View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by LJMU Research Online

Data Analysis Techniques to Visualise Accesses to Patient Records in Healthcare Infrastructures

Aaron Boddy, William Hurst, Michael Mackay, Abdennour El Rhalibi, Mutinta Mwansa
Department of Computer Science
Liverpool John Moores University
James Parsons Building, Byrom Street
Liverpool, UK, L3 3AF
Email: A.Boddy@2011.ljmu.ac.uk; {W.Hurst, M.I.Mackay, A.Elrhalibi}@ljmu.ac.uk; M.Mwansa@2017.ljmu.ac.uk

***Abstract*— Access to Electronic Patient Record (EPR) data is audited heavily within healthcare infrastructures. However, it is often left untouched in a data silo and only accessed on an ad hoc basis. Users with access to the EPR infrastructure are able to access the data of almost any patient without reprimand. Very Important Patients (VIPs) are an exception, for which the audit logs are regularly monitored. Otherwise, only if an official complaint is logged by a patient are audit logs reviewed. Data behaviour within healthcare infrastructures needs proactive monitoring for malicious, erratic or unusual activity. In addition, external threats, such as phishing or social engineering techniques to acquire a clinician's logon credentials, need to be identified. This paper presents research towards a system which uses data analysis and visualisation techniques deployed in a cloud setting. The system adds to the defence-in-depth of the healthcare infrastructures by understanding patterns of data for profiling users' behaviour to enable the detection and visualisation of anomalous activities. The results demonstrate the potential of visualising accesses to patient records for the situational awareness of patient privacy officers within healthcare infrastructures.**

Keywords-*Electronic Patient Records; Patient Privacy; Information Security; Data Analysis; Visualisation; Healthcare Infrastructures; Cloud Computing.*

I. INTRODUCTION

The health sector consistently constitutes the highest number of reported data security incidents according to the UK Information Commissioners Office [1]. The healthcare sector is an appealing target to attackers due to the high value of patient data on the black market. Patient data is valuable due to the wealth of detailed personal information held within, and the potential to commit identity fraud as a result.

Confidentiality and patient privacy within Electronic Patient Records (EPR) systems is typically managed through an agreed and signed code of practice between the organisation and its users. A healthcare organisation that collects, analyses, publishes or disseminates confidential patient data must commit to ensuring that the data is only accessed by relevant personnel and only when it is appropriate to do so [2]. However, in many cases, measures are not taken to detect and prevent patient privacy violations; any breaches of confidentiality are only brought to light once an investigation is launched, which is often too late. EPR systems are audited; however, the quantity of EPR audit data is significant and a challenge for regular analysis by an Information Security Analyst. Only a big data capable solution using a cloud platform is able to proactively monitor data for patient privacy violations.

Due to the increased need for 24-hour data access the boundaries for healthcare systems is evolving; General Practitioners (GPs) are progressively using Virtual Private Networks (VPN) and 3G connections to remotely access patient data. As a result, the number of access points for hackers is increasing [3] and healthcare organisations should have processes in place to identify data loss and wipe data remotely. Additionally, with many patients having the option of accessing their healthcare data from home PCs and mobile devices, the attack surface is increasing still further.

To detect abnormal data behaviours, visualisation techniques provide both awareness and modelling capabilities for the benefit of computing in critical infrastructures. This allows an analyst to understand data correlations and identify anomalies for investigation through the shape or colour of data patterns [4].

It is unfeasible to set rules for every single user and patient with an EPR as 1) Information Security teams are typically under resourced and 2) even if the resource is available to do this, it often would not provide meaningful information. Rules-based solutions cannot detect violations (such as Advanced Persistent Threats) in these contexts [3].

This paper proposes a cloud-based anomaly detection system which integrates data analysis and visualisation techniques. The system visualises relationships between users and patients in a novel and interactive way. Data analysis algorithms have the capability to explore complex datasets, detect hidden patterns and anomalies within them, and learn from analyst feedback. Visualisation techniques can be used to represent dense information visually, to augment the interpretation process. Cloud technology facilitates the extensive processing power and scalability a system would require to process the EPR audit data in real-time. The results demonstrate the potential for data analysis and visualisation techniques to aid the situational awareness of patient privacy officers within healthcare infrastructures.

The remainder of this paper is as follows. Section II presents a literature review of the background research on patient privacy within EPR systems. Section III outlines our system design. Section IV presents our results and a sample of test data. Section V discusses conclusions and the future work to be done.

II. BACKGROUND

Patient privacy within EPR systems is typically enforced through corrective mechanisms, such as two factor authentication, training and confidentiality agreements [5][6]. Approaches for detecting illegitimate access to EPRs [7] include restricting access control [8], applying patient-user matching algorithms [9], applying scenario-based rule extraction [10], and information gathering from EPR and non-EPR systems using a secure protocol [11]. This is in addition to commonly-used security mechanisms, such as secure networks with firewalls, encrypted devices and messages, strong user passwords, auditing and device timeouts [6].

When there is reason to suspect that unauthorised accesses have occurred, a review of the audit logs is undertaken by a security expert. However, this is inefficient because the overall process requires the information to be collated and reviewed by a security expert. It is also purely retrospective, and the process is only triggered when an anomaly is detected [7]. Therefore, there is a motivation to automate and alleviate the burden of this process [5].

The fundamental limitations in manual audit log reviewing are threefold [7]. 1) The volume of audit records means that audit logs are only practically useful as a supplementary information source to investigate suspected breaches, rather than a tool that can be utilised to proactively find illegitimate accesses; 2) Audit records can only provide data regarding the access itself, and contains no situational or relationship information or knowledge regarding the access. 3) The process is labour-intensive, without guidance of where to look for potential breaches. Subsequently, illegitimate accesses are buried amongst the audits of appropriate accesses.

The challenges facing healthcare security are as follows, a lack of labelled data from previous attacks; constantly evolving attacks and analyst's limited investigative time and budget [12]. Current solutions employ either analyst driven solutions, or unsupervised machine learning solutions but both of these solutions are insufficient on their own. Analyst driven solutions often lead to a high number of false negatives, due to their reliance on human judgement, in addition to delays between attack detection and the implementation of countermeasures [12]. Similarly, unsupervised machine learning solutions are insufficient due to their high number of false positive alarms, which leads to alarm fatigue and distrust by analysts [12]. To address the issue of patient privacy in healthcare infrastructures, hospitals establish a combination of access control solutions, and anomaly detection approaches.

A. Access Control

Healthcare systems commonly employ access control solutions [13]; where once an individual has been authenticated, they are allowed unhindered access inside the perimeter [7]. This means that it is a challenge to impose an access control policy on employees in a healthcare setting due to the dynamic and unpredictable patterns of hospital care [5]. Access control based approaches are limited due to several factors [6], including:

- Unpredictable and dynamic care patterns, including scheduled and unscheduled inpatient, outpatient and emergency department visits
- Varied workflows, with providers requiring access in unexpected areas
- A mobile workforce, with access required at unexpected locations and times
- The collaborative nature of clinical work and teaching environments
- A large number of users with varied job titles and roles
- Users job titles not directly relating to a list of patients whose records it would be appropriate to access

Due to these limitations, access control approaches are insufficient as the sole method of anomaly prevention within EPRs.

B. Detection Approaches

The following section examines several related common detection approaches to anomaly detection in large datasets:

The use of statistical and machine learning techniques have previously been used to detect fraud in financial reporting [14]. They detect fraud in credit card transaction data [15], construct spam email detectors [16] and solve fraud detection problems [17]. Their success is partly due to the fact that machine learning models can be trained on historical data access behaviours to identify future abnormal patterns [5].

In supervised anomaly detection approaches, a set of labelled training instances are provided, typically in the form of *anomaly* and *non-anomaly* [18]. The instances are then trained using a classification model based on their variable features. The resulting models are used to classify new actions. A clearly labelled training dataset, however, is too resource intensive to generate for EPRs, particularly in the context of a dynamic, evolving environment [18]. Supervised machine learning models, such as Support Vector Machines (SVMs), linear regression and logistic regression have been successfully applied to the challenge of detecting illegitimate access within EPR systems [5]–[7].

In unsupervised anomaly detection approaches, the inherent structure, or patterns in a dataset, are utilised in order to determine when a particular instance is sufficiently different [18]. Unsupervised techniques, such as *k*-nearest neighbour anomaly detection, are designed to measure the distances between instances using features such as social structures [19].

Collaborative filtering is a dyadic prediction method, where the task is to predict a label for the interaction of a pair of entities [5]. Within a hospital setting, these entities would be the system user, and the patient record. Collaborative filtering approaches for detecting unauthorised access to EPR data have been successful in recognising the identity of users and patients involved in patient record access [5]. Through the use of explicit and latent features for staff and patients, the following scenarios can be understood to be more likely to be involved in a future violation 1) a patient, whose record has previously been involved in a violation, or 2) a staff member who has performed a violation in the past [5]. In addition to the use of latent features of a dataset to

fingerprint, a user based on historical access data, collaborative filtering can collate data for reliable parameter estimation and create interaction-specific predictions [5].

Genetic algorithms are evolutionary algorithms intended to obtain more accurate solutions as time progresses [20]. The algorithms encode a potential solution to a problem on a chromosome-like data structure, and apply recombination and mutation operators to the structures so as to preserve critical information and improve the utility/objective function [21]. A number of initial solutions are generated (which act as 'parents'). Crossover and mutation operators are applied and new solutions are then generated, with the stronger solutions remaining and the weaker solutions being eliminated [20]. This process continues until the best solution has been found. Genetic algorithms have previously been applied successfully to the domains of credit card fraud, astronomy, optimisation problem and computer science [20]. There is therefore potential for the application of genetic algorithms to the field of anomalous access behaviour detection within healthcare infrastructures.

C. *Existing Approaches*

Monitoring Access Pattern phase 1 (MAP1) demonstrates that statistical and machine learning methods can assist in identifying potentially illegitimate accesses to EPRs [6]. One of the objectives of MAP1 is to identify illegitimate access to EPRs and score each access for appropriateness, so the top scoring cases can be prioritised and investigated by privacy officers. The production of scores indicating suspiciousness of access is preferable to simple rules-based patterns. A training set is created through labelling selected events as either suspicious or appropriate by privacy officers. Logistic Regression (LR) and Support Vector Machine (SVM) models is trained on 10-fold cross-validation sets of 1,291 labelled events [6]. MAP2 (Monitoring Access Pattern phase 2) is an extension of the work of MAP1 and relates to fine-tuning the detection algorithm [7]. MAP2 focuses on the construction of classifiers with appropriate filtering techniques to detect rare events. MAP2 uses a combination of Signature detection, Anomaly detection and Classifier detection, extending the capabilities of the previous MAP1 classifier algorithm. Privacy officers identified 78 illegitimate accesses to the EPR during the study period, and MAP2 identified 75 of those accesses independently, demonstrating that the technique has the capability to facilitate the detection of rare, but important events [7].

Security Information and Event Management (SIEM) systems are distributed systems, which collect and process logs generated by both network hardware and software assets, and perform real-time and centralised event analysis [22]. In doing so, event correlation mechanisms are implemented by the analysis server to identify the occurrence of malicious actions and foresee an attack. However there are a number of issues present in current SIEM solutions. Specifically, current SIEM solutions have processing constraints which limit the effectiveness of discovering violations within the business logic. Additionally, SIEMs cannot process data at the edge of the deployed architecture. This presents limits in addressing data disclosure and privacy issues, a particularly relevant problem within large scale deployments. Finally, no mechanisms are provided to improve the dependability of data storage systems that contain evidence of security breaches and maintain and store the sensitive data of involved parties [22].

III. System Design

As the background demonstrates, there is a clear need for a cloud-based anomaly detection system to ensure patient confidentiality within EPR systems. Our research to date has focused on the development of a system for modelling data flow within healthcare infrastructures [23][24]. The system assists information security officers, within healthcare organisations, to improve the situational awareness of patient data confidentiality risks. The issues of scalability require the system to be deployed on a cloud domain due to the requirements of storage of the EPR audit data and the processing of the machine learning algorithms.

A. *Approach*

The process follows the methodology order presented in Figure 1.

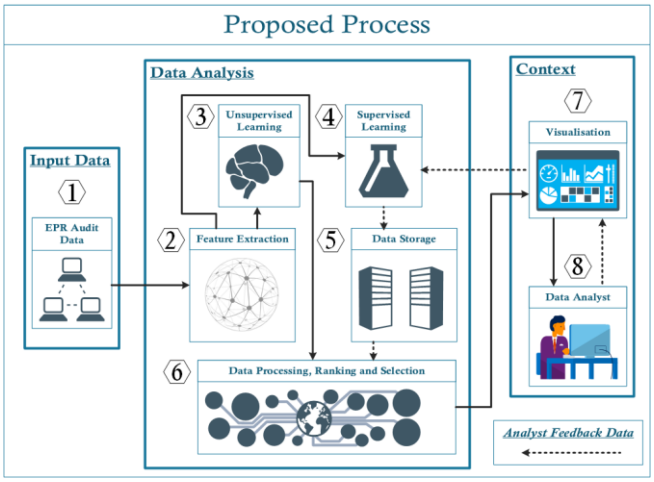

Figure 1. Methodology of the Proposed Process

The system put forward in this paper processes EPR audit data and presents it in such a way as to identify and highlight patterns and potentially anomalous behaviours within it. The data analysis element (components 2-6) is deployed in the cloud.

B. *Process Components*

The system components presented in Figure 1, are explained in this section.

1. *EPR Audit Data*: This audit data is stored by the EPR and captures every interaction with the EPR.
2. *Feature Extraction*: Features of the EPR audit data are extracted for machine learning purposes. Additionally during the testing phase, the data is split into training, test, and validation datasets.
3. *Unsupervised Learning*: Unsupervised machine learning techniques such as k-means clustering are applied in

order to extract unpredictable patterns and anomalies from the data.

4. *Supervised Learning*: Supervised machine learning techniques such as k Nearest Neighbour and Logistic Regression are applied in order to extract knowledge from the data.
5. *Data Storage*: This component stores the data in a database when not in use by the other components. Additionally, datasets of known attack vectors are stored.
6. *Data Processing, Ranking and Selection*: After the data has been pre-processed and subsequently analysed by the machine learning algorithms, the data is ranked and selected based on previous user interaction. This is to ensure the most notable data points are presented to the user first.
7. *Visualisation*: This component generates the visualisation for the user. The component uses the system operators input and calls upon the data stored in the database component, which is then processed and visualised within the generation engine and passed onto the UI Output.
8. *Data Analyst*: The operator interacts with and manipulates the visualisation in order to set their own data parameters. This increases their situational awareness of the data flow within the healthcare infrastructure.

IV. Evaluation

In this section, a case study of the EPR audit data is presented. This rich dataset contains 1,007,727 rows of audit logs of every user and their EPR activity in a UK hospital over a period of 18 months (28-02-16 – 21-08-17). Each User UID is tokenised through isolating the unique entries within the dataset, and assigning each value a unique random number between 1,000,0000 and 9,999,999. This process is done to ensure that once anonymised, each random identifier is still correlated accurately to the original User UID. The range selected is chosen due to information governance concerns regarding viewing the data un-tokenised. Therefore, the tokenizer script is written with the understanding that there are over 1,000,000 rows of audit logs, and therefore the potential for over 1,000,000 unique values. There are 1,515 unique User UIDs and 72,878 unique Patient UIDs. The same process is completed for the Patient UID field.

The dataset consists of the following fields:

- *Date* - The date the patient record is accessed
- *Time* - The time the patient record is accessed
- *Device* - The name of the device the user accessed the patient record
- *User UID (Tokenised)* - A tokenised representation of the User who accessed the patient record
- *Routine* - The routine performed whilst accessing the patient record (is the record updated, is a letter printed, etc.)
- *Patient UID (Tokenised)* - A tokenised representation of the patient record that is accessed
- *Duration* - The number of seconds the user accessed the patient record (this number counts for as long as the record is on the screen, so may not always be an accurate reflection of how long the User is actively interacting with the data)
- *Latest Adm Date* - The date the patient is last admitted to the hospital
- *Latest Dis Date* - The date the patient is last discharged from the hospital

A snapshot of the first 10 rows in the dataset is presented in **Table 1**:

Table 1 - EPR AUDIT SAMPLE DATA

Date	Time	Device	User UID	Routine	Patient UI	Durat	LocationL	Latest Dis
28-02-16	00:00	4Q7QF3J.1	U6199811	PHA.OR	P8290382	54	28-02-16	29-02-16
28-02-16	00:02	27ZKF5J.1	U5053689	ASF	P1591062	13	22-07-08	22-07-08
28-02-16	00:02	C0VLJ5J.2	U2151170	REC REC	P3126528	77	15-02-16	15-02-16
28-02-16	00:02	27ZKF5J.1	U5053689	ASF	P1591062	54	22-07-08	22-07-08
28-02-16	00:04	C0VLJ5J.2	U2151170	REC UK.	P8672400	147	08-02-16	08-02-16
28-02-16	00:04	BEDSIDE_0	U9786800	PHA.OR	P7076283	22	23-01-02	23-01-02
28-02-16	00:04	27ZKF5J.1	U5053689	ASF VH	P2718689	39	28-09-04	28-09-04
28-02-16	00:06	C0VLJ5J.2	U2151170	REC REC	P8526192	165	08-01-16	08-01-16
28-02-16	00:08	9P7QF3J.3	U4425924	NOTE	P5032341	75	25-01-12	25-01-12
28-02-16	00:10	7ZTLJ5J.1	U8857044	PHA.OR	P8705655	42	04-03-07	05-03-07

In Figure 2, a heatmap is presented of the dataset comparing Patient UID to the duration of the patient record access. The graph shows that there is consistent point density of up to 25,487 in the first row of the matrix, indicating that most patient records are only accessed for fewer than 300 seconds (5 minutes). This would represent normal and typical behaviour within the hospital.

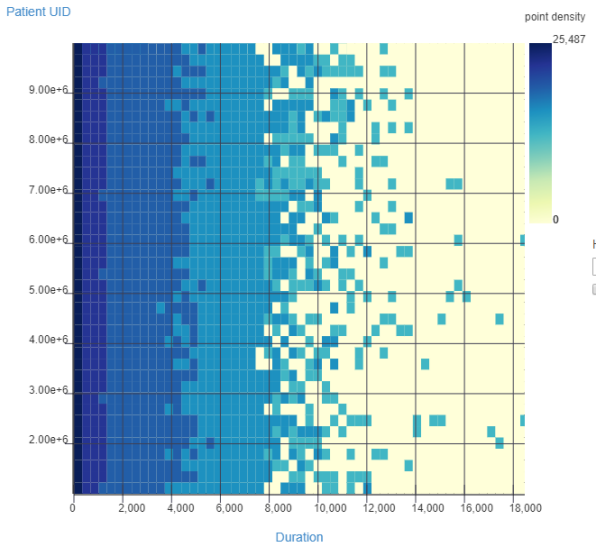

Figure 2 - Heatmap - Patient UID and Duration

Additionally, unsupervised machine learning techniques will be implemented to classify this data in future work as there is limited abnormal data and a lack of labelled training data. Representing the data as a heatmap does highlight some clear anomalies in the data. Notably, certain users are identified spending over 18,000 seconds (over 5 hours) accessing patient records. Extracting features from this data

(such as mean, median, mode and range of duration), will be used to train classifiers to autonomously learn normal and abnormal patterns through supervised learning techniques. This process will occur once the data has been clustered through the use of unsupervised learning algorithms. In combing both unsupervised and supervised machine learning techniques, the system will aid privacy officers in their situational awareness of access to patient records and identify outliers for investigation.

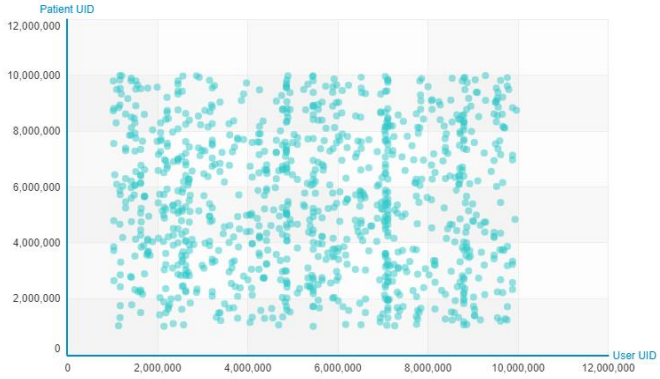

Figure 3 - Scattergraph displaying relationship between Patient UID and User UID

In Figure 3, a scatter graph displays the relationship between Patient UID and User UID. The User UID is displayed on the x-axis whilst the Patient UID is displayed on the y-axis. The graph is a high level representation of when users are accessing patient's details within the full dataset. The graph demonstrates the complexity of the data, as there is no clear structure to the data at face value. The data therefore needs to have meaningful features selected and users of interest identified. In doing so, legitimate accesses can be removed from the visualisation and illegitimate accesses highlighted to a privacy officer.

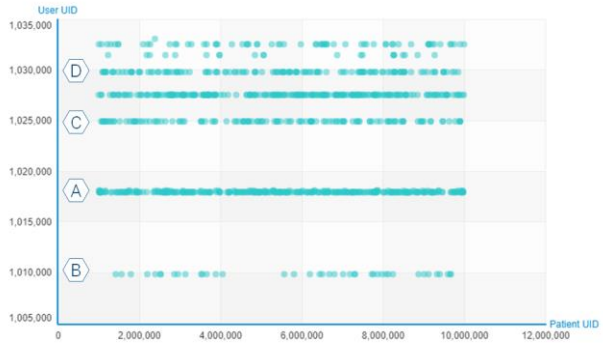

Figure 4 - Profile of 10 Users - Scattergraph of User UID and Patient UID

In Figure 4 a profile of 10 Users is presented as a case study. A scatter graph displaying the relationship between User UID and Patient UID is displayed. The 10 users are a random selection of users, as visualising 10 users represents a reflection of the dataset as a whole. Figure 4 is a representation of the potential of the system to filter the larger dataset of Figure 3 to users of interest. The selected users are the first 10 in the dataset (having been assigned a random number through the tokenisation process). User A is in the upper quartile of accesses to patient records, whereas user B accesses far fewer patients in comparison, indicating a different job role. Staff members C and D have a similar number of patients accessed and therefore likely share a similar job role within the hospital. In this way, roles can be clustered within the data and features extracted. Unusual or erratic spikes in activity would indicate illegitimate activity that would warrant further investigation.

This paper demonstrates the complexity and density of EPR audit data and the capability for a cloud-based anomaly detection system to enhance patient privacy and confidentiality. The system presented in this paper proposes a data analysis and visualisation approach, deployed on a cloud platform, to explore and describe the data in order to aid the situational awareness of privacy officers within healthcare infrastructures. Initial experiments have used a real-world dataset containing 1,007,727 audit records. The system aids privacy officers to find the 'needle in a haystack' of potential patient privacy violations within their vast data infrastructure.

V. Conclusion and Future Work

The results in Section IV display preliminary explorations of the dataset and the potential insights that a cloud-based detection system would provide. The results demonstrate the complexity and density of investigating EPR audit data for anomaly detection. Feature extraction and selection will be used for the benefit of machine learning models. These models are then used to explore the data further, with a particular emphasis on unsupervised learning, such as clustering. Once unsupervised algorithms have been selected, the system will be deployed in the cloud to process transactions and generate anomalies for investigation. This will allow initial patterns within the data to be identified to understand the data and identify illegitimate access to patient records within this real world EPR dataset. Following this process, these data points will be labelled as *anomaly* and *non-anomaly* and trained using a supervised learning model. Additionally, future work will reassign the random values of the User UID and Patient UID to a sequential number sequence for ease of use and visualisation.

Future work will focus on investigating unsupervised machine learning algorithms for clustering the data and identifying outliers within it. Additionally, the cloud infrastructure platform for the system will be explored further to understand the benefits it can provide to an anomaly detection system on real world EPR audit data. Once the system has been refined the system will be automated and tested on real-world data in a UK hospital to detect potential anomalies in real-time. Feedback from analysts during this testing period will inform supervised machine learning algorithms to pick up on patterns and trends within the data and tailor the system to the unique threat landscape of the healthcare infrastructure.

REFERENCES

[1] I. C. Office, Data security incident trends, [Online]. Available: https://ico.org.uk/action-weve-taken/data-security-incident-trends/. [Accessed: 02-Oct-2017], 2017.

[2] H. Chao, C. Hsu, and S. Miaou, "A data-hiding technique with authentication, integration, and confidentiality for electronic patient records", *IEEE Trans. Inf. Technol. Biomed.*, vol. 6, no. 1, pp. 46–53, Mar. 2002.

[3] R. Vargheese, "Dynamic Protection for Critical Health Care Systems Using Cisco CWS: Unleashing the Power of Big Data Analytics", in *2014 Fifth International Conference on Computing for Geospatial Research and Application*, pp. 77–81, 2014.

[4] N. Promrit, A. Minghkwan, S. Simcharoen, and N. Namvong, "Multi-Dimensional Visualization for Network Forensic Analysis", *Int. J. Adv. Comput. Technol.*, vol. 4, no. 5, pp. 222–232, Mar. 2012.

[5] A. K. Menon, X. Jiang, J. Kim, J. Vaidya, and L. Ohno-Machado, "Detecting Inappropriate Access to Electronic Health Records Using Collaborative Filtering"., *Mach. Learn.*, vol. 95, no. 1, pp. 87–101, Apr. 2014.

[6] A. A. Boxwala, J. Kim, J. M. Grillo, and L. Ohno-Machado, "Using statistical and machine learning to help institutions detect suspicious access to electronic health records", *J. Am. Med. Informatics Assoc.*, vol. 18, no. 4, pp. 498–505, Jul. 2011.

[7] J. Kim *et al.*, "Anomaly and signature filtering improve classifier performance for detection of suspicious access to EHRs.", *AMIA ... Annu. Symp. proceedings. AMIA Symp.*, vol. 2011, pp. 723–31, 2011.

[8] A. Ferreira, R. Cruz-Correia, L. Antunes, and D. Chadwick, "Access control: how can it improve patients' healthcare?", *Stud. Health Technol. Inform.*, vol. 127, pp. 65–76, 2007.

[9] J. Salazar-Kish, D. Tate, P. D. Hall, and K. Homa, "Development of CPR security using impact analysis.", *Proceedings. AMIA Symp.*, pp. 749–53, 2000.

[10] P. V Asaro, R. L. Herting, A. C. Roth, and M. R. Barnes, "Effective audit trails--a taxonomy for determination of information requirements"., *Proceedings. AMIA Symp.*, pp. 663–5, 1999.

[11] B. Malin and E. Airoldi, "Confidentiality preserving audits of electronic medical record access"., *Stud. Health Technol. Inform.*, vol. 129, no. Pt 1, pp. 320–4, 2007.

[12] K. Veeramachaneni, I. Arnaldo, V. Korrapati, C. Bassias, and K. Li, "AI2: Training a Big Data Machine to Defend", in *Proceedings - 2nd IEEE International Conference on Big Data Security on Cloud, IEEE BigDataSecurity 2016, 2nd IEEE International Conference on High Performance and Smart Computing, IEEE HPSC 2016 and IEEE International Conference on Intelligent Data and S*, pp. 49–54, 2016.

[13] D. F. Ferraiolo, R. Sandhu, S. Gavrila, D. R. Kuhn, and R. Chandramouli, "Proposed NIST standard for role-based access control", *ACM Trans. Inf. Syst. Secur.*, vol. 4, no. 3, pp. 224–274, Aug. 2001.

[14] T. B. Bell and J. V. Carcello, "A Decision Aid for Assessing the Likelihood of Fraudulent Financial Reporting", *Audit. A J. Pract. Theory*, vol. 19, no. 1, pp. 169–184, Mar. 2000.

[15] T. Guo and G. Li, "Neural data mining for credit card fraud detection", in *2008 International Conference on Machine Learning and Cybernetics*, pp. 3630–3634, 2008.

[16] K. Yoshida *et al.*, "Density-based spam detector", in *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '04*, p. 486, 2004.

[17] J. M. Pérez, J. Muguerza, O. Arbelaitz, I. Gurrutxaga, and J. I. Martín, "Consolidated Tree Classifier Learning in a Car Insurance Fraud Detection Domain with Class Imbalance", Springer, Berlin, Heidelberg, pp. 381–389, 2005.

[18] Y. Chen and B. Malin, "Detection of Anomalous Insiders in Collaborative Environments via Relational Analysis of Access Logs.", *CODASPY Proc. ... ACM Conf. data Appl. Secur. privacy. ACM Conf. Data Appl. Secur. Priv.*, vol. 2011, pp. 63–74, 2011.

[19] "Use of K-Nearest Neighbor classifier for intrusion detection", *Comput. Secur.*, vol. 21, no. 5, pp. 439–448, Oct. 2002.

[20] M. H. Ozçelik, E. Duman, M. Isik, and T. Cevik, "Improving a credit card fraud detection system using genetic algorithm", in *2010 International Conference on Networking and Information Technology*, pp. 436–440, 2010.

[21] D. Whitley, "A genetic algorithm tutorial", *Stat. Comput.*, vol. 4, no. 2, pp. 65–85, Jun. 1994.

[22] C. Di Sarno, V. Formicola, M. Sicuranza, and G. Paragliola, "Addressing security issues of electronic health record systems through enhanced SIEM technology", in *Proceedings - 2013 International Conference on Availability, Reliability and Security, ARES 2013*, pp. 646–653, 2013.

[23] A. Boddy, W. Hurst, M. MacKay, and A. El Rhalibi, "A Study into Detecting Anomalous Behaviours within HealthCare Infrastructures", *9th Int. Conf. Dev. eSystems Eng.*, pp. 111-117, 2016.

[24] A. Boddy, W. Hurst, M. Mackay, and A. El Rhalibi, "A Study into Data Analysis and Visualisation to increase the Cyber-Resilience of Healthcare Infrastructures", *Internet Things Mach. Learn.*, 2017.