IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

Density-Based Outlier Detection for Safeguarding Electronic Patient Record Systems (January 2019)

Aaron J. Boddy[1], Will Hurst[1], Michael Mackay and Abdennour El Rhalibi

[1]Faculty of Engineering and Technology, Liverpool John Moores University, Liverpool, L3 3AF, UK

Corresponding author: Aaron J. Boddy (e-mail: A.Boddy@2011.ljmu.ac.uk).

ABSTRACT This research concerns the detection of abnormal data usage and unauthorised access in large-scale critical networks, specifically healthcare infrastructures. Hospitals in the UK are now connecting their traditionally isolated equipment on a large scale to Internet-enabled networks to enable remote data access. This step-change makes sensitive data accessible to a broader spectrum of users. The focus of this research is on the safeguarding of Electronic Patient Record (EPR) systems in particular. With over 83% of hospitals adopting EPRs, access to this healthcare data needs to be proactively monitored for malicious activity. Hospitals must maintain patient trust and ensure that the information security principles of Integrity, Availability and Confidentiality are applied to EPR data. Access to EPR is often heavily audited within healthcare infrastructures. However, this data is regularly left untouched in a data silo and only ever accessed on an ad hoc basis. Without proactive monitoring of audit records, data breaches may go undetected. In addition, external threats, such as phishing or social engineering techniques to acquire a clinician's logon credentials, need to be identified. Data behaviour within healthcare infrastructures therefore needs to be proactively monitored for malicious, erratic or unusual activity. This paper presents a system that employs a density-based local outlier detection model. The system is intended to add to the defence-in-depth of healthcare infrastructures. Patterns in EPR data are extracted to profile user behaviour and device interactions in order to detect and visualize anomalous activities. The system is able to detect 144 anomalous behaviours in an unlabelled dataset of 1,007,727 audit logs. This includes 0.66% of the users on the system, 0.17% of patient record accesses, 0.74% of routine accesses, and 0.53% of the devices used in a specialist Liverpool (UK) hospital.

INDEX TERMS Data Analysis, Electronic Patient Records, Healthcare Infrastructures, Information Security, Patient Privacy, Visualisation,

I. INTRODUCTION

The health sector consistently receives the highest number of reported data security incidents [1], as the EPR data within represents some of the most sensitive and valuable data available. At the time of writing this paper, patient privacy within EPR systems is typically enforced through corrective mechanisms, managed through role-based access [2]. However, once a user has been authenticated, they are essentially afforded unhindered access. The wealth of personal information held is intrinsically valuable on the black market, often used for committing identity fraud.

There is also a tendency for organisational complacency within healthcare towards patient privacy violations [3]. Recent attacks, such as the WannaCry campaign [4], have further reduced the levels of public trust in security leading to widespread concern about the health sector's ability to maintain the privacy of patient data. Bell-LaPadula [5], and FairWarning [6], are the staple access control systems employed but are *i)* inflexible, presenting issues when considering the dynamic boundaries of many modern healthcare networks and *ii)* do not compensate for an attacker who has acquired the logon credentials of an approved clinician (e.g. through phishing or social engineering). This has been a challenge for security experts for many years, referred to as a *plain recognition* problem [7], Information Security Officers and IT Managers need to interpret disparate data behaviours to preserve privacy and safeguard EPR data [8]. They constantly balance privacy with a need for more intuitive security solutions. Therefore, confidentiality and patient privacy within EPR systems is typically managed through an agreed and signed code of practice between the organisation and its users [9].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2906503, IEEE Access

Patients need to be assured of three crucial security principles *1)* the data stored is trustworthy and accurate. *2)* Data can be reliably accessed by healthcare professionals when needed. *3)* Only authorised healthcare professionals have access to the data, and only access it when it is appropriate to do so. Issues also surround data being exchanged across multiple countries that have different laws and regulations concerning data traversal, protection requirements, and privacy laws.

The UK, specifically, is a significant contributor to data privacy and cyber security research with the establishment of 14 cyber-security Centres of Excellence from 2011 to 2017 [10], in addition to the formation of the Malvern Cyber Security Cluster in 2011 [11] and the North West Cyber Security Cluster in 2014 [12], as examples. The UK government does invest into cyber-security schemes, such as the £1.9billion investment into the national cyber security strategy, aiming to make the UK one of the safest places in the world to do business [13]. Yet within healthcare infrastructure, privacy and security are still seen as a secondary consideration, though the importance to establish data access regulations is imminent due to the geographical requirements for healthcare data being stored. Compliance with NHS guidelines, the Information Governance Toolkit, internal audit processes and information security standards (e.g. ISO27001 and ISO27002) is an additional concern to adhere to.

The research presented in this paper demonstrates a system that utilises density-based outlier detection techniques and an advanced visualisation approach to safeguard 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. 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 within EPR systems, the complexity of EPR data and the network structures in a typical UK hospital. Section III outlines the methodology and systematic approach. Section IV discusses our results and a case study. Section V outlines our conclusions and the future work to be done.

II. BACKGROUND

Machine learning algorithms observe and learn data patterns and profile users' behaviour, which can then be denoted. Combined with cloud infrastructure and data visualisation, the way large datasets are understood is being transformed, allowing extraction of otherwise unobtainable meaning from vast quantities of information. This is now a proven approach for detecting zero day attacks and uncovering unknown threats [14]. There is a large volume of literature concerning big-data-based privacy-preserving machine learning algorithms. Genetics-based machine learning (GBML) [15], clustering fuzzy rule-based classifiers [16] and Linear Support Vector Machines (SVMs) [17] are examples of the general conventional means of choice for researchers. Further to this, DarkTrace [18], based in the UK, is among the world's most advanced machine learning technologies for cyber defence and an advocate for using AI for safeguarding critical systems. Their Enterprise Immune System demonstrates the effectiveness of switching the security perimeter from an external 'wall' to an internal-facing adaptive model to improve security systems, threat detection and enhances the levels of data privacy.

DarkTrace is testament to the fact that cyber-security techniques are trending towards the use of reactive/proactive systems rather than passive detection in order to deter attacks. Machine learning and data visualisation techniques are the technique of choice for establishing this security evolution. The concept is that security systems should respond to unknown intrusions, much like an organic-immune system.

A. HOSPITAL NETWORKS

Introducing complex machine learning algorithms to interpret patterns of behaviour in hospital networks is a considerable challenge. With healthcare networks, devices (medical, clinical and personal) are connected to global networks for convenient access using platforms, such as HomeLinks. Typically, modern healthcare networks are overly complex systems, with hospitals having their own unique structure. As an example, Figure 1 displays the data connections for the Active Directory Domain Controller (DC), Electronic Prescribing (EP) and Patient Administration System (PAS) at a Liverpool-based hospital.

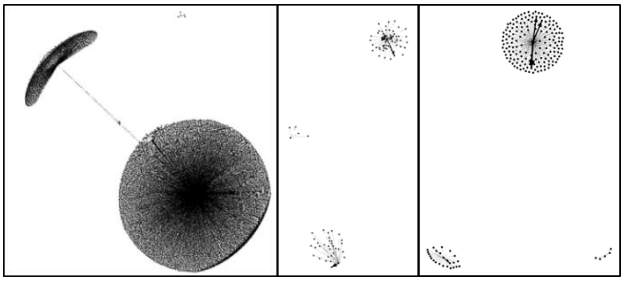

FIGURE 1. Data connections for DC, EP and PAS systems at a Liverpool (UK) specialist hospital depicted by the Yifan Hu algorithm

In Figure 1, a layout algorithm displays the data connections for DC, EP and PAS within a Liverpool Hospital network, demonstrating the complexity of the network data being analysed existing security applications (such as the IDS). In this case the Yifan Hu algorithm [19] is used to model the data connections. This is an approach typically used to present network data movement [20]. However, the data collected is only a snapshot of the network infrastructure using the network statistics (netstat) command-line in order to capture incoming and outgoing

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2906503, IEEE Access

Transmission Control Protocol (TCP) Data. The DC data comprises 590 established connections of 5688 total ports. The EP data comprises18 established connections of 88 total ports. The PAS data comprises 93 established connections of 173 total ports. The level of nodes and connectivity patterns demonstrate the challenge involved for data auditing and uncovering zero-day attacks, network weakness/flaws and emerging threats. The problem of enabling non-expert users to trust that the systems they use are secure when they do not have the technical capability to assess it themselves is not an easy problem to solve.

B. EPR DATA IN HOSPITAL NETWORKS

A sample of EPR data is presented in Table I. The full dataset contains 1,007,727 rows of audit logs.

TABLE I
EPR AUDIT SAMPLE DATA

Date & Time	Device ID	User ID	Routine Description	Patient ID	Duration (sec)	Adm Date	Dis Date
16/02/28 00:00	362	865	Pharmacy Orders	58991	54	28-02-16	29-02-16
16/02/28 00:02	923	199	Recent Clinical Results: (Departmental Reports).View Orders	17278	77	15-02-16	15-02-16
16/02/28 00:02	103	677	Assessment Forms	4786	13	22-07-08	22-07-08
16/02/28 00:02	103	677	Assessment Forms	4786	54	22-07-08	22-07-08
16/02/28 00:04	923	199	Recent Clinical Results. View Orders	62121	147	08-02-16	08-02-16
16/02/28 00:04	103	677	Assessment Forms | Visit History	14067	39	28-09-04	28-09-04
16/02/28 00:04	845	1489	Pharmacy Orders	49304	22	23-01-02	23-01-02
16/02/28 00:10	748	797	Recent Clinical Results: (Departmental Reports)	2166	20	28-01-16	28-01-16

The data used in this research is from a specialist hospital. 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*: 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.

From datasets such as this, usage patterns of the data access can be derived. For example, Figure 2 displays a comparison of the durations of routine activity for each user. The graph is extracted from a dataset of 1,515 unique User IDs and 72,878 unique Patient IDs. The visualisation is constructed using a logarithmic algorithm, outlined in *(6)*.

$$f(x) = log_b(x) \tag{6}$$

Where the base b logarithm of x is equal to $f(x)$. In this sense, a logarithmic heat-map is appropriate as the log scales enable a significant range of coefficients to be displayed.

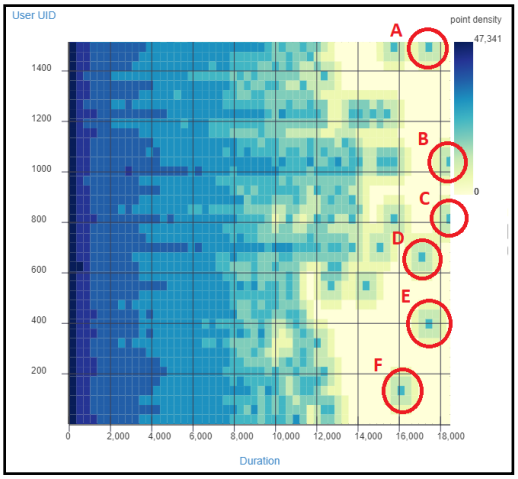

FIGURE 2. Heat-maps (logarithmic) comparing 1million rows of ID data to the duration of the patient record access

Lower-scale values are not compressed down into the congested section of the graph where the unique values would be challenging to identify.

The graph shows a consistent point density of up to 47,341 patient records in the first row of the matrix, indicating that the majority of patient records are only accessed for fewer than 300 seconds (5 minutes). This would represent normal (expected) behaviour within the hospital (as revealed in consultation with the hospital). Whereas, 6 clusters (A-F) require investigation, as they represent users performing routines for over 16,000 seconds (4.44 hours), which would be classed as abnormal (unexpected) behaviour. This observation was identified by the Information Security Manager at the hospital that provided the dataset.

Representing the data as a logarithmic heat-map is a clear approach for identifying data points of interest. However, the density of the dataset prohibits valuable insights from being

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2906503, IEEE Access

derived, and a real-time graph would be inefficient. The quantity of data prohibits all the data points from being visualised. In the following section, data normalisation, feature extraction and machine learning algorithms are applied to the dataset to detect abnormal EPR access. Once the dataset has been administered by these algorithms, visualisation techniques are applied. In doing so, the situational awareness of a patient privacy officer is enhanced.

III. METHODOLOGY

The research is timely due to *i)* a fundamental switch in the technology being used by beneficiaries within health care infrastructures; [21] *ii)* the increased need for 24-hour data access; *iii)* GPs increasingly using Virtual Private Networks (VPN) and 3G connections; *iv)* Most UK hospitals have/are upgrading online EMIS-web, EMIS Health is used by over half of GP practices across the country and EMIS-Web allows hospitals access to primary care, secondary care and mental health data *vi)* more patient remote monitoring is taking security outside hospitals. Such trends reduce security levels and increases access to hospital networks and exposed APIs.

The contribution of this research, (the novelty is further outlined in [22][4]) involves the use of Local Outlier Factor (LOF)-based data analytics techniques, an analyst-in-the-loop and visualisation to safeguard EPR data. The system provides contextual awareness to detect anomalous behaviour within EPR audit activity, using the following multi-stage process:

A. DATA PRE-PROCESSING

In order to provide a meaningful visualisation, the dataset first undertakes a pre-processing phase. The audit data is stored by the EPR and captures every user interaction. Data is extracted into comma separated values format and stored in a database.

1) FEATURE EXTRACTION

Features of the EPR audit data are extracted for the LOF classification process. During the pre-processing stage, a statistical features based approach is implemented [23]. Four measures of central tendency' are calculated through the Frequency, Mean, Median and Mode feature extraction process. Five measures of variability are calculated through the Standard Deviation, Minimum, Maximum, 1st Quartile and 3rd Quartile features. Finally, two measures of position are calculated through the 5th Percentile and 95th Percentile features.

The resulting eleven features are extracted from the dataset for each ID (User, Patient, Device and Routine). Table II displays the features selected, with an accompanying description.

TABLE II
DATASET FEATURE NAMES AND DESCRIPTIONS

Feature Name	Feature Description
Frequency	The number of times the ID featured in the dataset
Mean	The 'average' ID value in the dataset. The sum of the durations for all values for a particular ID, divided by the frequency of that ID.
Mode	The value that appears most in the ID range
Standard Deviation	The measure of the dispersion of the ID range from its mean
Minimum	The data value that is less than or equal to all other values in the ID range
5th Percentile	The value below which the lowest 5% of the data falls
1st Quartile	The median of the lower half of the data set
Median	The value that separates the higher and lower half of the ID range
3rd Quartile	The median of the upper half of the data set
95th Percentile	The value above which the upper 5% of the data falls
Maximum	The data value that is greater than or equal to all other values in the ID range

The mean (μ) is calculated using the equation outlined in (7).

$$\mu = \frac{1}{m}\sum_{i=1}^{m} x_i \tag{7}$$

From this, the standard deviation (σ) is calculated using the equation outlined in (8):

$$\sigma = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(x_i - \mu)^2} \tag{8}$$

The remaining frequency, mode, median, minimum, maximum, 5th percentile, 95th percentile, 1st quartile and 3rd quartile are calculated using sort functions. For example, the mode employs the computation outlined in the following pseudo code (9).

$$X = sort(x);$$
$$indices = find(diff([X; realmax]) > 0);$$
$$[modeL, i] = max\,(diff([0; indices]));$$
$$mode = X(indices(i));$$

(9)

2) DATA CLEANSING

Once the features are extracted, missing or null values (represented by an N/A in the dataset) are replaced with a 0 then the Median value for that feature class. However, within the raw EPR dataset used in this research, no null values are present.

3) FEATURE SCALING

At this stage of the pre-processing, an example of the pre-scaled features dataset is displayed in Table III. In order to ensure the data conforms to a common scale for the classification, the features are scaled.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2906503, IEEE Access

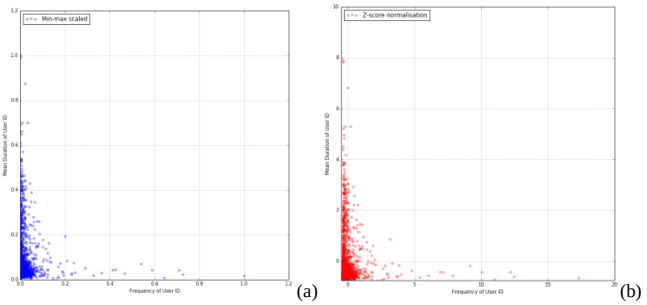

FIGURE 3. (a) Min-Max scaling (b) Z-score normalisation

The Min-Max approach scales the data to a fixed range, between 0-1. The normalised value is obtained using the method outlined in (10) and presented in Figure 3(a).

$$MM(x_{ij}) = \frac{x_{ij} - x_{min}}{x_{max} - x_{min}} \tag{10}$$

Having a bounded range results in lower standard deviations and suppresses the effect of outliers. Decimal scaling normalises by moving the decimal point of values of feature x. Therefore, a $DS(x)$ value is obtained using the method outlined in (11).

$$DS(x_{ij}) = \frac{x_{ij}}{10^c} \tag{11}$$

Where $max[\backslash(DS(X_{ij}))\backslash] < 1$ and c is the smallest integer. The *Z-score normalisation* approach rescales features so that they have the properties of a standard normalisation. The Z-score approach scales the data to a standard normal distribution. The scaled value is obtained using the method outlined in (12) and presented in Figure 3(b).

$$x_{ij} = Z(x_{ij}) = \frac{x_{ij} - \overline{x}_j}{\sigma_j} \tag{12}$$

Where $\overline{x}_j$ and σ_j are the sample mean and standard deviation of the *jth* attribute respectively [24].

B. MACHINE LEARNING

Typically, for the analytic process a machine learning approach is considered. Machine learning emphasises the design of self-monitoring systems, which self-diagnose and self-repair [24]. The technique is commonly used in web search algorithms, spam filters, recommender systems, ad placement, credit scoring, fraud detection, stock trading, drug design and a number of other real-world applications [25].

Machine learning techniques principally consist of combinations of three components, Representation, Evaluation and Optimisation [25] where the data is modelled as a set of variables [26]. The following metrics are employed, a particular task *T*, a performance metric *P*, and a type of experience *E*. If a system reliably improves its performance *P* at task *T*, following experience *E*, then it can be said to have '*learned*' [24].

1) LOCAL OUTLIER FACTOR

The system employs a density-based Local Outlier Factor algorithm. The Local Outlier Factor (LOF) process involves five stages [27]:

i) *k*-distance computation: The Euclidian distance of the *k*-th nearest object from an object **p** is calculated and defined as *k*-distance, where parameter *k* is the number of nearest neighbours.

ii) *k*-nearest neighbour set construction for **p**: Set *k*NN(**p**) is constructed by objects within *k-distance* from **p**.

iii) A *reachability distance* computation for **p**: *Reachability distance* of **p** to an object **o** in *k*NN(**p**) is defined as follows:

$$reach - distk(\boldsymbol{p}, \boldsymbol{o}) = max\{k - distance(\boldsymbol{o}), d(\boldsymbol{p}, \boldsymbol{o})\} \tag{13}$$

where d(**p**,**o**) is Euclidian distance of **p** to **o**.

iv) *lrd* computation for **p**: Local reachability density (*lrd*) of **p**, defined as follows:

$$lrd_k(\boldsymbol{p}) = \frac{k}{\sum_{o \in kNN(\boldsymbol{p})} reach - distk(\boldsymbol{p}, \boldsymbol{o})} \tag{14}$$

v) LOF computation for **p**: LOF of **p** is computed defined as follows:

$$LOF(\boldsymbol{p}) = \frac{\frac{1}{k}\sum_{o \in kNN(\boldsymbol{p})} lrd_k(\boldsymbol{o})}{lrd_k(\boldsymbol{p})} \tag{15}$$

The LOF process exposes anomalous data points by measuring the local deviation. In other words, patterns in data that do not conform to the expected behaviour are revealed. In the case of EPR data, employing a LOF process is effective in that it recognises points, which are outliers from similar/related points in one area of the dataset. Therefore, the algorithm is particularly applicable to a dataset, where multiple job types/roles are present. It considers the relative density of points and can detect data in biased datasets. This means that it is advantageous over proximity-based clustering. LOF employs the relative-density of a coefficient against its neighbours as the indicator of the degree of the object being an outlier [28].

If a global outlier is employed, the detection of irregular behaviours would not be possible without correlating the different hospital roles (as demonstrated in Table I) with each other, adding an extra stage to the detection process – one which might not be possible. This is due to the process that a global outlier detection process undertakes in identifying data points that are far from other points in the dataset.

C. FEATURE TESTING

Given the mean expressed in (7), the scatter matrix is the *m*-by-*m* positive semi-definite matrix. Where T denotes matrix transpose, and multiplication is with regards to the outer product [29], as expressed in (16).

$$S = \sum_{i=1}^{m}(x_i - \mu)(x_i - \mu)^T = \sum_{i=1}^{m}(x_i - \mu) \otimes (x_i - \mu)^T = \left(\sum_{i=1}^{m} x_i x_i^T\right) - m\mu\mu^T \tag{16}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2906503, IEEE Access

The scatter matrix, displayed in Figure 4 (all features have been abbreviated in the graph labels) visualises the relationship between the features to predict the most appropriate for the LOF classification.

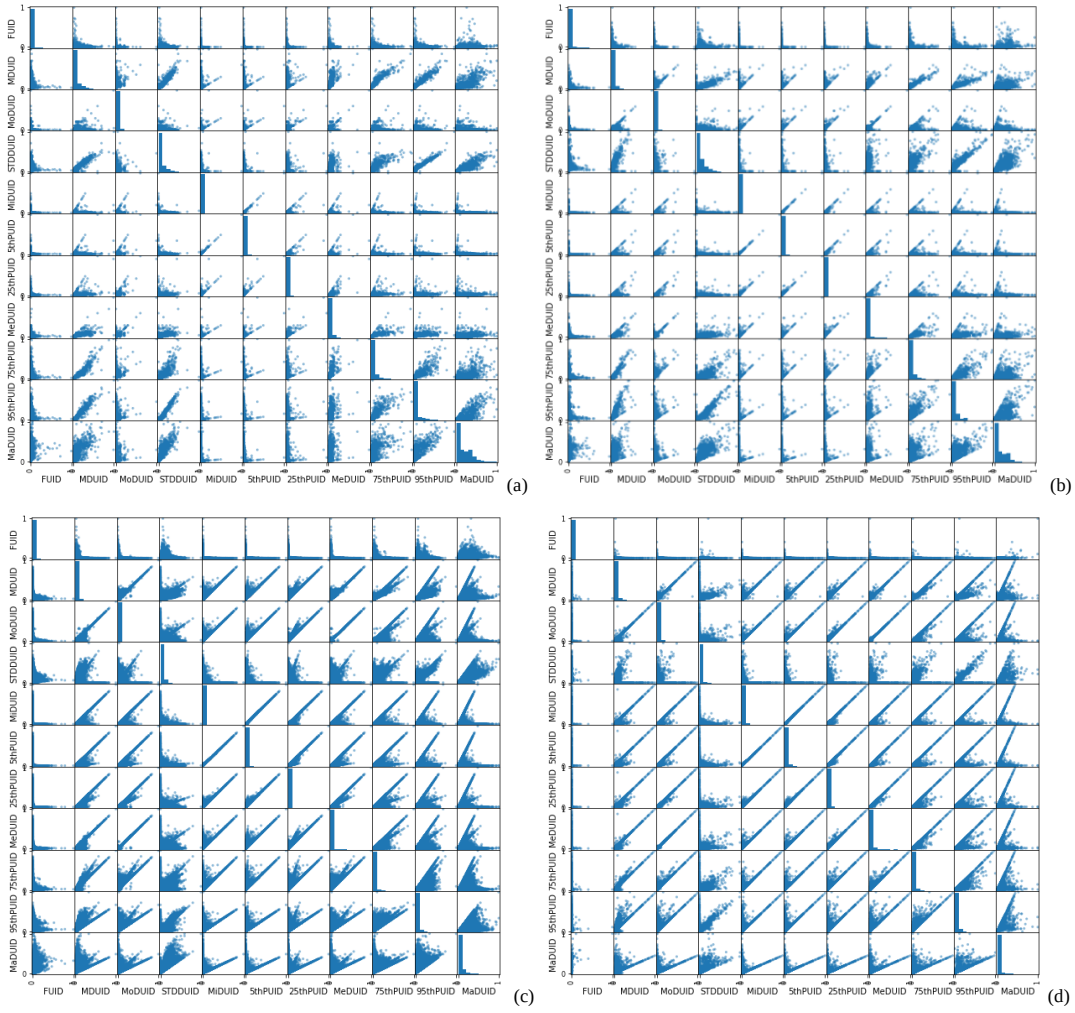

FIGURE 4. (a) Scatter Matrix of extracted features for UserID, (b) Scatter Matrix of extracted features for DeviceID, (c) Scatter Matrix of extracted features for PatientID, (d) Scatter Matrix of extracted features for Routine

The scatter matrix displays the positive and negative correlation between the features. In this case, from the visual inspection, the majority of features have a positive correlation. However, based on Figure 4, the consideration would be to remove the feature Frequency for each Unique Identifier (FUID) for the UserID, Routine and Device Interaction classification but retain it for PatientID.

Referring to the Routine and Device Interaction, the data collected relates predominately to unique routine combinations, so logically the FUID feature is less significant, as confirmed by the scatter matrix.

IV. EXPERIMENT AND RESULTS

A case study of actual EPR audit data is presented as an evaluation of the system methodology. 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). The dataset contains four distinct ID types, User, Patient, Device and Routine. Each User ID, Patient ID and Device ID is tokenised by isolating the unique entries and assigning each value an incrementing number. This is done to anonymise the dataset. The Routine ID was not tokenised as it denotes the tasks performed by the User on the EPR for the interaction. For example, in the first row of Table 1, User 865 accesses the 'Pharmacy Orders' function of the EPR on Patient 58991 whilst using Device 362.

For every value of each of the four IDs, a LOF anomaly score was calculated. The LOF anomaly score measures the local deviation of density through determining how isolated the value given by *k*-nearest neighbours (k is set to 5). A LOF anomaly score 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.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2906503, IEEE Access

A LOF anomaly score is calculated by taking the number of variants according to the mathematical combination and is calculated using the equation in (17). As there are ten features, 45 LOF scores are calculated to account for all the feature combinations for every ID in the dataset. There are 90,385 unique IDs in the dataset in total (for user, patient, device and routine combined), and a LOF score is calculated for the 45 combinations (of the 10 features) for each of the unique IDs in the dataset. Therefore 4,067,325 unique LOF scores are calculated in total. Data cleaning is then performed on the LOF scores in order to convert the 'NaN' and 'Inf' values. A NaN value indicates that a point has many neighbours in the same location, therefore the ratio of densities is undefined, and the points are not outliers. An Inf value occurs when a point is next to several identical points, but is not itself a member of that cluster, they are therefore 'infinite' and can be classified as anomalous. The NaN values are therefore assigned a value of 1, to indicate it is not anomalous, and the Inf values are assigned a value of 2, to indicate they are anomalous. The mean LOF scores for each ID is then calculated and the highest anomaly scores are presented in Table III and IV.

$$\binom{n}{k} = \frac{n(n-1)\dots(n-k+1)}{k(k-1)\dots 1} \tag{17}$$

A. USER, PATIENT AND DEVICE ID

There are 1,515 unique User IDs, 72,878 unique Patient IDs and 2,270 unique Devices within the dataset. In Table III, IV, and V LOF identifies anomalous User IDs, Patient IDs and Device IDs. The neighbourhood radius is defined in stage 3 of the LOF algorithm (Section B, 1), the density score is defined in stage 4, and the anomaly score is the final LOF value, as defined in stage 5.

TABLE III
LOF (MEAN) ANOMALY SCORES FOR USER ID

User ID	Density Score	Anomaly Score	Neigbourhood Radius
685	3.03	4.36	0.546
260	5.73	3.54	0.518
1037	69.14	2.80	0.251
1002	46.81	2.61	0.051
1401	16.55	2.56	0.153
707	19.05	2.28	0.207
1311	83.73	2.23	0.016
242	77.78	2.13	0.024
1493	47.75	2.03	0.134
507	28.66	2.00	0.103

Within the User ID range, the most notable ID is #685, with an anomaly score of 4.36. There are 10 User IDs with an anomaly score above 2. Therefore LOF has indicated that 0.66% of the User IDs are anomalous.

TABLE IV
LOF (MEAN) ANOMALY SCORES FOR PATIENT ID

Patient ID	Density Score	Anomaly Score	Neigbourhood Radius
35888	371.74	9.41	0.006
19327	175.92	8.81	0.018
58816	794.70	8.58	0.003
69053	51.59	8.55	0.053
51280	765.21	7.61	0.003
41306	150.01	7.53	0.014
46695	647.07	7.32	0.008
13704	1315.64	6.99	0.002
34419	23.12	6.97	0.101
56428	2570.47	6.94	0.003

Similarly, the most notable Patient ID is #35888, with an anomaly score of 9.41. There are 122 Patient IDs with an anomaly score above 2, indicating 0.17% of the Patient IDs are anomalous.

TABLE V
LOF (MEAN) ANOMALY SCORES FOR DEVICE ID

Device ID	Density Score	Anomaly Score	Neigbourhood Radius
2258	374.85	4.86	0.003
1082	2.26	4.75	0.730
1557	168.84	2.92	0.009
729	5.26	2.80	0.303
499	29.58	2.52	0.048
527	6.84	2.43	0.206
896	10.35	2.32	0.170
2014	6.75	2.29	0.216
1104	107.50	2.28	0.033
523	17.38	2.25	0.077

Finally, the most notable Device ID is #2258, with an anomaly score of 4.86. There are 12 Device IDs with an anomaly score above 2, indicating that 0.53% of the Device IDs are irregular. Overall therefore, LOF identifies 0.45% of IDs as anomalous, which would be highlighted to a patient privacy officer for investigation.

Examples of audit log data classified as inlier, outlier and abnormal data for User ID is presented in Table VI. Audit log data classified as an inlier within the dense region (<1) is User ID 571, with a LOF score of 0.95. Audit log data classified as an outlier within the normal region (>1 and <2) is User ID 1486, with a LOF score of 1.12. Audit log data classified as an outlier within the abnormal region (>2) is User ID 707, with a LOF score of 2.28.

TABLE VI
EPR AUDIT LOG DATA EXAMPLES FOR INLIER, OUTLIER AND ABNORMAL DATA POINTS FOR USER ID

Date & Time	Device ID	User ID	Routine Description	Patient ID	Duration (sec)	Adm Date	Dis Date
17/03/08 01:32	2046	571	Visit History	33727	28	08/03/2017	08/03/2017
17/08/07 15:37	396	1485	Current Medication Orders | Pharmacy Orders	62584	58	16/10/2001	16/10/2001
16/05/30 11:09	936	707	Visit History | Radiology Reports | Maternity Data | Cancelled Account.UK.Letter | Cancelled Account.UK.Scheduling UK.View Orders	28160	385	26/01/2016	26/01/2016

The results presented here demonstrate a technique for uncovering anomalous or irregular behavioural patterns from a complex dataset that would otherwise not be possible from either a visual inspection/visualisation of the whole dataset (such as the heatmap presented in Figure 2).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2906503, IEEE Access

B. ROUTINE ID

However, the LOF technique cannot be applied as effectively to the Routine ID. Table VII presents a sample of the highest LOF anomaly scores for the Routine ID dataset.

The EPR audit logs calculate a string of routines performed on the same patient as a unique Routine ID. The differing routines are delimited with a pipe (|). Therefore there are 13,722 Routine IDs in the dataset, whereas there are more accurately approximately 100 unique routines a user could perform.

TABLE VII
LOF (MEAN) ANOMALY SCORES FOR ROUTINE ID

Routine Set Description	Density Score	Anomaly Score	Neigbourhood Radius
Assessment Forms | Maternity Data | Care-Area Administrative Data | Admissions Demographic Data	1043.094	13.34	0.003
***** | UK.View Orders | Admissions Demographic Data | Pharmacy Orders**	1649.703	11.64	0.005
***** | Cancelled Account.UK.Letter | Admissions Demographic Data**	2213.821	11.41	0.004
Maternity Data | Theatre Management | Assessment Forms | Visit History	581.246	11.35	0.004
Theatre Management | Cancelled Account.UK.Letter | Cancelled Account.UK.Scheduling | Admissions Demographic Data	632.774	9.70	0.005
Recent Clinical Results | Recent Clinical Results:(Departmental Reports) | Pharmacy.Medication Order History | UK.View Orders	70.561	9.54	0.035
Assessment Forms | Admissions Demographic Data | Visit History | Alerts	601.429	9.29	0.004
Cancelled Account.UK.Letter | Pharmacy Orders | Admissions Demographic Data	470.423	8.81	0.005
Assessment Forms | Cancelled Account.UK.Letter | Cancelled Account.UK.Scheduling | Medication Order History	646.410	8.32	0.006
Internet Access | Alerts | Assessment Forms	693.934	8.22	0.005

There are 102 routine sets with an anomaly score above 2. Therefore LOF has indicated that 0.74% of the routine sets are anomalous. The most notable routine set is the combination 'Assessment Forms | Maternity Data | Care-Area Administrative Data | Admissions Demographic Data', with an anomaly score of 13.34. This specific routine combination only occurs twice in the audit logs of over 1,000,000 rows. However, in order for the LOF scores for routine to be of value, each routine (rather than the routine combination) would need to be calculated. Unfortunately, this cannot be differentiated within the dataset. For example, if the LOF scores for each routine are calculated individually (rather than as a routine set), such as 'Assessment Forms' and 'Maternity Data', then these values can be compared with other instances of that routine, to determine whether certain log accesses are anomalous. However, as these cannot be separated within the combinations of routines, then an informative LOF score cannot be determined for Routine ID.

C. VISUALISATION OF RESULTS

A visualisation of the LOF results for each ID is presented in Figure 5.

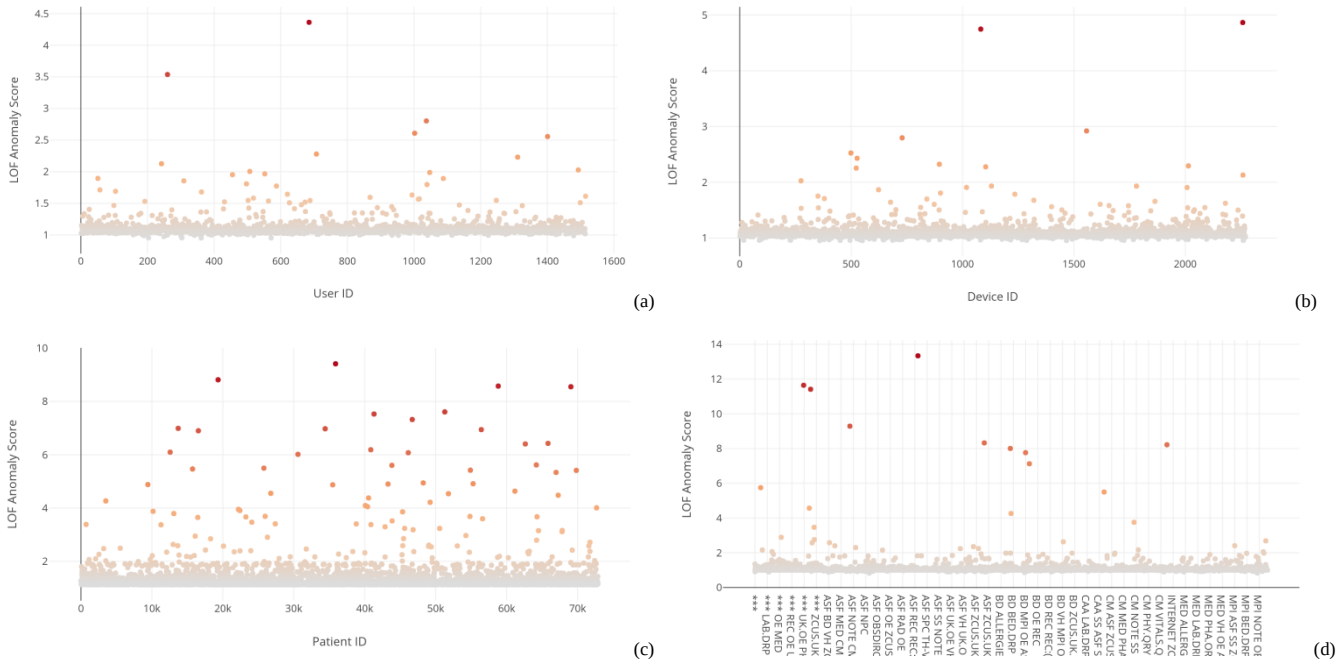

FIGURE 5. (a) Scattergraph of LOF results for UserID, (b) Scattergraph of LOF results for DeviceID, (c) Scattergraph of LOF results for PatientID, (d) Scattergraph of LOF results for Routine

Through visualising the anomalies in this way, outliers can be highlighted to an analyst for scrutiny. In our visualisation engine, outliers in the top quarter of each ID range are highlighted as red, to be investigated as a priority.

Outliers in the 3rd quarter appear orange, and outliers in the 2nd quarter appear yellow. This creates an interactive live task list for the analyst, with an anomaly priority ordering. Clicking on a point displays the ID number, which allows the analyst to investigate the activity associated with the ID. The display updates when new data is input and new LOF scores are calculated, providing a current view of anomalous EPR activity within a hospital. Activity such as insider threats (a staff member misusing their access privileges), or external threats (such as credentials accessed through social engineering and utilised for data exfiltration), can be investigated. In this way, the system provides situational awareness to aid patient privacy officers to monitor for malicious or unusual activity proactively.

V. CONCLUSION AND FUTURE WORK

The far-reaching consequences of this work are illustrated with a prediction: This research project will increase the situational awareness of data flow and actively address this issue of data misuse. Machine learning algorithms have the capability to observe and learn patterns of data and profile users' behaviour, which can then be represented visually. The far reaching consequences of this work will result in the development of a system that can be used by healthcare practitioners to increase the protection of their EPR records. This will make the UK, not only one of the safest places to conduct business, but also one of the securest in protecting patient privacy in healthcare systems.

Future Work will involve normalising the data further with a case study of the routine 'Pharmacy Orders'. This routine accounts for approximately 21.27% of the actions performed on the EPR. It is therefore possible to use this as a case study to understand user roles within the dataset and compare similar actions, in order to identify anomalous behaviours. Factors other than solely the duration of the routine (such as the date and time an action is performed) will be considered. Additionally, a quantitative model-based approach that takes into account the duration and the sequence of events during the interaction of the user with the EPR will be explored.

The features discussed in the paper compare every activity performed associated with each ID, but without detail. For example, for each User it compares the duration of all actions performed for that user. This can broadly identify anomalous behaviour, but for a more nuanced approach, other factors can be taken into consideration. For example how long a user typically spends performing a certain task, or accesses a specific device, or with a particular patient. By calculating the local outlier factor for these behaviours, and assigning each a weighted score, these can be factored together to provide data-driven insight of potential EPR misuse. Additionally, currently inputting new data to calculate their LOF values is a manual process and not in real-time. This will be explored further with an aim to automate this and improve update efficiency within the big data context of EPR audit logs.

Future work will also incorporate game theory through the use of an interactive visualisation. The vision is that 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. Additionally, The Theory of Gamified Learning infers that gamification can positively affect learning and decision making through a more direct mediating process and a less direct moderating process [30]. Firstly, gamification affects learning via mediation when a user's behaviour is encouraged in such a way that it itself improves learning outcomes, such as a fitness app [31]. The theory therefore mediates the relationship between game elements and learning. Secondly, gamification affects learning via moderation when pre-existing information is improved through strengthening the relationship between instructional design quality and outcomes [32]. For the moderation theory, the moderator does not influence the outcome construct independently of the causal construct, therefore the pre-existing information must be of high quality, or the addition of gamification techniques would be of no benefit. Through the use of visualisation techniques to enhance the results of the local outlier factor results, gamification moderation theory is implemented.

Supervised learning techniques will be implemented to compliment the unsupervised LOF scores. Access to labelled data for EPR audit logs is often not available or comprehensive. However, through displaying LOF results to an analyst, upon investigation the analyst can label the data as legitimate or illegitimate. Through this process, the combined use of unsupervised and supervised machine learning algorithms results in a semi-supervised approach to the challenge of detecting EPR misuse. Additionally, once semi-supervised techniques are employed, the accuracy of the algorithms in detecting outliers can be quantified through feedback from analysts.

REFERENCES

[1] ICO, "Data security incident trends," 2016. [Online]. Available: https://ico.org.uk/action-weve-taken/data-security-incident-trends/. [Accessed: 02-Oct-2017].

[2] 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.

[3] J. Stoll and R. Z. Bengez, "Visual structures for seeing cyber policy strategies," in *2015 7th International Conference on Cyber Conflict: Architectures in Cyberspace*, 2015, pp. 135–152.

[4] 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.

[5] G. Zhao and D. W. Chadwick, "On the modeling of Bell-LaPadula security policies using RBAC," in *Proceedings of the Workshop on Enabling Technologies: Infrastructure for Collaborative Enterprises, WETICE*, 2008, pp. 257–262.

[6] Fair Warning, "How Privacy Considerations Drive Patient Decisions and Impact Patient Care Outcomes Purpose of the

Study and Executive Overview Report," 2011.

[7] J. J. Walker, T. Jones, and R. Blount, "Visualization, modeling and predictive analysis of cyber security attacks against cyber infrastructure-oriented systems," in *2011 IEEE International Conference on Technologies for Homeland Security (HST)*, 2011, pp. 81–85.

[8] W. Hurst and C. Dobbins, "Guest Editorial Special Issue on: Big Data Analytics in Intelligent Systems," *J. Comput. Sci. Appl. Spec. Issue Big Data Anal. Intell. Syst.*, vol. 3, no. 3A, pp. 1–9, 2015.

[9] H. M. Chao, C. M. Hsu, and S. G. 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.

[10] Engineering and Physical Science Research Council, "Scheme to Recognise Academic Centres of Excellence for Cyber Security Research," 2014. [Online]. Available: https://epsrc.ukri.org/research/centres/acecybersecurity/. [Accessed: 04-Apr-2018].

[11] Malvern Cyber Security Cluster, "Cyber Valley: Malvern Cyber Fuzzy Rule Based Big Data Analytics in Cloud Computing," *IEEE J. Biomed. Heal. Informatics*, vol. 22, no. 5, pp. 1605–1618, Sep. 2018.

[17] A. Barnett *et al.*, "Image Classification using non-linear Support Vector Machines on Encrypted Data," *IACR Cryptol. ePrint Arch.*, p. 857, 2017.

[18] 2017. Dean, T. and Stockdale, J., DARKTRACE Ltd, "Anomaly alert system for cyber threat detection," 06-Feb-2017.

[19] Y. Hu, "The Mathematica ® Journal Efficient, High-Quality Force-Directed Graph Drawing," 2006.

[20] M. Jacomy, T. Venturini, S. Heymann, and M. Bastian, "ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software," *PLoS One*, vol. 9, no. 6, p. e98679, Jun. 2014.

[21] D. Rose and N. Joshi, *The Digital Doctor: Hope, Hype, and Harm at the Dawn of Medicine's Computer Age*, vol. 71, no. 1. 2018.

[22] 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.

[23] B. Esmael, A. Arnaout, R. K. Fruhwirth, and G. Thonhauser, "A Statistical Feature-Based Approach for Operations Recognition in Drilling Time Series," *Int. J. Comput. Inf. Syst. Ind. Manag. Appl.*, vol. 5, pp. 2150–7988, 2013.

[24] T. M. Mitchell, "The Discipline of Machine Learning," *Mach. Learn.*, vol. 17, no. July, pp. 1–7, 2006.

[25] P. Domingos, "A few useful things to know about machine

Security Cluster." [Online]. Available: https://www.malvern-cybersecurity.com/. [Accessed: 04-Apr-2018].

[12] North West Cyber Security Cluster, "North West Cyber Security Cluster | Helping Prevent Cyber Crime." [Online]. Available: http://www.nwcsc.org.uk/. [Accessed: 04-Apr-2018].

[13] H. Government, "National Cyber Security Strategy 2016-2021," 2016.

[14] 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.

[15] V. Stanovov, C. Brester, M. Kolehmainen, and O. Semenkina, "Why don't you use Evolutionary Algorithms in Big Data?," in *IOP Conference Series: Materials Science and Engineering*, 2017, vol. 173, no. 1, p. 012020.

[16] A. Jindal, A. Dua, N. Kumar, A. K. Das, A. V. Vasilakos, and J. J. P. C. Rodrigues, "Providing Healthcare-as-a-Service Using learning," *Commun. ACM*, vol. 55, no. 10, p. 78, Oct. 2012.

[26] L. Buitinck *et al.*, "API design for machine learning software: experiences from the scikit-learn project," *arXiv Prepr. arXiv ...*, pp. 1–15, Sep. 2013.

[27] J. Lee, B. Kang, and S. H. Kang, "Integrating independent component analysis and local outlier factor for plant-wide process monitoring," *J. Process Control*, vol. 21, no. 7, pp. 1011–1021, Aug. 2011.

[28] M. M. Breunig *et al.*, "LOF," in *Proceedings of the 2000 ACM SIGMOD international conference on Management of data - SIGMOD '00*, 2000, vol. 29, no. 2, pp. 93–104.

[29] C. Croux and G. Haesbroeck, "Influence Function and Efficiency of the Minimum Covariance Determinant Scatter Matrix Estimator," *J. Multivar. Anal.*, vol. 71, no. 2, pp. 161–190, Nov. 1999.

[30] R. N. Landers, "Developing a Theory of Gamified Learning: Linking Serious Games and Gamification of Learning," *Simul. Gaming*, vol. 45, no. 6, pp. 752–768, Dec. 2014.

[31] J. Hamari, J. Koivisto, and H. Sarsa, "Does gamification work? - A literature review of empirical studies on gamification," in *Proceedings of the Annual Hawaii International Conference on System Sciences*, 2014, pp. 3025–3034.

[32] R. M. Baron and D. A. Kenny, "The Moderator-Mediator Variable Distinction in Social The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations," *J. Pers. Soc. Psychol.*, vol. 51, no. 6, pp. 1173–1182, 1986.