_i \)) and a part for analysis (samples denoted as \( X^d \)).

Let \( time(X'_i) \) be the function returning timestamp of the sample \( X'_i \) passed as the argument.

Let \( t'_i = time(X'_{i+1}) - time(X'_i), i = 1, \ldots, n - 1 \) be the time interval of two samples in the reference part.
Then average time interval for the reference part \( \frac{1}{n - 1} \sum_{i=1}^{n-1} t_i \) must belong to \( [T' - \Delta'; T' + \Delta'] \), where \( T' \) is the expected period, \( \Delta' \) is allowed variation of the expected period. Both are user-defined external parameters represented as positive datetime values.

Consider \( t_i^j = time(X_i^j) - time(X_i^{j-1}) \), \( j = 1, ..., m-1 \) as the time interval of two samples in the analyzed part. The average value \( \frac{1}{m - 1} \sum_{j=1}^{m-1} t_i^j \) must also belong to \( [T' - \Delta'; T' + \Delta'] \) and \( \forall j \in \{1, 2, ..., m\} \left( k_i \min_{i=1,2,...,n} \left( t_i^j \right) \leq t_i^j \leq k_u \max_{i=1,2,...,n} \left( t_i^j \right) \right) \) for the period of the analyzed part to be considered as valid, where \( k_i \) and \( k_u \) are user-defined external parameters represented by positive real values.

### 4. Architecture of the network filter

The network filter is intended for DICOM traffic interception and off-line analysis of incoming data series of observed patients. Low-level software components rely on libpcap functionality to capture traffic from network interfaces on Linux machines. All the Linux functionality and libpcap facilities are denoted as LinuxKernel component on further diagrams for simplicity.

The class diagram on figure 1 illustrates the anatomy of the filtering and analysis components.

![Class diagram of medical data analyzer](image-url)

**Figure 1.** Class diagram of medical data analyzer
Table 1. DICOMInterceptor class is for analyzing DICOM packets and initiating content checking. It is the main class interacting with LinuxKernel.

| Method              | Description |
|---------------------|-------------|
| inspectPacket       | 1. Extracts timestamp from the data packet.  
                            2. If data is valid DICOM file with timestamp, run the analyzing task asynchronously. |
| passThroughAndLog   | Passes the packet through and logs this event. |
| extractTimestamp    | Extracts timestamp from DICOM file. |
| postAlert           | Posts alert for specified DICOM file. |
| analysisTaskProceduce| 1. Waits for the file with the specified timestamp appears in the DICOM archive.  
                                2. Runs the analysis of this file.  
                                3. If there are violations, posts alerts. |

Table 2. LegalityCheckingProvider class is for configuration management and providing checking functionality for the filtering component.

| Method               | Description |
|----------------------|-------------|
| retrainPatientSeriesModel | Retrain ML model for specified series with specified data interval for training. |
| testPatientSeries    | Returns hashmap with samples IDs as keys and their legality as <boolean,boolean> values. The first boolean in the tuple is period legality, the second boolean in the tuple is content legality. |
| testPatientFile      | Returns period and content legality of the single specified patient file. |
| getSeriesOfPatient   | Return observed series IDs for the patient. |
| releasePatientSeries | Remove the patient's series from the observation list. |
| lookOnPatientSeries  | Add a series of a patient to the observation list. |
| observations         | Hashmap with patient IDs as keys and lists of series IDs as values. |

Table 3. SeriesTester class loads and uses various plugins to check legality of data for different modalities (types of data series). It also uses the instance of PeriodAnalyzer to test periodicity of incoming data.

| Method              | Description |
|---------------------|-------------|
| testSeries          | Returns hashmap with samples IDs as keys and their legality as <boolean,boolean> values. The first boolean in the tuple is period legality, the second boolean in the tuple is content legality. |
| getCheckerForSeries | Get the entry point in the appropriate plugin for specified data series. |
| periodAnalyzer      | An instance of PeriodAnalyzer. |
Table 4. PeriodAnalyzer class is for analyzing the periodicity of data incomings.

| Method          | Description                                                                 |
|-----------------|-----------------------------------------------------------------------------|
| setEthalonPeriod| Sets the value of the period to check conformance of the time series with.   |
| getSuspiciousFiles| Returns hashmap with samples IDs and keys and their legality as boolean values. |

Table 5. ILegalityChecker is a common interface each modality-checking plugin must implement.

| Method          | Description                                                                 |
|-----------------|-----------------------------------------------------------------------------|
| trainModel      | Train ML model for specified series with specified data interval for training. |
| testSeries      | Returns hashmap with samples IDs as keys and their legality as boolean values.  |

MRChecker is an MR modality sample checking plugin implementing ILegalityChecker interface.
RGChecker is an RG modality sample checking plugin implementing ILegalityChecker interface.
ECGChecker is an ECG modality sample checking plugin implementing ILegalityChecker interface.
All the classes except checkers for different modalities (MRChecker, RGChecker, ECGChecker) are implemented in the DICOMAnalyzer component. Checkers are implemented in separate components each. The component diagram on figure 2 illustrates relationships of components in the system and other software it works with, e.g. dcm4che providing DICOM archive facilities.

Figure 2. Component diagram of medical data analyzer
DICOMAnalyzer, MRAnalyzer, ECGAnalyzer, RGAnalyzer (and other modality checkers) are to be packaged as separate artifacts (with the same names, figure 3). All of them should be deployed on the FilteringFrontend node. These nodes should be clustered for high availability. Nginx server is an open-source robust solution for HTTP load balancing to be deployed on HTTPLoadBalancer node. dcm4che, which provides DICOM archive facilities, and PostgreSQL which provides database for it, are to be deployed on separate nodes too.

![Deployment diagram of medical data analyzer](image)

**Figure 3.** Deployment diagram of medical data analyzer

The sequence diagram on figure 4 illustrates the process of inspecting and analyzing DICOM packets within the system on the filtering frontend node.

The process illustrated on the diagram assumes that the model for observing series is already trained (training was initiated manually or automatically via scheduling facilities). Only valid DICOM packets may be analyzed. When the packet arrives, it is inspected and timestamp value is extracted. If the packet being inspected is a valid DICOM packet, the analysis procedure is run asynchronously. It waits for the DICOM file to appear in the DICOM archive and starts its analysis for period and contents legality using PeriodAnalyzer and [Modality]Checker instances. Which checker to choose for contents analysis is decided by getCheckerForSeries of SeriesTester class. In all the cases network activity is logged.
5. Conclusion
An architecture of a BigData-ready system for offline validation of medical data within DICOM archives is proposed. Failover condition is the presence of clustered instances of DICOMAnalyzer components deployed on different physical machines with or without virtualization or containerization layer. Analyzers for different modalities should be deployed with the same redundancy. Their implementation relates to large-scale research, which is a subject of a future work.

References
[1] Magomedov S G 2020 Security analysis of computer networks and applications of the healthcare organizations information processes Cloud of Science 7 pp 685–704
[2] Hewett, Andrew J et al 1997 Conformance testing of DICOM image objects Medical Imaging PACS Design and Evaluation: Engineering and Clinical Issues Vol 3035
[3] https://www.dcm4che.org/ (circulation date 22.08.2020)
[4] Coatrieux, Gouenou et al 2007 Mixed reversible and RONI watermarking for medical image reliability protection 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society
[5] Abd-Eldayem, Mohamed M 2013 A proposed security technique based on watermarking and encryption for digital imaging and communications in medicine Egyptian Informatics Journal 14.1 pp 1–13

Figure 4. Analyzing Incoming DICOM files
[6] Al-Haj A 2015 Providing integrity, authenticity, and confidentiality for header and pixel data of DICOM images Journal of digital imaging 28.2 pp 179–87

[7] Singla, Sanjoli and Jasmeet Singh 2013 Cloud data security using authentication and encryption technique Global Journal of Computer Science and Technology

[8] Dorgham, Osama et al 2018 Enhancing the security of exchanging and storing DICOM medical images on the cloud International Journal of Cloud Applications and Computing 8.1 pp 154–172

[9] Dehlinger J 2011 Challenges and Research Directions MobiCASE

[10] Kassinen O, Harjula E, Koskela T and Ylianttila M 2010 Guidelines for the Implementation of Cross-platform Mobile Middleware International Journal of Software Engineering and Its Applications, of Oulu 4

[11] Petrov AV et al 2013 Application template method to increase mobility of distributed systems for collecting and relaying information from biomedical sensors Journal of radio electronics 5

[12] Lebedev A S, Bolshakov O S and Petrov A V 2013 Designing distributed retransmission system with the mobile clients based on cross-platform software development methods Current problems of science and education 1

[13] Akatkin Yu M, Konyavsky V A, Shishkanov D V and Yasinovskaya E D 2017 Digital health in a digital society. ecosystem and cluster p 220

[14] Komisaruk O, Nikulchev E and Malykh S 2020 Neural network model for artifacts marking in EEG signals Cloud of Science 7

[15] Benssalah M and Rhaskali Y 2020 A Secure DICOM Image Encryption Scheme Based on ECC, Linear Cryptography and Chaos 1st International Conference on Communications, Control Systems and Signal Processing (CCSSP) pp 131–136

[16] Mortajez S et al 2020 A novel chaotic encryption scheme based on efficient secret keys and confusion technique for confidential of DICOM images Informatics in Medicine Unlocked

[17] Shini S G, Thomas T and Chithararanjan K 2012 Cloud based medical image exchange-security challenges. Procedia Engineering