Smart Health 26 (2022) 100354

Contents lists available at ScienceDirect

Smart Health

journal homepage: www.elsevier.com/locate/smhl

Securing electronic health records against insider-threats: A supervised machine learning approach

William Hurst [a,*], Bedir Tekinerdogan [a], Tarek Alskaif [a], Aaron Boddy [b], Nathan Shone [c]

[a] Information Technology Group, Wageningen University and Research, Leeuwenborch, Hollandseweg 1, 6706 KN Wageningen, the Netherlands
[b] Aintree Hospital, Liverpool, L9 7AL, United Kingdom
[c] Department of Computer Science, Liverpool John Moores University, Liverpool, L3 3AF, United Kingdom

ARTICLE INFO

Keywords:
Electronic health record
Machine learning
Anomaly detection
Security

ABSTRACT

The introduction of electronic health records (EHR) has created new opportunities for efficient patient data management. For example, preventative medical practice, rather than reactive, is possible through the integration of machine learning to mine digital patient record datasets. Furthermore, within the wider smart cities' infrastructure, EHR has considerable environmental and cost-saving benefits for healthcare providers. Yet, there are inherent dangers to digitising patient records. Considering the sensitive nature of the data, EHR is equally at risk of both external threats and insider attacks, but security applications are predominantly facing the outer boundary of the network. Therefore, in this work, the focus is on insider data misuse detection. The approach involves the use of supervised classification (decision tree, random forest and support vector machine) based off pre-labelled real-world data collated from a UK-based hospital for the detection of EHR data misuse. The results demonstrate that by employing a machine learning approach to analyse EHR data access, anomaly detection can be achieved with a 0.9896 accuracy from a test set and 0.9908 from the validation set using a support vector machine classifier. The emphasis of this research is on the detection of EHR data misuse, through the detection of anomalous behavioural patterns. Based on the results, the recommendation is to adopt an SVM for data misuse/insider threat detection.

1. Introduction

The last ten years have seen an increasing uptake of digital records for documenting patient health care data, offering instantly available information, regardless of the patient location. This change is part of the exponential rise in smart amenities, providing real-time services within the smart cities' domain (Parah et al., 2020). The advantages of this approach are clear; providing efficiency, location independent access to information, enhanced communication, as well as general improved patient care and experience. Subsequently, Electronic Health Records (EHR) are widely recognised for improving the healthcare infrastructure (Bujnowska-Fedak & Wysoczański, 2020). Their core benefits, as presented in Fig. 1, are outlined as follows. 1) Time value created by a reduction in documentation time for admin staff and nurses. This is crucial in a role with a large amount of paperwork (e.g. 10 pieces of paper per

* Corresponding author.
E-mail addresses: will.hurst@wur.nl (W. Hurst), bedir.tekinerdogan@wur.nl (B. Tekinerdogan), tarek.alskaif@wur.nl (T. Alskaif), dr.aaronboddy@gmail.com (A. Boddy), n.shone@ljmu.ac.uk (N. Shone).

https://doi.org/10.1016/j.smhl.2022.100354
Received 11 April 2022; Received in revised form 31 August 2022; Accepted 16 October 2022
Available online 20 October 2022

patient per visit); 2) Information quality, where EHR have shown to offer a greater completeness of information, which also means less time spent searching for missing data; 3) Financial benefits including savings for both inpatient and outpatient services. Research shows EHR produces greater revenue through greater patient health tracking; 4) Environmental impact benefits in the form of a reduction in paper usage. For example, thousands of tons of paper are consumed by the healthcare industry across the globe each year. Removing the need for paper documentation has clear environmental benefits; and 5) Improved health care by tracking of medical history and integration of advanced health care applications.

Aside from these clear advantages, the data within is being used increasingly to systematise and increase the efficiency in which clinical decisions are made and processed. This is because the nature of EHR means that the data can be subjugated to Machine Learning (ML) applications for automated decision making (Tort et al., 2020). Furthermore, data mining of historical records enables medical staff to examine patient conditions with high flexibility (Cook et al., 2018) (Jose et al., 2020).

For example, Gallagher et al. detail how hospital readmission has a tremendous cost impact (as well as dissatisfaction and increased mortality) of unplanned hospital readmissions. With just over a quarter of readmissions preventable, there is a significant benefit to the identification of those at risk of being resubmitted. EHRs are placed suitably for mitigation against readmission of patients through the automated analysis and prediction of individuals or groups which are at greater risk of readmission (Gallagher, Zhao, & Brucker, 2020). EHR also supports making data findable, accessible, interoperable and reusable. However, with these principles it is important to preserve the privacy and security requirements, as EHRs are comprised primarily of demographic data, administrative information, and clinical data (Tort et al., 2020).

The result is, EHRs have become a lucrative target for exploitation through cyber/insider attacks - Where an insider threat refers to any malicious actor that performs actions designed to negatively impact on a system, of which they have prior knowledge, access and/or authorisation. A common instance is disgruntled employees sharing their access credentials with external parties, or internal intruders pretending to be real users in the network. In contrast to an insider attack, an outsider attack (within the cyber-security domain) refers to an unauthorised attempt by an individual(s) without direct access to any of the nodes in a network, yet with potential physical access to the property (Medhi & Huang, 2008, chap. 14). For clarity, Medhi et al. define four grades of outsider attack, including *Sniffing*, which is a passive process involving collecting and recording of unsecured data transmitted on a network; *Falsification*, referring to the substitution and insertion of false data on a router; *Obstruction*, where an example would include overloading by means of routing dummy traffic in high volume and; *Replay*, which involves fraudulently routing valid data transmission repeatedly. Typically, outsider attacks are prevented by means of an outward-facing security system, as discussed in Section 2.1.

With the focus of the research presented in this article on insider-threats, we consider the three gradings of insider attack by Janjua et al. (Janjua et al., 2020). Namely this includes *malicious*, *careless* and *compromised*. The first being an individual or organisation who deliberately attempt to steal information, cause a disruption or access sensitive personal data. The second refers to individuals working for an organisation who simply do not comply to existing security procedures, thus placing the organisation at risk. The final classification relates to a legitimate user whose login credentials have been hacked/stolen, and this is, of course, linked to the first grading. The focus of this work is on gradings one and three: malicious and compromised.

The objective of this work is to adopt a supervised machine learning approach (decision tree, random decision forest and Support Vector Machine (SVM) classification) for anomalous data usage detection, which supports the malicious and compromised insider threats. SVM in particular is an ideal choice for this approach as it has been a main-stay and consistently high-scoring solution for anomaly detection approaches within a security setting (Li, 2018). This is because one desirable characteristic of an SVM is that it's hyperplanes are more resilient to outliers, which justifies its suitability for this classification task. Supervised is also selected over unsupervised as the data used for experimentation is labelled. Whilst SVMs, DTs and RFs have been widely used in related works (some of which are outlined in Section 2), to the best of our knowledge, this is the first time this approach has been conducted on this dataset and the data extraction process - involving the use of the four data groups from within the EHR dataset (routine, user, device and

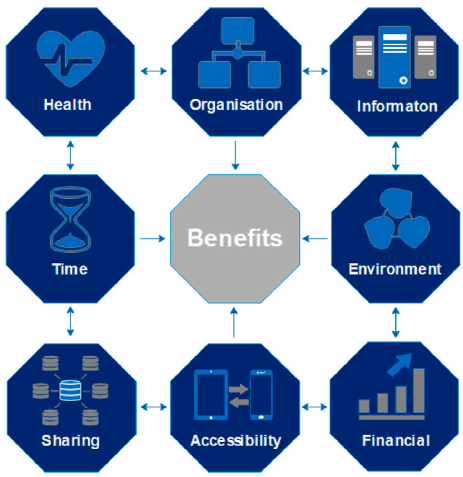

Fig. 1. Electronic health record benefits.

patient) - makes this work stand apart from other approaches. The closest existing work aligned to this investigation is provided by Menon et al., (who adopt an SVM methodology (Menon et al., 2014) among others). Thus, a discussion of the results achieved compared with the approach by Menon et al. is provided in Section 4.3. Whilst both approaches involve an SVM and focus on EHR data, the data pre-processing approaches differ significantly.

Aside from the approach by Menon et al. the increasing level of risks to EHR systems have led to numerous other emerging research investigations into novel techniques to protect both the systems and the information within the digital records. For example, Perumal et al. focus their research on master-key management by proposing a novel architecture for the key management system within an e-health infrastructure (Perumal & Nadar, 2020). Their technique involves the use of key generation and encryption modules, coupled with a multi-key server approach to provide faster access to healthcare content. The approach focuses on login verification and encrypted data transmission to secure healthcare network access (Perumal & Nadar, 2020). This differs from the work poised in this article, where the attention is on anomaly detection and data misuse by means of behavioural analysis. Whereas Saha et al. discuss a framework for preserving the privacy of EHR (Saha et al., 2019). Their approach employs a combination of multiple end-user devices (e.g. health monitoring systems, laptops, a layer of fog, access points, servers and routers). The layers then provide a cryptographic exchange process, a consensus approach to control the view of EHRs and a query handling process (Saha et al., 2019). The framework is tested within a cloud-fog infrastructure, in which a reduced transaction time occurs when compared with similar approaches. This approach differs significantly to that which is poised in this article.

Whilst SVM, KNN and DTs are already widely used, there is a strong argumentation for their application in this setting. As Mehta et al. and Zheng et al. discuss, evidence suggests that machine learning is an ideal technique for a wide variety of EHR-based data analysis applications (Mehta, 2018) (Zheng, Xie, Xu, & He, 2017); with kNN, DTs and RFs as notable approaches for an anomaly detection process. Further, the aforementioned research by Janjua et al. which makes use of linguistic analysis to determine an employee's risk level, focusing primarily on emails also adopts an SVM for insider threat detection, yet other techniques are benchmarked (e.g. NB, KNN, LR, etc.). There is an argument for adopting more advanced techniques. For example, Anakath et al. employ deep belief neural networks, however, their focus is not specifically on EHR data protection and hospital networks; but rather on generic cloud solutions (Anakath et al., 2022). Their approach is also tested by means of the Cooja simulator rather than real-world data, as employed in this article. Crucially, as Adil et al. discuss, there are very few articles adopting a machine learning approach for EHR security despite the success of its application in related security domains (Seh et al., 2021). This is also confirmed by Qayyum et al. who question the levels of robustness for the use of machine learning within healthcare-based security applications (Adnan Qayyum et al., 2020).

Clearly existing research indicates that there is still a need for experimentation with machine learning for EHR security application. Thus, we offer the following contributions to knowledge: 1) The research in this article is the first to test well-established techniques (i.e. SVM, DT and RF) on an actual EHR dataset. Few other approaches use real world data on this scale and many adopt simulation for testing; yet none have tested the approach on this specific dataset. For clarity, the data used for the analysis is provided by a UK-based hospital (which the authors have made available on EASY-DANS (Hurst, 2021)), with the anomalous readings pre-labelled by means of density-based classification outlined in our related work in (Hurst et al., 2020) where anomalous points were confirmed in verbal consultation. 2) The focus of this approach is on insider threats, rather than external, and specifically focused on EHR datasets. Many other works offer generic solutions to insider-threats (discussed in Section 2.2) in which hospitals are provided as example applications, yet the approach in this article is bespoke for EHR. 3) The data pre-processing approach involves four dataset groupings. This data structuring process produces effective results (as detailed in Section 4) and also makes the approach stand out compared to related works. The authors found that division of the EHR into these four groupings allowed the classifiers to perform optimally, both in terms of Area Under the Curve (AUC) and speed. However, this also meant that the anomaly detection process could be conducted with only a partition of the full EHR records and without the use of highly sensitive information (such as names, addresses and personal health-related information). The remainder of this article is organised as follows. Section 2 provides a background discussion of EHR data, its applications and related work. Section 3 details the research methodology with the results outlined in Section 4. The paper is concluded in Section 5.

2. Background and related work

The digital smart cities revolution has also penetrated the healthcare sector, where e-health services are becoming increasingly prevalent. Their crucial role has been particularly apparent during the Covid-19 pandemic (Bujnowska-Fedak & Wysoczański, 2020), by facilitating healthcare efficiency to provide timely access to comprehensive and organised patient information; thus, enabling the delivery of high-quality patient care (Jacquemard et al., 2020). EHR systems function by collating a patient's health data, digitising the information and then serving as the data source for healthcare providers and medical institutions (Shi et al., 2020). EHRs are an amalgamation of clinical data repositories, clinical decision support, controlled medical vocabulary, order entry, computerised provider order entry, pharmacy, and clinical documentation applications' (Shiells, Alejandra Diaz Baquero, Štěpánková, & Holmerová, 2020). Features of EHRs include electronic diaries, automated text messages, clinic letter verification, prescriptions and the reviewing of investigative results (Misra et al., 2017). The implementation of EHR has been shown to influence the administration of clinical care, relationships between clinicians, and professional autonomy, affecting how health professionals operate (Ihlebæk, 2020).

2.1. EHR security

EHR portals enable patients to access their personal health information (Pho, 2019), view laboratory and imaging results, contact clinical staff with questions and updates, schedule future appointments and request medication refills (Girault, 2015). This unique

access to their own data enhances transparency and increases patient satisfaction, while breaking down barriers and improving communication between patients and clinical staff. However, the size and complexity of healthcare records is increasing (Tanwar et al., 2020) and the widespread adoption of EHR has resulted in the collection of a significant growing volume of clinical data (Lin et al., 2020). There is now a corresponding interest in exploiting analytical applications regarding this granular data repository (Khennou et al., 2018); where emerging secondary applications include (Cano et al., 2017) 1) epidemiological and pharmacovigilance studies; 2) facilitating recruitment to randomised controlled trials, audits and benchmarking studies; 3) financial and service planning; 4) enhanced clinical decision support for patient evaluation and treatment (Shoolin, 2017); and 5) supporting the generation of novel biomedical research outcomes. Where, for example, clinicians are provided with relevant healthcare information in a timely manner and offset risks from factors such as time-constraints practicing in stressful environments (Gil et al., 2019).

The applications for EHRs are further increasing in the UK in particular, largely due to the NHS Spine, which is a collection of national applications, services and directories that enable the exchange of information in national and local IT systems (Misra et al., 2017). The NHS Spine is a result of the National Programme for IT, which failed to deliver an integrated EHR, but instead achieved the creation of the Spine, the N3 Network, choose and book, picture archiving, communication systems and standards which have allowed integration (Peckham, 2016).

However, there are various risks with the digitisation of health records. The introduction of EHR systems can also result in unintended negative consequences (Jacquemard et al., 2020). At a practical level, in some cases, the overuse of reminders and pop-up prompts has led to desensitisation and a tendency to ignore the provided information (Cecil, Dewa, & Ma, 2019). With more serious implications, many EHR applications also do not follow standardised protocols for data storage and application programming interfaces (Evans, 2016). This makes it challenging to interface EHRs with other clinical systems and impacts on data extraction (Schwartz et al., 2019). Other undesirable values have also emerged, ranging from installation errors (leading to erroneous health status reports), poor cybersecurity practices, sharing data with commercial parties affecting patient trust and a failure to appreciate the limitation. Even biases in datasets unfairly privileging or discriminating against certain ethnic groups have been known to occur (Ledford, 2019).

As a common challenge for all EHR providers, the security risk is a core topic of research, especially given the complex nature of healthcare systems. Relying on human vigilance alone is not a practical method for security records (Millard, 2017); as reflected in a report by the European Union Agency for Network and Information Security (ENISA), which outlines how cyber-attacks on e-health systems have a high societal impact (Liveri, Sarri, & Skouloudi, 2015). New systems are needed for safeguarding EHRs. A combination of behavioural variables and standard security measures should be employed to safeguard systems.

2.2. Related work

Argaw et al. discuss that the healthcare industry is among the top three sectors to be most affected by ransomware attacks (Medhi & Huang, 2008, chap. 14). The year 2016, in particular, saw a sharp rise in attacks on EHR repositories, with over 12 million record breaches (a 300% increase from 2015), many of which are documented as for sale on the dark web (Millard, 2017). High profile cyber-attacks, such as the well-documented WannaCry ransomware (which affected more than 100 countries (Millard, 2017)), demonstrates the potential that successful attacks have to disrupt access to millions of records of patients.

The mainstay technology currently in place involves the use of a firewall for safeguarding measures, a typical approach to counter the aforementioned four grades of outsider attacks. As Kruse et al. discuss, the technique is a proven approach for maintaining the health of the network (Kruse et al., 2017) and restricting access for unauthorised users. Standard firewall types include 1) filtering, e.g. assessing the internal/external electronic feeds; 2) inspection, where the incoming electronic feed is correlated with previously filtered feeds; and 3) application level gateway involving the firewall acting as an intermediary (Kruse et al., 2017). Yet, the premise of their limitations is basing the functionality on pre-defined rule sets (such as IP filtering). Illegitimate access by means of stolen account details (e.g. insider threats) remains a challenge for firewalls. Other major risks to EHR systems include 1) mass scale/system-wide shut down through ransomware attacks (Millard, 2017), many of which have taken placed globally. It is clear that many hospitals are still unprepared for the sophistication of attacks on their networks as, 2) the use of current standards is not yet suitable for mass deployment of EHR usage (Sanchez-Guerrero, Almenarez Mendoza, & Diaz-Sanchez, 2017); and 3) there are insufficient techniques to manage, store and control the sensitive data (Sanchez-Guerrero, Almenarez Mendoza, & Diaz-Sanchez, 2017). Furthermore, 4) within the e-health environment, as the goal is to allow multiple access points to EHR data for the benefit of the patients, this creates inherent risks within the Mobile Cloud Computing (MCC) environment. Particularly privacy leakage is a great concern. Wang et al. therefore, discuss the general architecture of using mobile cloud in digital health care environments (Wang & Jin, 2019). They outline how more users are becoming increasingly reliance on MCC services for remote access to EHR data. Where, 5) it is apparent, that due to the need to access data remotely, a successful attack would also have an impact on healthcare service staff being able to fulfil their roles from outside the healthcare facility.

Zhou et al. also consider the need for flexible data access, and propose two anomaly schemes involving 1) hiding the patient's identity using role-based access control and 2) the use of a scheme built on a bilinear grouping (Zhou et al., 2016). Their approach achieves effective results, with the authors able to produce a flexible access control method with encapsulated EHR data. Their technique has clear benefits for hiding patient/doctor interactions documentation. Al-Zubaidie et al., also focus on preserving patient identifies by means of pseudonymization and anonymisation (Al-Zubaidie et al., 2019) for safe patient data access within a PAX framework. Their approach has the advantage of not requiring continuous mining of patient data; and is both inward/outward facing. However, the work makes use of a simulation environment. Whilst this is effective for experimentation, the system is yet to be tested in a real-world environment. Whereas, in this paper a real-world hospital data is employed for the analysis process, which is

advantageous. Further, the use of blockchain for securing EHR records is also becoming more prominent. Many works, for example (Hang et al., 2019) (Shen et al., 2019), assess the use of blockchain technologies for securing communication channels, cloud infrastructures and general security of the data transfer process. Blockchain is an effective technology for improving the sharing of healthcare data. However, the use of blockchain falls out the scope of the work poised in this article, as the focus is on illegitimate data access in the case of stolen or mis-used system credentials. Blockchain offers an effective solution for the security of data against external cyber-threats, however analysis of the data patterns is not provided.

Novel security applications are paramount for safeguarding EHR and, in this article, the authors propose an approach for detecting anomalous EHR record access, by means of an inward-facing detection methodology to counter insider attacks specifically as there is a need for advanced anomaly detection approaches to ensure privacy, integrity and access to EHR data. Machine learning for this application has, to-date, been relatively under-used for EHR security despite its success in other cyber security applications (). This is confirmed in the detailed findings by Adil et al. who conduct an SLR-based analysis of existing applications of machine learning for safeguarding EHR (Seh et al., 2021). Their analysis covers 7 digital libraries, and findings indicate that there are 19 related articles in this domain (3 of which are by the authors of this article), from which the K-nearest neighbour algorithm is predominantly employed. From their findings, notable related articles include the aforementioned article by Menon et al. who investigation inappropriate access to EHR by means of collaborative filtering (Menon et al., 2014) and the works by Wesołowski et al. who adopt and ensemble-model (Wesołowski et al., 2016). The approach by Menon et al. involves use of the use of 34.1 million EHR accesses and a file-access dataset from Amazon. Their collaborative filtering model, which predicts a label for the interaction of a pair of entities, differs from the approach put forward in this article. The process involves predicting preferences for users based on historical preferences, whereas we adopt the use of four data groupings and focus on wider EHR content rather than accesses. Wesolowski et al. make use of data from 15 users acting in a legitimate capacity and 14 potential intruders. From which their architecture ensembles for classifiers, including C4.5, Bayesian network, SVM and RF. SVM and RF are both employed in this article but not in an ensemble capacity, making our approach different to that of Wesolowski et al. However, their ensemble approach offers a suitable commendation for the use of both algorithms in an EHR security-based application.

3. Anomaly detection approach

The EHR dataset used in this paper contains four distinct ID types, 1) routine-based actions, 2) user identification, 3) patient identification and 4) device interaction, as depicted in Fig. 2. Other fields are also commonly present within EHR datasets (e.g. Patient First Name, Patient Last Name, Position Title, Department, etc.) but these are not made available to this project due to information governance and staff privacy concerns. However, this also ensures the approach put forward is efficient, in that it functions by means of a vastly reduced dataset. Yet, the choice of fields available for the experimentation also ensure that this investigation is focused on insider-based data misuse detection. No external access to systems is considered within the investigation.

In the EHR dataset, the routine actions are unique activities performed whilst accessing the patient record (e.g. pharmacy orders, assessment form history, etc.). These differ to the user (medical practitioner) actions, which relates to who accessed the patient record and the access duration. The device data relates to how the record was accessed (e.g. which device number that data was accessed on), and the patient user information corresponds to an identifier for which record was accessed. A sample of the raw EHR data is displayed in Table 1 for clarification. The disaggregation of the EHR is a necessary step for benchmarking normal and anomalous behaviours for each of the four action groups in the record for the supervised learning; as the core activities within each of the groupings differ to a high extent - as outlined in detail in previous work (Hurst et al., 2020).

3.1. EHR dataset

Normal and abnormal behaviours are defined using density-based classification algorithms (such as local outlier factor and density-based spatial clustering of applications with noise), as outlined in detail in previous work (Hurst et al., 2020). The anomalous points are validated through a cross-check with the corresponding data record (Hurst et al., 2020). Once labelled, a feature extraction process

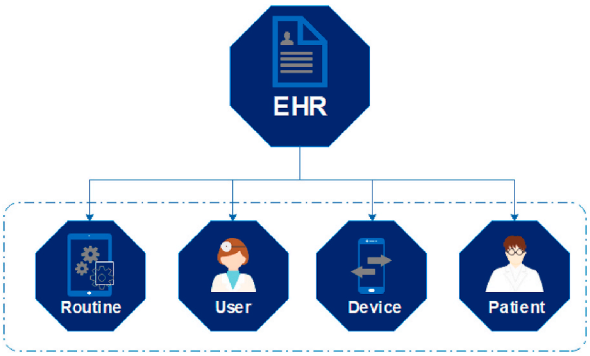

Fig. 2. EHR disaggregation.

Table 1
Example of EHR raw data.

Date&Time	Device	User ID	Routine	Patient ID	Duration
28/02/16 00:04	103	677	Assessment Forms Visit History	14,067	39
28/02/16 00:04	845	1489	Pharmacy Orders	49,304	22
28/02/16 00:06	923	199	Recent Clinical Results Recent Clinical Results: (Departmental Reports) UK.View Orders	60,948	165
28/02/16 00:08	775	568	Patient Care Notes	32,826	75
28/02/16 00:10	748	797	Recent Clinical Results Recent Clinical Results: (Departmental Reports)	2166	20

involving the frequency, mean, mode, standard deviation, min, 5th percentile, 25th percentile, median, 75th percentile, 95th percentile, max and outlier score is conducted (the feature descriptors are provided in Table 2). The data is then aggregated, and vectors are categorised as normal or anomalous, as in Table 3. A supervised machine learning process is then adopted for the anomaly detection, as presented in Fig. 3.

The dataset is unbalanced for the classification. Therefore, the machine learning challenge in this research is both an imbalanced and binary class problem. This is a common challenge in most real-world classification problems, where classes do not make up an equal portion of the dataset. Therefore, we have adhered to this data structure for the experimentation rather than adopting approaches, such as Synthetic Minority Over-sampling Technique (SMOTE) for balancing the dataset prior to classification.

Fig. 4 displays the count of the distribution between normal and anomalous behaviour in the dataset as a whole. However, it is of course possible to split this data to the different classes (i.e. Device, Patient, User, Routine), as displayed in Fig. 5.

3.2. Classification methodology

The algorithms selected for the anomaly detection process include i) decision tree, ii) random decision forest and iii) Support Vector Machine (SVM). Decision trees are a well-known classification tool for modelling/predicting decisions and their possible consequences based on chance event outcomes. It functions by means of a binary splitting technique, where feature dominance plays a crucial role in the prediction outcome. Decision trees provide an effective benchmark experiment due to their commonality and their efficiency. Random decision forests typically achieve a predictive accuracy by generating bootstrapped trees. A final predicted outcome is achieved through combining the results across all of the trees by using an average in regression/majority vote. The random forest approach offers an interesting comparison with the decision tree. As the decision tree makes its prediction from the entire dataset, whereas the random forest selects observations at random (and selects specific features) to build multiple decision trees. An SVM is primarily a discriminative classifier. It is concerned with optimal hyperplane calculation for categorising data points. It is an ideal technique for high-dimensionality data and working with binary decisions (Saha et al., 2019). Using the classifiers, the experiments involve training the algorithms on all labels, converted to either normal or abnormal with outlier score removed. The benefit of adopting this approach is that the challenge is a binary classification problem, i.e. normal compared with anomalous data.

4. Results

The results are achieved using a 70:30 training-test split in the dataset. For each classification experiment, the split is implemented

Table 2
Feature descriptions.

Measures of Central Tendency	
Feature ID	Description
Mean	The value of the most commonly found ID in the dataset. Calculated by dividing the total of the number of durations of the ID value by the frequency.
Mode	The most commonly occurring value in the ID data range
Measures of Variability	
Feature ID	**Description**
Standard Deviation	The measure of variation in the ID range
Max	The data value that is $\geq$ all other values in the ID range
Min	The data value that is $\leq$ all other values in the ID range
Frequency	The number of reoccurrences in the dataset of the ID
Measures of Position	
Feature ID	**Description**
5th Percentile	The value $<$ lowest 5% of the overall dataset
25th Percentile	The first (lower) quartile ($\leq$ 75% of values in the dataset)
75th Percentile	Third quartile ($\geq$ 75% of values in the dataset)
95th Percentile	$\geq$ 95% of values in the dataset
Outlier Score	The local outlier score as labelled in (Hurst et al., 2020)

Table 3
Example of labelled data.

Record	Label
1082	Anomaly
1734	Normal
24,603	Anomaly
805	Normal
1510	Normal
706	Anomaly
12,917	Anomaly
11,370	Normal

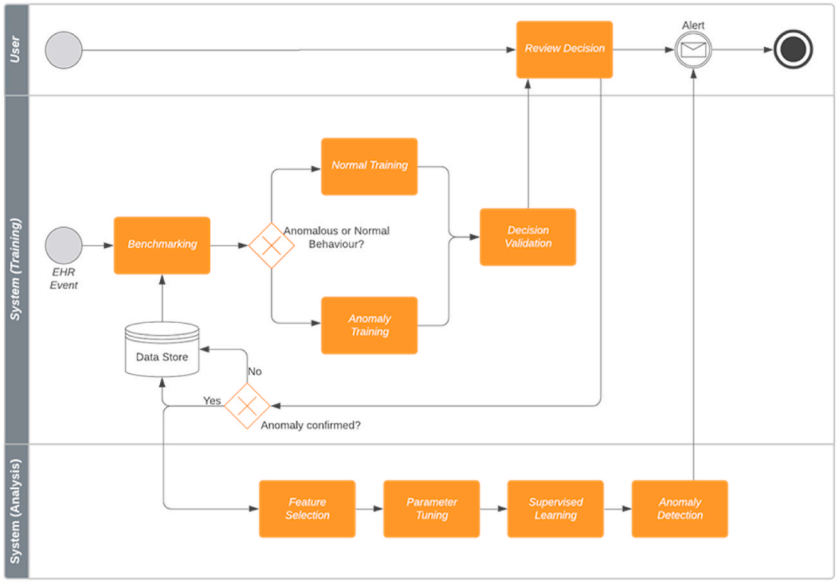

Fig. 3. Anomaly detection process.

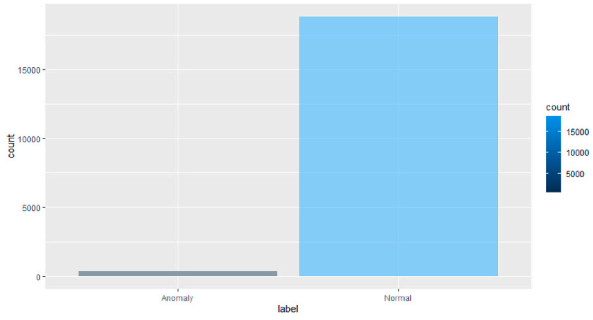

Fig. 4. Normal and abnormal dataset distribution.

using the caTools R package. The decision tree and random forest classification process offers a benchmark classification score for the anomaly detection process for comparison with the SVM. The performance of the classifiers is assessed using a confusion matrix, where the accuracy ((TP + TN)/Total), and misclassification error ((FP + FN)/Total) are calculated.

4.1. Benchmark experiment: decision tree and random forest

The decision tree algorithm is implemented using the rpart R package. Based on the test data, the classifier achieves a 0.018622 root node error (which is 250/13,425 nodes). The results are displayed in Table 4.

As presented in the above confusion matrix, the accuracy is 0.98783 and misclassification error rate is 0.01216. The classifier's

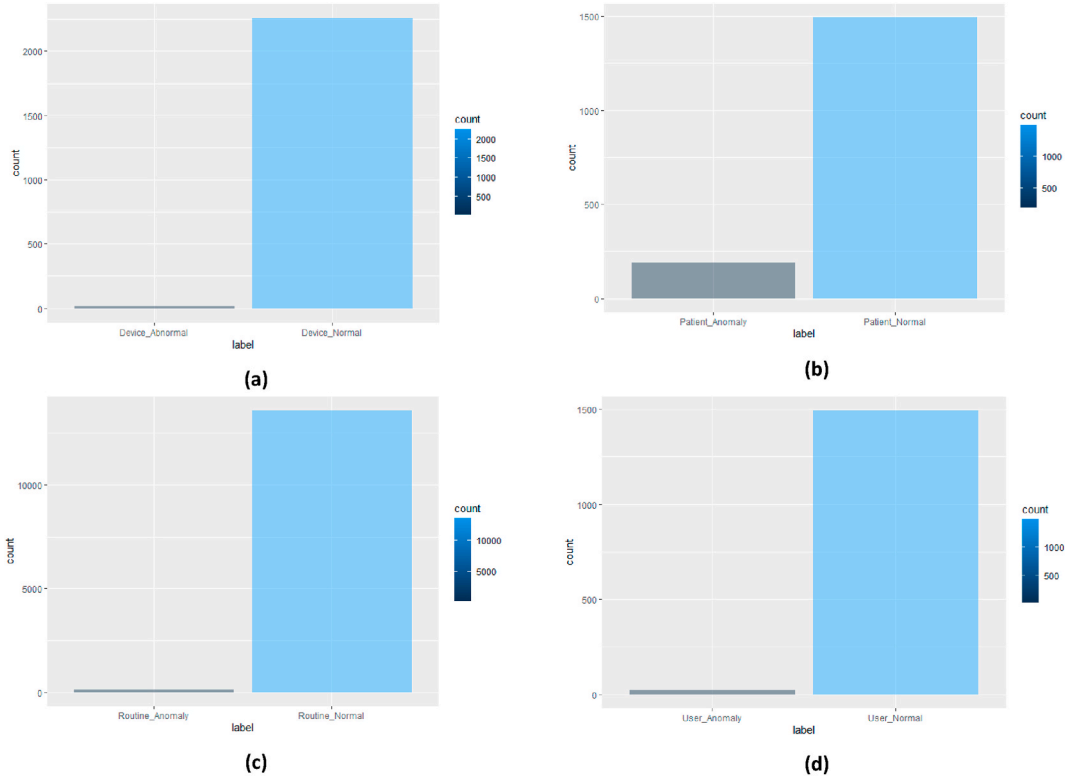

Fig. 5. Dataset Balance. a) Device Class b) Patient Class c) Routine Class d) User Class.

performance is visualized in Fig. 6; where (a) displays the R-square against the number of splits and (b) shows the relative error score.

The random decision forest is implemented using the randomForest R package. The classification results are outlined in Table 5; where the accuracy is 0.987487 and the error rate is 0.012513. The classifier produced 500 trees with 3 variables tried at each split. The random forest performance is outlined in Fig. 7. Where (a) displays the error score against the number of trees (where the black line indicates the Out-of-bag (OOB) error, the red and green line indicate the class errors) and (b) shows the prediction accuracy for each selected feature.

4.2. SVM experiment

SVM is implemented using the e1071 R package. If the data is classified as a whole with the 8 class labels outlined in Table 2 (Section 3.1), without converting the labels to a binary normal/abnormal format, the real-time implementation of the detection process is greatly affected by the calculation of a large number of support vectors (6491), as outlined in Table 6. Given this data is only a snapshot of what would be a much larger dataset that would require continuous processing, an optimal support vector count is more appropriate.

The following outlines the results of the parameter tuning process on the entire dataset using the 10-fold cross validation sampling method, the best parameters are cost: 0.1, gamma: 0.5, with a best performance of 0.2819751 error; displayed in Table 7 and Fig. 8.

However, for the experiment, the parameters are as follows; the SVM-Type is a C-classification with a radial SVM-Kernel, with cost set to 100 and gamma set to 0.5. The parameter tuning is outlined in Table 8 and Fig. 9.

The results were the most successful when cost is set to 100 and gamma set to 0.5, where the best performance is 0.012461 error and 0.002730 dispersion with 582 support vectors. Fig. 10 displays a plot of the results achieved, with a visualisation of each of the predictions for normal compared with anomalous behaviour for each of the 9 features using the SVM classifier. Fig. 11 displays a plot of four of the most dominant features from the classification process. Fig. 11a displays the means vs the median and Fig. 11b shows the max vs min values. Normal data is represented by the red colour, with anomalies in black. With the circle representing correct

Table 4
Decision tree confusion matrix.

	Predicted Anomaly	Predicted Normal
Anomaly (True Anomaly)	38	1
Normal (True Normal)	69	5646

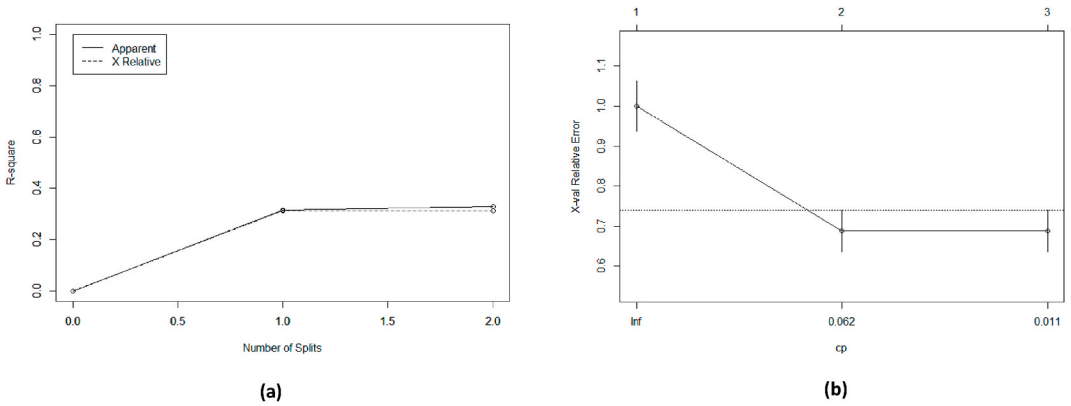

Fig. 6. a) R-squared against splits, b) Relative Error.

Table 5
Random forest confusion matrix.

	Anomaly	Normal
Anomaly (True Anomaly)	42	65
Normal (True Normal)	7	5640

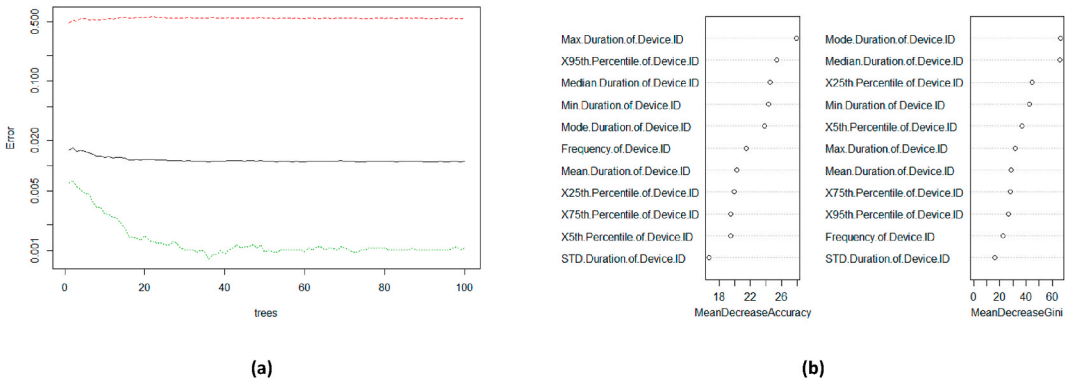

Fig. 7. (a) Error plot (b) prediction accuracy of individual features.

Table 6
SVM parameters.

Parameters	Anomaly Score
SVM-Type:	C-classification
SVM-Kernel:	radial
Cost:	1
Num. Support Vectors	6491

predictions and the cross showing incorrect predictions. The values on the x/y axis refers to seconds, following min-max normalisation.

To assess the effectiveness of the SVM classification, research standard performance evaluation metrics are adopted. Specifically, sensitivity, specificity, kappa and accuracy are considered. The results from the test set are presented in the confusion matrix in Table 9.

As the algorithm adopts a 70–30 split, the kappa provides a metric for the evaluation by normalizing the baseline of random chance in the dataset. Accuracy refers to the percentage of correct predictions out of all instances and is ideal for a binary classification problem. Sensitivity (recall) is the true-positive rate, so conversely, specificity is the true negative rate; in other words, sensitivity relates to how many instances from the positive class which are predicted correctly, whereas specificity relates to the number of instances from the negative class (second) predicted correctly. Balanced accuracy is an assessment of the average of the proportion corrects of each class and is useful to consider given that the dataset has an imbalance in the class distribution. Table 10 provides a

Table 7
SVM parameter tuning whole data – 10-fold cross validation.

Cost	Gamma	Error	Dispersion
0.1	0.5	0.281975	0.006078
1.0	0.5	0.282027	0.006009
10.0	0.5	0.282079	0.006126
100.0	0.5	0.282288	0.006133
0.1	1.0	0.282027	0.006001
1.0	1.0	0.282027	0.006009
10.0	1.0	0.282236	0.006147
100.0	1.0	0.282184	0.006099
0.1	2.0	0.282236	0.005976
1.0	2.0	0.282236	0.005998
10.0	2.0	0.282132	0.006046
100.0	2.0	0.282079	0.006221

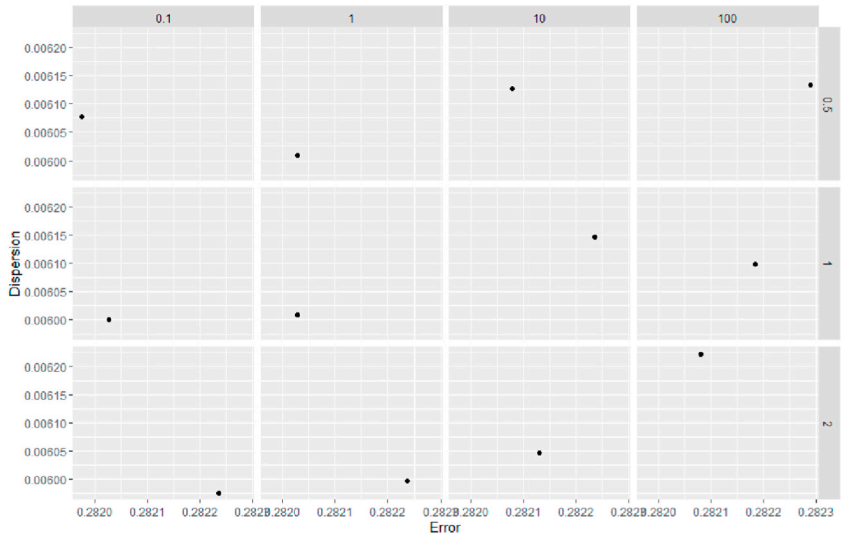

Fig. 8. Cost vs Gamma Results Plot all Data.

Table 8
SVM parameter tuning.

Cost	Gamma	Error	Dispersion
0.1	0.5	0.016216	0.002719
1.0	0.5	0.012722	0.002359
10.0	0.5	0.012670	0.002306
100.0	0.5	0.012461	0.002730
0.1	1.0	0.017519	0.002532
1.0	1.0	0.013452	0.002417
10.0	1.0	0.012826	0.002603
100.0	1.0	0.013191	0.002613
0.1	2.0	0.018562	0.002435
1.0	2.0	0.013869	0.002591
10.0	2.0	0.013087	0.002559
100.0	2.0	0.012879	0.002624

breakdown of the performance evaluation for the classifier.

4.3. Discussion

The best results in the test set included 0.9896 accuracy, with a 0.6108 kappa and 0.9846 balance accuracy. The sensitivity and specificity are 0.9796 and 0.9897 respectively. With regards to the validation of the approach on the raw dataset, the trained classifier is able to detect with a 0.9908 accuracy, 0.6722 and 0.9823 balanced accuracy. As such, these results validate the use an SVM for an

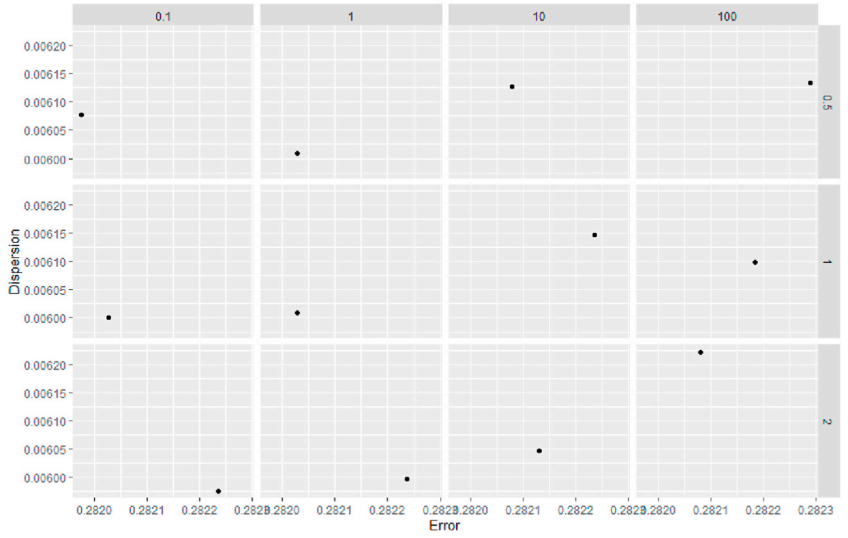

Fig. 9. Cost vs Gamma Results Plot Exp. 1.

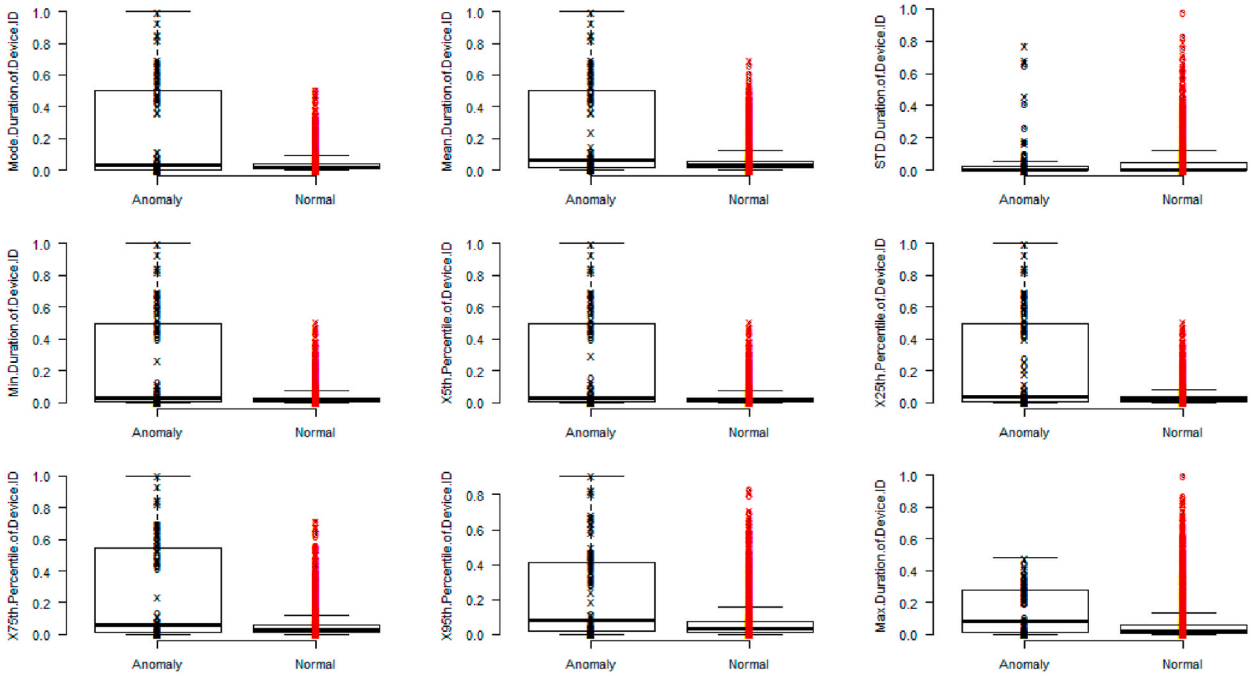

Fig. 10. Boxplot of SVM Individual Predictions for Normal vs Anomaly for Each Feature.

internal-facing anomaly detection system. The decision tree classification performs with high accuracy (0.98783), however, it has a tendency to predict normal values as anomalies.

Where the decision tree sensitivity is 0.9998, and specificity is 0.9886. The random forest, whilst also having a high classification accuracy (0.98783), has the inverse in that it is prone to predicating a high level of anomalies as normal. The resulting sensitivity and specificity are 0.97436 and 0.392523, respectively. This would not be ideal for an anomaly detection method as the algorithm would be likely to miss anomalies. A plot of the results comparison is displayed in Fig. 12. The SVM also classifiers some normal behaviour as anomalies; however, at a more optimal rate than the decision tree, which is why it has been selected as a suitable approach (but the difference is minor). The SVM scores a higher specificity of 0.9897 compared with the 0.9886 of the decision tree process, in addition to a higher overall accuracy score. It can be surmised that performance of the SVM is due to creation of appropriate feature spaces by tuning the kernel in Table 6. The aforementioned work by Saha et al. differs to the approach put forward in this paper, in that the focus is on signature exchange with high transaction time (Saha et al., 2019). Further, as previously mentioned in Section 2, Menon et al.

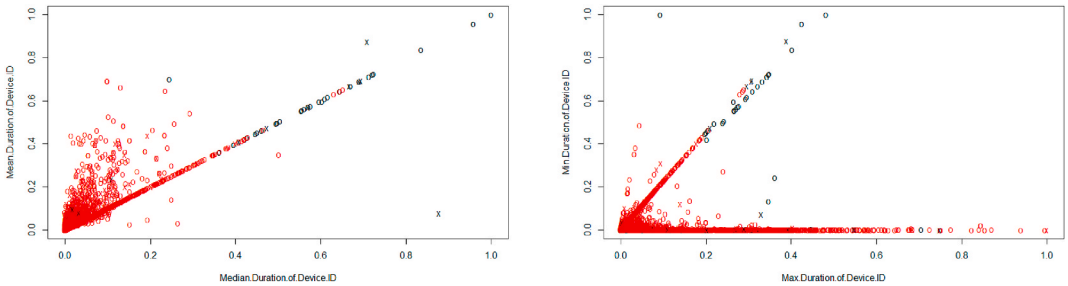

Fig. 11. a) Mean vs Median Feature Plot. b) Max vs Min Feature Plot.

Table 9
SVM parameters.

	Anomaly	Normal
Anomaly (True Anomaly)	48	1
Normal (True Normal)	59	5646

Table 10
SVMModel Performance (Positive class: Anomaly).

	Test Data	Validation Data
Parameters	Performance	Performance
Accuracy:	0.9896	0.9908
Kappa	0.6108	0.6722
Sensitivity	0.9796	0.9737
Specificity	0.9897	0.9909
Pos. Predicted Value	0.4486	0.5182
Neg. Predicted Value	0.9998	0.9997
Balanced Accuracy	0.9846	0.9823

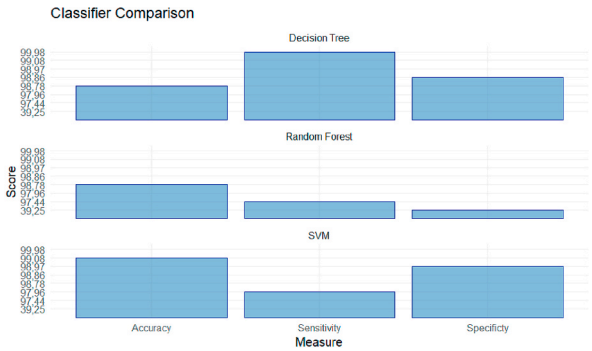

Fig. 12. Classifier comparison plot.

adopt an SVM methodology, with successful results of 0.9658 accuracy documented in (Menon et al., 2014), yet this is lower than the 0.9908 in this research. This score is boosted by collaborative filtering to 0.9900, and the 0.9908 in this article is achieved without collaborative filtering, with a reduced feature and anomaly label set during training. However, it should be clarified that the dataset employed in this article is different to that which Menon et al. adopt in their approach. Direct comparison of the results is therefore not possible. For example, the dataset by Menon is 34x larger and integrates two data sources to serve as panel data for the experimentation. Their focus is also on access logs and association rule mining. Whereas the dataset employed in this article has granularity and offers the ability to extract the 4 data groupings for analysis in the machine learning approach. Despite it not being possible to compare the findings of the two approaches, both scores provide a benchmark of the potential machine learning has for enhancing the access security concerning EHR.

5. Conclusion

EHR is making the healthcare industry more efficient, reducing costs and helping the environment. However, with the many benefits brought about by its introduction, new risks have also emerged. Advanced anomaly detection approaches are paramount to maintaining confidentiality, integrity and availability of the data. In this paper, the authors presented an anomaly detection approach, which assessed the effectiveness of three supervised learning classifiers for the identification of anomalous data access. Based on the results achieved, the recommendation of this work is to adopt an SVM for the detection process. The unique focus of this work is on the detection of insider threat based on anomalous behavioural patterns when using the EHRs. To the best of our knowledge, this is the first time this methodology has been applied to the dataset used in this paper with a focus on insider threat detection. Therefore, in future work, we will assess the effectiveness of the approach when operating in real time and compare with deep learning approaches to improve the overall accuracy of the anomaly detection prediction. Both a larger data set and an improved balance (between anomaly and normal readings) will also be considered and compared with the results achieved in this manuscript.

Author Credit

William Hurst: Investigation, Methodology, Software, Supervision, Visualisation; William Hurst, Aaron Boddy: Conceptualization, Data curation, Formal analysis, Methodology, Project administration, Validation, Writing – original draft; William Hurst, Bedir Tekinerdogan, Tarek Alskaif, Aaron Boddy and Nathan Shone: Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

Adnan Qayyum, J. Q., Bilal, M., & Al-Fuqaha, A. (2020). *Secure and robust machine learning for healthcare: A survey*. arXiv, no. arxiv.org/abs/2001.08103.

Al-Zubaidie, M., Zhang, Z., & Zhang, J. (2019). PAX: Using pseudonymization and anonymization to protect patients' identities and data in the healthcare system. *MDPI Int. J. Environ. Res. Public Health, 16*(9), 1490.

Anakath, A. S., Kannadasan, R., Joseph, N. P., Boominathan, P., & Sreekanth, G. R. (2022). Insider attack detection using deep belief neural network in cloud computing. *Computer Systems Science and Engineering, 41*(2), 479–492.

Bujnowska-Fedak, M. M., & Wysoczański, Ł. (2020). Access to an electronic health record: A polish national survey. *MDPI Int. J. Environ. Res. Public Health, 17*(17), 6165.

Cano, I., Tenyi, A., Vela, E., Miralles, F., & Roca, J. (2017). Perspectives on Big Data applications of health information. *Current Opinion in Structural Biology, 3*, 36–42.

Cecil, Elizabeth, Dewa, Lindsay, Ma, Richard, et al. (2019). *Primary health care professionals views of reminders in electronic patient records*. J. Epidemiol. Community Health.

Cook, D. J., Duncan, G. E., Sprint, G., & Fritz, R. (2018). Using smart city technology to make healthcare smarter. *Proceedings of the IEEE, 106*(4), 708–722.

Evans, R. S. (2016). Electronic health records: Then, now, and in the future. *Yearb. Med. Inform, 1*, 48–61.

Gallagher, David, Zhao, Congwen, Brucker, Amanda, et al. (2020). Implementation and continuous monitoring of an electronic health record embedded readmissions clinical decision support tool. *Journal of Personalisaed Medicine, Special Issue Use of Clinical Decision Support Software within Health Care Systems), 10*(3), 103.

Gil, M., Sherif, R. E., Pluye, M., Fung, B. C. M., Grad, R., & Pluye, P. (2019). Towards a knowledge-based recommender system for linking electronic patient records with continuing medical education information at the point of care. *IEEE Access, 7*, 15955–15966.

Girault, A. (2015). Internet-based technologies to improve cancer care coordination: Current use and attitudes among cancer patients. *European Journal of Cancer, 51*(4), 551–557.

Hang, L., Choi, E., & Kim, D. (2019). A novel EMR integrity management based on a medical blockchain platform in hospital. *MDPi Electronics, 8*(4), 467.

Hurst, W. (2021). Electronic patient record dataset - UK hospital. *Easy Dans*. https://doi.org/10.17026/dans-znf-sh4q

Hurst, W., Boddy, A., Merabti, M., & Shone, N. (2020). Patient privacy violation detection in healthcare critical infrastructures: An investigation using density-based benchmarking. *MDPI Future Internet, 12*(6), 100.

Ihlebæk, H. M. (2020). Lost in translation - silent reporting and electronic patient records in nursing handovers: An ethnographic study. *International Journal of Nursing Studies, 109*, Article 103636.

Jacquemard, T., Doherty, C. P., & Fitzsimons, M. B. (2020). Examination and diagnosis of electronic patient records and their associated ethics: A scoping literature review. *BMC Medical Ethics, 21*, 76.

Janjua, F., Masood, A., Abbas, H., & Rashid, I. (2020). Handling insider threat through supervised machine learning techniques. *Procedia Computer Science, 177*, 64–71.

Jose, T., Hays, J. T., & Warner, D. O. (2020). Improved documentation of electronic cigarette use in an electronic health record. *MDPI Environment Research and Public Health, 17*(16), 5908.

Khennou, F., Khamlichi, Y. I., & Chaoui, N. E. H. (2018). Improving the use of big data analytics within electronic health records: A case study baseD OpenEHR. *Procedia Computer Science, 127*, 60–68.

Kruse, C. S., Smith, B., Vanderlinden, H., & Nealand, A. (2017). Security techniques for the electronic health records. *Journal of Medical Systems, 41*, 127.

Ledford, H. (2019). Millions of black people affected by racial bias in health-care algorithms. *Nature, 574*, 608.

Li, J.-h. (2018). Cyber security meets artificial intelligence: A survey. *Frontiers of Information Technology & Electronic Engineering, 19*, 1462–1474.

Lin, W. C., Chen, J. S., Chiang, M. F., & Hribar, M. R. (2020). Applications of artificial intelligence to electronic health record data in ophthalmology. *Transl. Vis. Sci. Technol, 9*(2), 13.

Liveri, Dimitra, Sarri, Anna, & Skouloudi, Christina (2015). *Security and resilience in eHealth*. Brussels: European Union Agency for Network and Information Security.

Medhi, D., & Huang, D. (2008). Secure and resilient routing: Building blocks for resilient network architectures," information assurance. *The Morgan Kaufmann Series in Networking*, 417–448 (Chapter 14).

Mehta, N. (2018). Machine learning, natural language programming, and electronic health records: The next step in the artificial intelligence journey? *The Journal of Allergy and Clinical Immunology, 141*(6).
Menon, A. K., Jiang, X., Kim, J., Vaidya, J., & Ohno-Machado, L. (2014). Detecting inappropriate access to electronic health records using collaborative filtering. *Machine Learning, 95*, 87–101.
Millard, W. B. (2017). Where bits and bytes meet flesh and blood. *Annals of Emergency Medicine, 70*(3), 1–17.
Misra, S., Dyer, F., & Sandler, P. P. (2017). Persecution complex: I am getting one over electronic patient records. *Florida Dental Journal, 8*(2), 78–81.
Parah, S. A., Sheikh, J. A., Akhoon, J. A., & Loan, N. A. (2020). Electronic health record hiding in images for smart city applications: A computationally efficient and reversible information hiding technique for secure communication. *Future Generation Computer Systems, 108*, 935–949.
Peckham, D. (2016). Electronic patient records, past, present and future. *Paediatric Respiratory Reviews, 20*, 8–11.
Perumal, A. M., & Nadar, E. R. S. (2020). Architectural framework of a group key management system for enhancing e-healthcare data security. *Healthcare Technology Letters, 7*, 13–17.
Pho, K. K. (2019). Mobile device applications for electronic patient portals in oncology. *JCO Clin. Cancer Informatics, 3*, 1–8.
Saha, R., Kumar, G., Rai, M. K., Thomas, R., & Lim, S. (2019). Privacy ensured e-healthcare for fog-enhanced IoT based applications. *IEEE Access, 7*, 44536–44543.
Sanchez-Guerrero, Rosa, Almenarez Mendoza, Florina, Diaz-Sanchez, Daniel, et al. (2017). Collaborative eHealth meets security: Privacy-enhancing patient profile management. *IEEE Journal of Biomedical and Health Informatics, 21*(6), 1741–1749.
Schwartz, J. T., Gao, M., Geng, E. A., Mody, K. S., Mikhail, C. M., & Cho, S. K. (2019). Applications of machine learning using electronic medical records in spine surgery. *Neurospine, 16*, 643–653.
Seh, A. H., Al-Amri, J. F., Subahi, A. F., Agrawal, A., Pathak, N., Kumar, R., & Khan, R. A. (2021). An analysis of integrating machine learning in healthcare for ensuring confidentiality of the electronic records. *Computer Modeling in Engineering and Sciences, 130*(3), 1387–1422.
Shen, B., Guo, J., & Yang, Y. (2019). MedChain: Efficient healthcare data sharing via blockchain. *MDPi Applied Sciences, 9*(6), 1207.
Shi, S., He, D., Li, L., Kumar, N., Khan, M. K., & Choo, K. K. R. (2020). Applications of blockchain in ensuring the security and privacy of electronic health record systems: A survey. *Computers & Security, 97*, Article 101966.
Shiells, Kate, Alejandra Diaz Baquero, Angie, Štěpánková, Olga, & Holmerová, Iva (2020). Staff perspectives on the usability of electronic patient records for planning and delivering dementia care in nursing homes: A multiple case study. *BMC Medical Informatics and Decision Making, 20*, 1–14.
Shoolin, J. S. (2017). Clinical decision support and the electronic health record—applications for physiatry. *Elsevier PM and R, 9*(5), 34–40.
Tanwar, S., Parekh, K., & Evans, R. (2020). Blockchain-based electronic healthcare record system for healthcare 4.0 applications. *Journal of Information Security and Applications, 50*, Article 102407.
Tort, C. G., Pulido, V. A., Ulloa, V. S., Boedo, F. D., Gestal, J. M. L., & Loureiro, J. P. (2020). Electronic health records exploitation using artificial intelligence techniques. In *MDPI 3rd XoveTIC conference*. Spain: A Coruña.
Wang, X., & Jin, Z. (2019). An overview of mobile cloud computing for pervasive healthcare. *IEEE Access, 7*, 667744, 66791.
Wesołowski, T. E., Porwik, P., & Doroz, R. (2016). Electronic health record security based onensemble classification of keystroke dynamics. *Applied Artificial Intelligence, 30*(6), 521–540.
Zheng, Tao, Xie, Wei, Xu, Liling, He, Xiaoying, et al. (2017). A machine learning-based framework to identify type 2 diabetes through electronic health records. *International Journal of Medical Informatics, 97*, 120–127.
Zhou, X., Liu, Q. W. J., & Zhang, Z. (2016). Privacy preservation for outsourced medical data with flexible access control. *IEEE Access, 6*, 14827–14841.