A Hybrid Density-Based Outlier Detection Model for Privacy in Electronic Patient Record Systems

Aaron Boddy, William Hurst, Michael Mackay, Abdennour El Rhalibi
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

Abstract—**This research concerns the detection of unauthorised access within hospital networks through the real-time analysis of audit logs. Privacy is a primary concern amongst patients due to the rising adoption of Electronic Patient Record (EPR) systems. There is growing evidence to suggest that patients may withhold information from healthcare providers due to lack of Trust in the security of EPRs. Yet, patient record data must be available to healthcare providers at the point of care. Ensuring privacy and confidentiality of that data is challenging. Roles within healthcare organisations are dynamic and relying on access control is not sufficient. Through proactive monitoring of audit logs, unauthorised accesses can be detected and presented to an analyst for review. Advanced data analytics and visualisation techniques can be used to aid the analysis of big data within EPR audit logs to identify and highlight pertinent data points. Employing a human-in-the-loop model ensures that suspicious activity is appropriately investigated and the data analytics is continuously improving. This paper presents a system that employs a Human-in-the-Loop Machine Learning (HILML) algorithm, in addition to a density-based local outlier detection model. The system is able to detect 145 anomalous behaviours in an unlabelled dataset of 1,007,727 audit logs. This equates to 0.014% of the EPR accesses being labelled as anomalous in a specialist Liverpool (UK) hospital.**

Keywords-Electronic Patient Records; Healthcare Infrastructures; Information Security; Patient Privacy; Visualisation; Machine Learning.

I. INTRODUCTION

Electronic Patient Record systems can support clinical operations within healthcare organisations [1] and improve the safety and efficiency [2] of healthcare delivery, whilst reducing costs [3]. The shift from paper-based to computer-based patient records has improved the availability to patient data without limitations of time or place [4]. Additionally, availability is one of the three principles of information security. However, this shift in making health information accessible and useable by a range of health professionals conflicts with public perception of patient confidentiality and autonomy [5]. To ensure patient privacy in this landscape, there is a requirement for focus on the other two principles of information security, confidentiality and integrity [6]. A continued focus on trustworthy security and privacy mechanisms for health information sharing is necessary due to public concern regarding privacy of EPRs [7].

Patient data privacy, security and confidentiality concerns are validated through numerous reports of patient information being stolen, lost, misplaced, or released without authorisation [8]. Hacking and identity theft is often cited as a cause for concern regarding EPR security, alongside unauthorised access [9]. In particular, patients in the UK are often sceptical regarding the ability of the NHS to safeguard medical information and manage large technological projects, due to failed programs such as NPfIT (National Programme for IT) [10]. This view is particularly held amongst those who had worked in the NHS themselves [9].

The research presented in this paper demonstrates a system that utilises a HILML and density-based outlier detection model, in addition to an advanced visualisation approach, to ensure patient privacy within EPR systems. Density-based outlier detection can identify when a user's behaviour has changed, by comparing behaviours, such as the type of actions being taken and the patients they are viewing. HILML models employ active learning techniques to leverage human expertise and iterate training of the machine learning model. In this way, potentially illegitimate access to patient records can be highlighted and investigated.

The remainder of this paper is as follows. Section II presents background research on patient privacy and audit logs within EPR systems, in addition to a discussion of relevant machine learning and visualisation techniques. Section III outlines the methodology and systematic approach. Section IV discusses our results and a case study. Section V outlines the conclusions and the future work to be done.

II. BACKGROUND

There are 13 features required for security and privacy in EPRs [11]. These include system and application access control, compliance with security requirements, interoperability, integration and sharing, consent and choice mechanism, policies and regulation, applicability and scalability and cryptography techniques. Additionally there are 3 primary focuses of HIPAA (Health Insurance Portability and Accountability Act of 1996) regulations for attaining security in an EPR [12]. 1) Provide sufficiently anonymous release of information for research purposes. 2) Provide appropriate controls to prevent unauthorised people from gaining access to an organisations information systems and control of external communications links and access. 3)

Provide mechanisms for controlled and user-differentiated access to individual patient records.

Traditional methods for defining security policies within organisations are problematic within the context of healthcare organisations due to their reliance on the knowledge of domain experts, or observations of external specialists. Within healthcare the number of security policies are large, defined in an ad hoc manner and can be revised at a moment's notice [13]. A primary feature of patients desire for widespread EPR adoption is transparency, with patients enquiring who has the ability to access their medical records, in addition to determining who has viewed them [14].

Patient Privacy concerns within EPRs is resulting in loss of trust of healthcare providers by patients [15]. This is evidenced by the following studies:

- A 2015 study found that 78.9% of participants would worry about the security of their record if it was part of a national EPR system and 71.3% felt the NHS was unable to guarantee EPR safety [9].
- A 2014 study found that 64.5% of patients expressed concerns regarding data breaches when personal health information was being transferred between healthcare professionals electronically [16].
- A 2012 study found that approximately 60% of respondents believed that the widespread adoption of EPR systems will lead to more personal information being lost or stolen [17].

Additionally, concerns about patient privacy can lead to patients being selective about the information they provide to healthcare providers, or offering incomplete or misleading information [15]. Withholding information due to privacy concerns among patients is evidenced in the following studies:

- A 2014 study by the Office of the National Coordinator for Health Information Technology (ONC) found 7% of patients have withheld information from their healthcare provider due to privacy of security concerns, with this percentage increasing to 33% among those who strongly disagree that there are reasonable protections in place for EPRs [18].
- A 2014 study found that 12.3% of patients withhold information out of concern for a data breach, with the likelihood of withholding information higher among respondents who perceived they had little say regarding how their medical records were used [16].
- A 2011 study by FairWarning in Canada found that 43.2% of patient's withhold information based on privacy concerns, and 31.3% would postpone seeking care for a sensitive medical condition [18]. Additionally, 61.9% reported that if there were serious or repeated breaches at a hospital where they had treatment it would reduce their confidence in the quality of healthcare at the hospital.

A. *Audit Logs*

Without audit mechanisms, EPR systems are vulnerable to undetected misuse, as users could modify or delete health information without their actions being traceable [19]. Audit Logs are usually recorded and stored for the purposes of access management [20]. However, they can also be used for the benefits of monitoring employee behaviour and system failures [21]. Audit Logs should have at least the following elements: 1) Time; 2) Date; 3) Information Accessed and 4) User ID.

Thorough and frequent analysis of audit logs have been shown to discourage abuse [22]. Yet, this analysis often consists of manual audit log review. Motives for a breach of confidentiality within an EPR that may be detected through audit log analysis include: 1) Characteristics of the patient or patient record (such as a VIP). 2) A relationship between the user and the patient. 3) A relationship between the user and another person represented in the patient record (such as a spouse or child).

Indicators of confidentiality breaches can also be separated into positive and negative indicators, where positive indicators are evidence of a potential breach and negative indicators are evidence of expected behaviour, typically based on the established provider role [22]. Probability scoring and an indicator weighting mechanism can aid in prioritising possible breaches for further investigation.

B. *Machine Learning*

There are three primary challenges facing information security that can be addressed through the use of machine learning 1) A lack of labelled data, 2) constantly evolving attacks and 3) limited investigative time and budget [23].

To address these problems, a solution should use analysts' time effectively, detect new and evolving attacks in early stages, reduce response times between detection and prevention and have a low false positive rate [23].

Machine Learning workflows require iterative experimentation in order to attain a desired accuracy. Through analysis of an existing model, the workflow is modified to improve performance with a developer-in-the-loop during the development cycle. Such iterations include adding/removing features, introducing new data sources, changing the machine learning model, adding ensemble averaging to the model and adding a HILML [24].

Ensemble and semi-supervised machine learning techniques involve the combination of both labelled and unlabelled data to change learning behaviour [25]. Through the application of active learning, outlier detection is improved. Due to the lack of labelled data for patient privacy violations within EPRs, semi-supervised learning has been applied for healthcare fraud detection [26].

1) *Ensemble Averaging*

Committee methods operate on the principal that combining the output of a group of machine learning algorithms can achieve a decision function superior to any individual output [27]. Ensemble averaging is a committee method in artificial neural networks that averages the output of a collection of outputs. Ensemble averaging concerns the following two properties of artificial neural networks [28]. Firstly, in a network, bias can be reduced at the cost of increased variance. Secondly, in a group of networks, the variance can be reduced at no bias cost.

The ensemble average can be calculated through the following, where each expert is y_i, and the overall result $\tilde{y}$ can be defined as:

$$\tilde{y}(x;\alpha) = \sum_{j=1}^{p} \alpha_j y_j(x) \tag{1}$$

For a given input, x, the output of the combined model, $\tilde{y}$, is the weighted sum of the corresponding outputs of the component neural networks, y_j, $j = 1,\cdots,p$, and the α_j's are the associated combination-weights [29].

2) Human-in-the-Loop

A HILML model must be able to generalise across use-cases and accept a declarative or semi-declarative specification [30]. Due to the declarative specification, a HILML system should capture a model of a Directed Acyclic Graph (DAG) of intermediate data items. Through a declarative specification, HILML can identify the logical operator for each node in the workflow, such as data preparation or model training [24].

The key advantages of a HILML model are as follows [23]: 1) Overcoming limited analyst bandwidth: An analyst can only feasibly examine less than 1% of the overall event volume. Therefore the use of outlier detection can present the most pertinent events for investigation. 2) Overcoming weaknesses of unsupervised learning: An events rarity, or status as an outlier, does not necessarily constitute maliciousness. Therefore an events score does not capture intent. Using a HILML model can include an analyst's subjective assessment of malicious intent. 3) Actively adapts and synthesises new models: Analyst feedback provides labelled data regularly, creating a positive feedback loop. The more attacks the machine learning model detects, the more feedback it receives from an analyst, which then improves the accuracy of future predictions.

III. Methodology

Our previous work is able to detect 246 anomalous behaviours in an unlabelled dataset of 1,007,727 audit logs. This includes 10 users on the system (0.66%), 122 patient records (0.17%), 102 of routines (0.74%), and 12 devices (0.53%) [31]. This served to highlight specific user or patient IDs for further investigation. The system provides contextual awareness to detect anomalous behaviour within EPR audit activity. The contribution of this research, (the novelty is further outlined in [31] [32]) involves the use of ensemble averaging and a human-in-the-loop model. Our previous work used Local Outlier Factor (LOF)-based data analytics techniques, and visualisation to safeguard EPR data. The process identifies abnormal User IDs, Patient IDs, Device IDs and Routine IDs and highlighting them to an analyst. In this paper, we present a model of amalgamating LOF values into an ensemble averaged LOF value in order to identify individual audit logs for review. Through highlighting a specific audit log for review, the analyst can review the context around the EPR access and determine whether the access was appropriate or inappropriate. By including the HILML model, an active learning approach employs analyst feedback to train the machine learning model.

A. Audit Log Ensemble Averaging

Our previous work does not indicate exactly when a potential inappropriate access has occurred. In order to assign an anomaly score to a specific audit log, rather than a specific ID, the LOF anomaly scores need to calculate the ensemble average. In order to achieve this, a weighted average is applied to each audit log. An additional column is added next to each of the IDs with that IDs associated anomaly score. For every audit log, a weighted average of the four anomaly ID scores is calculated. The calculated ensemble average anomaly score can then be plotted against the Date & Time stamp and visualised to the analyst.

IV. Experiment and Results

A case study of actual EPR audit data is presented as an evaluation of the system methodology.

A sample of EPR data is presented in Table I. This rich dataset contains 1,007,727 rows of audit logs of every user and their EPR activity in a single UK specialist hospital over a period of 18 months (28-02-16 – 21-08-17).

TABLE I
EPR AUDIT SAMPLE DATA

Date & Time	Device ID	User ID	Routine ID	Patient ID	Duration (sec)	Adm Date	Dis Date
16/02/28 00:00	362	865	PHA.ORDS	58991	54	28-02-16	29-02-16
16/02/28 00:02	923	199	REC REC:(DRP) UK.OE	17278	77	15-02-16	15-02-16
16/02/28 00:02	103	677	ASF	4786	13	22-07-08	22-07-08
16/02/28 00:02	103	677	ASF	4786	54	22-07-08	22-07-08
16/02/28 00:04	923	199	PHA.ORDS	62121	147	08-02-16	08-02-16

A large teaching hospital would have approximately 4 times the number of staff and would therefore have a proportional increase in data quantity. The task of navigating this data for anomalous activity is therefore considerable.

The dataset presented consists of the following fields. 1) Date & Time: The date/time the patient record was accessed; 2) Device (Tokenised): The name of the device the patient record was accessed on; *3) User ID (Tokenised)*: A tokenised representation of the User who accessed the patient record; *4) Routine ID*: The routine performed whilst accessing the patient record (was the record updated, was a letter printed etc.); *5) Patient ID (Tokenised)*: A tokenised representation of the patient record that was accessed; *6) Duration*: The number of seconds the patient record is accessed for (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 was actively interacting with the data); *7) Latest Adm Date*: The date the patient is last admitted to the hospital and *8) Latest Dis Date*: The date the patient is last discharged from the hospital.

A. Anomaly Score Ensemble Averaging

A sample of EPR data with a calculated ensemble average LOF anomaly score is presented in Table II.

TABLE II

EPR AUDIT DATA WITH ENSEMBLE AVERAGING APPLIED TO LOF ANOMALY SCORE

Date & Time	Device	Device Anomaly Score	User ID	User Anomaly Score	Routine ID	Routine Anomaly Score	Patient ID	Patient Anomaly Score	Duration (sec)	Adm Date	Dis Date	Ensemble Averaging Anomaly Score
16/09/26 17:02	1284	1.05	435	1.087	ASF SPC CAA MPI	13.339	71272	1.081	853	15/03/2016	15/03/2016	**4.139**
16/11/25 03:39	102	1.084	1487	1.044	ASF SPC CAA MPI	13.339	29971	1.047	901	02/09/2015	02/09/2015	**4.129**
16/08/15 20:56	531	1.161	358	1.052	*** UK.OE MPI PHA.ORDS	11.643	23637	1.066	1180	08/01/2016	08/01/2016	**3.730**
16/11/21 21:46	369	1.088	1021	1.125	SPC SS ASF VH	11.350	41661	1.090	970	24/01/1997	24/01/1997	**3.663**
17/08/09 11:39	1537	1.123	77	1.048	SPC SS ASF VH	11.350	57108	1.030	1041	30/12/2015	30/12/2015	**3.638**
16/11/21 17:38	1052	1.094	809	1.087	SS ZCUS.UK.LETTER ZCUS.UK.SCH MPI	9.701	43065	1.054	723	29/12/1997	29/12/1997	**3.234**
16/04/01 01:12	49	1.151	117	1.031	SS ZCUS.UK.LETTER ZCUS.UK.SCH MPI	9.701	52200	1.028	861	29/09/2015	29/09/2015	**3.228**
16/12/19 20:03	293	1.067	992	1.090	REC REC:(DRP) PHA.MEDS UK.OE	9.538	41375	1.054	2454	28/11/2016	28/11/2016	**3.187**
17/02/07 00:18	566	1.164	262	1.074	ZCUS.UK.LETTER	1.084	35888	9.414	1182	18/04/2013	18/04/2013	**3.184**
16/12/27 18:50	293	1.067	992	1.087	ASF SPC CAA MPI	9.538	46862	1.020	1691	07/12/2016	07/12/2016	**3.179**

The table is ordered by the highest LOF anomaly scores. Within the date range, the most notable audit log occurred on 26th Sep 2016 at 17:02. User #435 accessed Patient #71272 on Device #1284 performing the following Routine combination 'Assessment Forms Maternity Data Care-Area Administrative Data Admissions Demographic Data', with an anomaly score of 4.14. There are 145 audit logs with an anomaly score above 2. Therefore LOF has indicated that 0.014% of the EPR Audit Logs are anomalous.

B. *Visualisation of Results*

In Figure 1, a visualisation of the LOF results of calculated ensemble averaged anomaly scores is displayed for all 1,007,727 audit logs. The x-axis displays the date, and the y-axis displays the calculated ensemble average anomaly score. The LOF anomaly score measures the local deviation of density through determining how isolated the value given by k-nearest neighbours. A value of 1 indicates that an object is comparable to its neighbours and represents an inlier. A value below 1 indicates a dense region, and would therefore also be an inlier. A value significantly above 1 therefore indicates an outlier (anomaly). As all values within the range 0-1 are classified as inliers, values within the range 1-2 were also classified as inliers. Any value above 2 was considered to indicate an outlier for the purposes of this experiment.

The visualisation clearly displays the key logs of interest to be investigated by an Analyst. Hovering over a data point displays the date and time of the EPR access (in yy/mm/dd hh:mm format), in addition to its anomaly score.

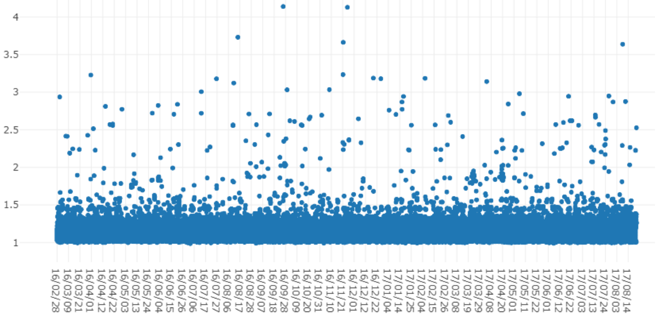

Figure 1 - Visualisation of LOF Results for Ensemble Averaged Anomaly Scores

C. *Discussion*

Our previous work used density based outlier techniques to identify IDs of interest to be investigated by an analyst. Highlighting an ID of interest, such as a user, is useful in some cases, where repeated inappropriate behaviour is evident. However, if the inappropriate behaviour occurred only once then an analyst would need to investigate the user's entire behaviour for patterns, which is not feasible. Figure 1 uses ensemble averaging to identify specific audit logs of interest for an analyst. Highlighting a single audit log as an outlier instead allows an analyst to review it within the context of the other audit logs and determine intent. Additionally, an event being an outlier does not constitute maliciousness. Focusing attention to a single event of interest allows analyst intuition to be leveraged in determining context and intent. Therefore employing a HILML method overcomes the limitations of an unsupervised learning model and incorporates analyst feedback to adapt and use new models. In doing so, analyst attention can be focused to the most pertinent events within the dataset

V. CONCLUSION AND FUTURE WORK

Electronic Patient Record systems represent a fundamental shift for healthcare through increasing availability of healthcare data to providers. However, this ubiquity of data is causing privacy concerns among patients who feel there data is less secure electronically. Current rules-based models are insufficient and most information security incidents are detected by the patient, or staff member, whose privacy has been violated, causing reputational damage to the hospital. Therefore this paper represents research towards a system to ensure confidentiality and privacy of EPR systems. Through the use of machine learning techniques which employ a human-in-the-loop and density-based outlier detection techniques, proactive monitoring of EPR audit logs is achieved. Proactive monitoring allows for inappropriate behaviour to be detected and managed, in addition to prompting a cultural shift among employees to refrain from such behaviour in future.

Future Work will involve gathering feedback and testing the system with information security analysts in a hospital. This will validate the concept on real world non-anonymised data. Using non-anonymised data will allow for other factors to be taken into consideration to determine motivational indicators. Determining a user's role may provide valuable insight. Admin staff and doctors may both have access to the EPR. If an admin staff member is accessing clinical data, this would achieve a higher anomaly score than a doctor, and may indicate a breach. Additionally a patient's characteristics, such as a VIP or a relation to the patient, may provide context to determine whether a patient's confidentiality has been breached. Accounting for additional factors such as these will continuously improve the system.

References

[1] Y. Chen, N. Lorenzi, S. Nyemba, J. S. Schildcrout, and B. Malin, "We work with them? Healthcare workers interpretation of organizational relations mined from electronic health records," *Int. J. Med. Inform.*, vol. 83, no. 7, pp. 495–506, Jul. 2014.

[2] C. Chen, T. Garrido, D. Chock, G. Okawa, and L. Liang, "The Kaiser Permanente electronic health record: Transforming and streamlining modalities of care," *Health Aff.*, vol. 28, no. 2, pp. 323–333, Mar. 2009.

[3] J. A. Zlabek, J. W. Wickus, and M. A. Mathiason, "Early cost and safety benefits of an inpatient electronic health record," *J. Am. Med. Informatics Assoc.*, vol. 18, no. 2, pp. 169–172, Mar. 2011.

[4] N. USENIX Association., D. ACM SIGMOBILE., M. ACM Digital Library., and S. Moulton, "A sensor-based, web service-enabled, emergency medical response system," in *Proceedings of the 2005 workshop on End-to-end, sense-and-respond systems, applications and services*, 2005, p. 48.

[5] J. Sheather and S. Brannan, "Patient confidentiality in a time of care.data," *BMJ (Online)*, vol. 347, British Medical Journal Publishing Group, p. f7042, 27-Nov-2013.

[6] S. A. Hameed, H. Yuchoh, and W. F. Al-Khateeb, "A model for ensuring data confidentiality: In healthcare and medical emergency," in *2011 4th International Conference on Mechatronics: Integrated Engineering for Industrial and Societal Development, ICOM'11 - Conference Proceedings*, 2011, pp. 1–5.

[7] C. Papoutsi, J. E. Reed, C. Marston, R. Lewis, A. Majeed, and D. Bell, "Patient and public views about the security and privacy of Electronic Health Records (EHRs) in the UK: results from a mixed methods study," *BMC Med. Inform. Decis. Mak.*, vol. 15, no. 1, p. 86, Oct. 2015.

[8] G. Laurie, M. L. Stevens, K. H. Jones, and C. Dobbs, "A Review of Evidence Relating to Harms Resulting from Uses of Health and Biomedical Data," Nuffield Council on Bioethics, Feb. 2014.

[9] C. Papoutsi, J. E. Reed, C. Marston, R. Lewis, A. Majeed, and D. Bell, "Patient and public views about the security and privacy of Electronic Health Records (EHRs) in the UK: results from a mixed methods study," *BMC Med. Inform. Decis. Mak.*, vol. 15, no. 1, p. 86, Oct. 2015.

[10] L. Presser, M. Hruskova, H. Rowbottom, and J. Kancir, "Care data and access to UK health records: patient privacy and public trust," *Technol. Sci.*, pp. 1–31, 2015.

[11] F. . F. Rezaeibagha, K. T. . K. T. Win, and W. . W. Susilo, "A systematic literature review on security and privacy of electronic health record systems : technical perspectives," *Heal. Inf.*, vol. 44, no. 3, pp. 1–16, 2015.

[12] J. Salazar-Kish, D. Tate, P. D. Hall, and K. Homa, "Development of CPR security using impact analysis," *Proc AMIA Symp*, pp. 749–753, 2000.

[13] B. Malin, S. Nyemba, and J. Paulett, "Learning relational policies from electronic health record access logs," *J. Biomed. Inform.*, vol. 44, no. 2, pp. 333–342, Apr. 2011.

[14] R. J. Gallagher, D. Ph, S. Sengupta, G. Hripcsak, R. C. Barrows, and P. D. Clayton, "An Audit Server for Monitoring Usage of Clinical Information Systems," *Proc. AMIA Symp.*, p. 10032, 1997.

[15] T. Glenn and S. Monteith, "Privacy in the Digital World: Medical and Health Data Outside of HIPAA Protections," *Current Psychiatry Reports*, vol. 16, no. 11. 2014.

[16] I. T. Agaku, A. O. Adisa, O. A. Ayo-Yusuf, and G. N. Connolly, "Concern about security and privacy, and perceived control over collection and use of health information are related to withholding of health information from healthcare providers," *J. Am. Med. Informatics Assoc.*, vol. 21, no. 2, pp. 374–378, Mar. 2014.

[17] . The National Partnership for Women & Families, "Making IT Meaningful : How Consumers Value and Trust Health IT," no. February, pp. 1–278, 2012.

[18] N. London and C. December, "Canada : How Privacy Considerations Drive Patient Decisions and Impact Patient Care Outcomes," 2011.

[19] J. King, B. Smith, and L. Williams, "Modifying without a trace: general audit guidelines are inadequate for open-source electronic health record audit mechanisms," *Proc. 2nd ACM SIGHIT Int. Heal. Informatics Symp.*, pp. 305–314, 2012.

[20] S. Nunn, "Managing audit trails," *J. AHIMA*, vol. 80, no. 9, pp. 44–45, Sep. 2009.

[21] R. Cruz-Correia *et al.*, "Analysis of the quality of hospital information systems audit trails," *BMC Med. Inform. Decis. Mak.*, vol. 13, no. 1, p. 84, Aug. 2013.

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

[23] 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*, 2016, pp. 49–54.

[24] D. Xin, L. Ma, J. Liu, S. Macke, S. Song, and A. Parameswaran, "Accelerating Human-in-the-loop Machine Learning: Challenges and Opportunities," 2018.

[25] V. M. Gélvez-Ordoñez, F. Mendoza-Galvis, and J. O. Delgado, "Efecto del tratamiento con ultrasonido sobre algunas propiedades funcionales de la clara de huevo," *Rev. Cient. la Fac. Ciencias Vet. la Univ. del Zulia*, vol. 19, no. 1, pp. 71–76, Jan. 2009.

[26] J. Li, K.-Y. Huang, J. Jin, and J. Shi, "A survey on statistical methods for health care fraud detection," *Health Care Manag. Sci.*, vol. 11, no. 3, pp. 275–287, Sep. 2008.

[27] L. Wei, Y. Yang, R. M. Nishikawa, and Y. Jiang, "A study on several machine-learning methods for classification of malignant and benign clustered microcalcifications," *IEEE Trans. Med. Imaging*, vol. 24, no. 2, pp. 371–380, 2005.

[28] U. Naftaly†, N. Intrator‡, and D. Horn§, "Optimal ensemble averaging of neural networks," *Netw. Comput. Neural Syst.*, vol. 8, no. 3, pp. 283–296, 1997.

[29] S. Hashem, "Optimal linear combinations of neural networks," *Neural Networks*, vol. 10, no. 4, pp. 599–614, Jun. 1997.

[30] A. Ghoting *et al.*, "SystemML: Declarative machine learning on MapReduce," in *Proceedings - International Conference on Data Engineering*, 2011, pp. 231–242.

[31] 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.

[32] 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.*, 2016.