Securing Electronic Health Records against Insider-Threats: A Supervised Machine Learning Approach

William Hursta[[1]](#footnote-1)\*, Bedir Tekinerdogana, Tarek Alskaifa, Aaron Boddyb and Nathan Shonec

a Information Technology Group, Wageningen University and Research, Leeuwenborch, Hollandseweg 1, 6706 KN Wageningen, the Netherlands, will.hurst@wur.nl, bedir.tekinerdogan@wur.nl, tarek.alskaif@wur.nl

b Aintree Hospital, Liverpool, L9 7AL, United Kingdom, dr.aaronboddy@gmail.com

c Department of Computer Science, Liverpool John Moores University, Liverpool, L3 3AF, United Kingdom, n.shone@ljmu.ac.uk

\* will.hurst@wur.nl

|  |  |
| --- | --- |
| ARTICLE INFO | ABSTRACT |
| **Keywords**:Electronic Health RecordMachine LearningAnomaly DetectionSecurity | 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 [1]. 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 [2]. Their core benefits, as presented in Figure 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 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 [3]. Furthermore, data mining of historical records enables medical staff to examine patient conditions with high flexibility [4] [5].



**Fig. 1.** Electronic Health Record Benefits

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 [6]. 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 [3]. 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 example is disgruntled employees sharing their access credentials with external parties.

Argaw et al. discuss that the healthcare industry is among the top three sectors to be most affected by ransomware attacks [7]. 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 [8]. High profile cyber-attacks, such as the well-documented WannaCry ransomware (which affected more than 100 countries [8]), demonstrates the potential successful attacks have to disrupt access to millions of records of patients. Thus, 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. There is a need for advanced anomaly detection approaches to ensure privacy, integrity and access to EHR data. 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. To the best of our knowledge, this is the first time this approach has been conducted on this dataset and the use of the four data groups from within the EHR dataset (routine, user, device and patient) makes this work stand apart from other approaches. To substantiate the investigation, the data used for the analysis is provided by a UK-based hospital, with the anomalous readings pre-labelled by means of density-based classification. 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 [9] [10]; with kNN, Decision Tree and Random Forest as notable approaches for an anomaly detection process. The research in this article is organised as follows. Section 2 provides a background discussion of EHR data and its applications. Section 3 details the research methodology with the results outlined in Section 4. The paper is concluded in Section 5.

1. Background

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 [2], by facilitating healthcare efficiency to provide timely access to comprehensive and organised patient information; thus enabling the delivery of high quality patient care [11]. 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 [12]. 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’ [13]. Features of EHRs include electronic diaries, automated text messages, clinic letter verification, prescriptions and the reviewing of investigative results [14]. 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 [15].

* 1. EHR and Security

EHR portals enable patients to access their personal health information [16], view laboratory and imaging results, contact clinical staff with questions and updates, schedule future appointments and request medication refills [17]. 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 [18] and the widespread adoption of EHR has resulted in the collection of a significant growing volume of clinical data [19]. There is now a corresponding interest in exploiting analytical applications regarding this granular data repository [20]; where emerging secondary applications include [21] 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 [22]; 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 [23].

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 [14]. 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 [24].

However, there are various risks with the digitisation of health records. The introduction of EHR systems can also result in unintended negative consequences [25]. 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 [26]. With more serious implications, many EHR applications also do not follow standardised protocols for data storage and application programming interfaces [27]. This makes it challenging to interface EHRs with other clinical systems and impacts on data extraction [28]. 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 [29].

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 [8]; 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 [30]. New systems are needed for safeguarding EHRs. A combination of behavioural variables and standard security measures should be employed to safeguard systems.

* 1. EHR Related Work

The mainstay technology currently in place involves the use of a firewall fort safeguarding measures. As Kruse et al. discuss, the technique is a proven approach for maintaining the health of the network [31]. 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 [31]. 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 [8], 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 are not yet suitable for mass deployment of EHR usage [32]; and 3) there are insufficient techniques to manage, store and control the sensitive data [32]. 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 [33]. Wang et al., therefore, discuss the general architecture of using mobile cloud in digital health care environments [33]. 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.

The increasing level of risks to EHR systems have led to numerous research investigations into novel techniques to protect both the systems and the information within the digital records. Many of which differ to the approach put forward in this article. 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 [36]. Their technique involves the use of a 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 [36]. This differs from the work poised in this paper, 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 (referred to as electronical medical records in this instance) [37]. Their approach employs a combination of multiple end-user devices. A selection of which include 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 [37]. The framework is tested within a cloud-fog infrastructure, in which the framework demonstrates that it has a reduced transaction time when compared with similar approaches.

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 [39]. 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 [40] 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 [41] [42] [43], 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 paper, 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. As a final note, existing work closely aligned to this investigation is provided by Menon et al., (who adopt an SVM methodology [38]). Thus, a comparison of the results achieved compared with the approach my Menon et al., is provided in Section 4.3.

1. 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 Figure 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.



**Fig. 2.** EHR Disaggregation

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 [34].

**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 | 14067 | 39 |
| 28/02/16 00:04 | 845 | 1489 | Pharmacy Orders | 49304 | 22 |
| 28/02/16 00:06 | 923 | 199 | Recent Clinical Results Recent Clinical Results:(Departmental Reports) UK.View Orders | 60948 | 165 |
| 28/02/16 00:08 | 775 | 568 | Patient Care Notes | 32826 | 75 |
| 28/02/16 00:10 | 748 | 797 | Recent Clinical Results Recent Clinical Results:(Departmental Reports) | 2166 | 20 |

* 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 [34]. The anomalous points are validated through a cross-check with the corresponding data record [34]. Once labelled, a feature extraction process 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 Figure 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. 3.** Anomaly Detection Process

**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 >= all other values in the ID range |
| Min | The data value that is <= 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 (<= 75% of values in the dataset) |
| 75th Percentile | Third quartile (>= 75% of values in the dataset) |
| 95th Percentile | >= 95% of values in the dataset |
| Outlier Score | The local outlier score as labelled in [42] |

**Table 3** - Example of labelled data

|  |  |
| --- | --- |
| **Record** | **Label** |
| 1082 | Anomaly |
| 1734 | Normal |
| 24603 | Anomaly |
| 805 | Normal |
| 1510 | Normal |
| 706 | Anomaly |
| 12917 | Anomaly |
| 11370 | Normal |

Figure 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 Figure 5.



**Fig. 4.** Normal and Abnormal Dataset Distribution

|  |  |
| --- | --- |
| **(a)** | **(b)** |
| **(c)** | **(d)** |

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

* 1. 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 tress 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 [35]. 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.

1. Results

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

* 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 250/13425 nodes). The results are displayed in Table 4.

**Table 4** – Decision Tree Confusion Matrix

|  |  |  |
| --- | --- | --- |
|  | **Predicted Anomaly** | **Predicted Normal** |
| **Anomaly (True Anomaly)** | 38 | 1 |
| **Normal (True Normal)** | 69 | 5646 |

As presented in the above confusion matrix, the accuracy is 0.98783 and misclassification error rate is 0.01216. The classifier’s performance is visualized in Figure 6; where (a) displays the R-square against the number of splits and (b) shows the relative error score.

|  |  |
| --- | --- |
| **A close up of a map  Description automatically generated****(a)** | **A screenshot of a social media post  Description automatically generated****(b)** |

**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 |

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 Figure 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.

|  |  |
| --- | --- |
| **A screenshot of a map  Description automatically generated****(a)** | **A screenshot of a cell phone  Description automatically generated****(b)** |

**Fig 7**. (a) Error Plot (b) Prediction Accuracy of Individual Features

* 1. 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.

**Table 6** – SVM Parameters

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

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 Figure 8.

**Table 7** – SVM Parameter Tuning Whole Data – 10-Fold Cross Validation

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Cost** | **Gamma** | **Error** | **Dispersion** | A picture containing group, water, large, lot  Description automatically generated**Fig. 8**. Cost vs Gamma Results Plot all Data |
| 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 |

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 Figure 9.

**Table 8** – SVM Parameter Tuning

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Cost** | **Gamma** | **Error** | **Dispersion** | **Fig. 9**. Cost vs Gamma Results Plot Exp. 1 |
| 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 |

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. Figure 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. Figure 11 displays a plot of four of the most dominant features from the classification process. Figure 11a displays the means vs the median and Figure 11b shows the max vs min values. Normal data is represented by the red colour, with anomalies in black. With the circle representing correct predictions and the cross showing incorrect predictions. The values on the x / y axis refers to seconds, following min-max normalisation.



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

|  |  |
| --- | --- |
| A close up of a map  Description automatically generated | A close up of a map  Description automatically generated |

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

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.

**Table 9** – SVM Parameters

|  |  |  |
| --- | --- | --- |
|  | **Anomaly** | **Normal** |
| **Anomaly (True Anomaly)** | 48 | 1 |
| **Normal (True Normal)** | 59 | 5646 |

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 breakdown of the performance evaluation for the classifier.

**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 |

* 1. 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 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.



**Fig. 12**. Classifier Comparison Plot

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 Figure 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. Further, as previously mentioned in Section 2, Menon et al. adopt an SVM methodology, with successful results of 0.9658 accuracy documented in [38], yet this is lower than the 0.9908 in this research. This score is boosted by collaborative filtering to 0.9900, however the 0.9908 in this article is achieved without collaborative filtering, with a reduced feature and anomaly label set during training.

1. 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.

References

1. S. A.Parah, J. A.Sheikh, J. A.Akhoon and N. A.Loan, 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, vol. 108, pp. 935-949, 2020.
2. M. M. Bujnowska-Fedak and Ł. Wysoczański, Access to an Electronic Health Record: A Polish National Survey,” MDPI Int. J. Environ. Res. Public Health , vol. 17, nr. 17, p. 6165, 2020.
3. C. G. Tort, V. A. Pulido, V. S. Ulloa, F. D. Boedo, J. M. L. Gestal en J. P. Loureiro, Electronic Health Records Exploitation Using Artificial Intelligence Techniques, in MDPI 3rd XoveTIC Conference, A Coruña, Spain, 2020.
4. D. J. Cook, G. E. Duncan, G. Sprint and R. Fritz, Using Smart City Technology to Make Healthcare Smarter,” Proceedings of the IEEE, vol. 106, nr. 4, pp. 708-722, 2018.
5. T. Jose, J. T. Hays and D. O. Warner, Improved Documentation of Electronic Cigarette Use in an Electronic Health Record, MDPI Environment Research and Public Health, vol. 17, nr. 16, p. 5908, 2020.
6. D. Gallagher, C. Zhao, A. Brucker, J. Massengill, P. Kramer, E. G. Poon and B. A. Goldstein, Implementation and Continuous Monitoring of an Electronic Health Record Embedded Readmissions Clinical Decision Support Tool, MDPI Journal of Personalisaed Medicine, Special Issue Use of Clinical Decision Support Software within Health Care Systems), vol. 10, nr. 3, p. 103, 2020.
7. S. T. Argaw, N.-E. Bempong, B. Eshaya-Chauvin and A. Flahault, The state of research on cyberattacks against hospitals and available best practice recommendations: a scoping review, BMC Medical Informatics and Decision Making, vol. 19, p. 10, 2019.
8. W. B. Millard, Where Bits and Bytes Meet Flesh and Blood, Annals of Emergency Medicine, vol. 70, nr. 3, pp. 1-17, 2017.
9. N. Mehta, Machine learning, natural language programming, and electronic health records: The next step in the artificial intelligence journey?, Journal of Allergy and Clinical Immunology, vol. 141, nr. 6, 2018.
10. T. Zheng, W. Xie and L. Xu, A machine learning-based framework to identify type 2 diabetes through electronic health records, International Journal of Medical Informatics, vol. 97, pp. 120-127, 2017.
11. C. P. D. a. M. B. F. T. Jacquemard, Examination and diagnosis of electronic patient records and their associated ethics: a scoping literature review, BMC Med. Ethics, vol. 21, p. 76, 2020.
12. S. Shi, D. He, L. Li, N. Kumar, M. K. Khan and K. K. R. Choo, „Applications of blockchain in ensuring the security and privacy of electronic health record systems: A survey,” Computers and Security, vol. 97, p. 101966, 2020.
13. A. A. D. B. O. Š. a. I. H. K. Shiells, Staff perspectives on the usability of electronic patient records for planning and delivering dementia care in nursing homes: A multiple case study, BMC Med. Inform. Decis. Mak, vol. 20, pp. 1-14, 2020.
14. S. Misra, F. Dyer and P. P. Sandler, Persecution complex: I am getting one over electronic patient records!, Fac. Dent. J, vol. 8, nr. 2, pp. 78-81, 2017.
15. H. M. Ihlebæk, Lost in translation - Silent reporting and electronic patient records in nursing handovers: An ethnographic study, Int. J. Nurs. Stud., vol. 109, p. 103636, 2020.
16. K. K. Pho, Mobile Device Applications for Electronic Patient Portals in Oncology, JCO Clin. Cancer Informatics, vol. 3, pp. 1-8, 2019.
17. A. Girault, Internet-based technologies to improve cancer care coordination: Current use and attitudes among cancer patients, Eur. J. Cancer, vol. 51, nr. 4, pp. 551-557, 2015.
18. S. Tanwar, K. Parekh and R. Evans, Blockchain-based electronic healthcare record system for healthcare 4.0 applications, J. Inf. Secur. Appl., vol. 50, p. 102407, 2020.
19. W. C. Lin, J. S. Chen, M. F. Chiang and M. R. Hribar, Applications of artificial intelligence to electronic health record data in ophthalmology, Transl. Vis. Sci. Technol, vol. 9, nr. 2, p. 13, 2020.
20. F. Khennou, Y. I. Khamlichi and N. E. H. Chaoui, Improving the use of big data analytics within electronic health records: A case study baseD OpenEHR, Procedia Computer Science, vol. 127, pp. 60-68, 2018.
21. I. Cano, A. Tenyi, E. Vela, F. Miralles and J. Roca, Perspectives on Big Data applications of health information, Current Opinion in Systems Biology, vol. 3, pp. 36-42, 2017.
22. J. S. Shoolin, Clinical Decision Support and the Electronic Health Record—Applications for Physiatry, Elsevier PM and R, vol. 9, nr. 5, pp. 34-40, 2017.
23. M. Gil, R. E. Sherif, M. Pluye, B. C. M. Fung, R. Grad and P. Pluye, Towards a Knowledge-Based Recommender System for Linking Electronic Patient Records with Continuing Medical Education Information at the Point of Care, IEEE Access, vol. 7, pp. 15955-15966, 2019.
24. D. Peckham, Electronic patient records, past, present and future, Paediatric Respiratory Reviews, vol. 20, pp. 8-11, 2016.
25. T. Jacquemard, C. P. Doherty and M. B. Fitzsimons, Examination and diagnosis of electronic patient records and their associated ethics: a scoping literature review, BMC Med. Ethics, vol. 21, p. 76, 2020.
26. E. Cecil, L. Dewa, R. Ma, A. Majeed, R. Glover and A. Fraser, Primary health care professionals views of reminders in electronic patient records, J. Epidemiol. Community Heal, 2019.
27. R. S. Evans, Electronic Health Records: Then, Now, and in the Future, Yearb. Med. Inform., vol. 1, pp. 48-61, 2016.
28. J. T. Schwartz, M. Gao, E. A. Geng, K. S. Mody, C. M. Mikhail and S. K. Cho, Applications of machine learning using electronic medical records in spine surgery, Neurospine, vol. 16, pp. 643-653, 2019.
29. H. Ledford, Millions of black people affected by racial bias in health-care algorithms, Nature, vol. 574, p. 608, 2019.
30. D. Liveri, A. Sarri and C. Skouloudi, Security and Resilience in eHealth, European Union Agency for Network and Information Security, Brussels, 2015.
31. C. S. Kruse, B. Smith, H. Vanderlinden and A. Nealand, Security Techniques for the Electronic Health Records, Journal of Medical Systems, vol. 41, p. 127, 2017.
32. R. Sánchez-Guerrero, F. A. Mendoza, D. Díaz-Sánchez, P. A. Cabarcos and A. M. López, Collaborative eHealth Meets Security: Privacy-Enhancing Patient Profile Management, IEEE Journal of Biomedical and Health Informatics, vol. 21, nr. 6, pp. 1741-1749, 2017.
33. X. Wang and Z. Jin, An Overview of Mobile Cloud Computing for Pervasive Healthcare, IEEE Access, vol. 7, pp. 66774-66791, 2019.
34. A. M. Perumal and E. R. S. Nadar, Architectural framework of a group key management system for enhancing e-healthcare data security, Healthcare Technology Letters, vol. 7, pp. 13-17, 2020.
35. R. Saha, G. Kumar, M. K. Rai, R. Thomas and S. Lim, Privacy Ensured e-Healthcare for Fog-Enhanced IoT Based Applications, IEEE Access, vol. 7, pp. 44536-44543, 2019.
36. X. Zhou, J. Liu, Q. Wu and Z. Zhang, Privacy Preservation for Outsourced Medical Data With Flexible Access Control, IEEE Access, vol. 6, pp. 14827-14841, 2018.
37. M. Al-Zubaidie, Z. Zhang and J. Zhang, PAX: Using Pseudonymization and Anonymization to Protect Patients’ Identities and Data in the Healthcare System, MDPI Int. J. Environ. Res. Public Health, vol. 16, nr. 9, p. 1490, 2019.
38. A. K. Menon, X. Jiang, J. Kim, J. Vaidya and L. Ohno-Machado, “Detecting inappropriate access to electronic health records using collaborative filtering,” Machine Learning, vol. 95, pp. 87-101, 2014.
39. M. Kim, S. Yu, J. Lee, Y. Park and Y. Park, Design of Secure Protocol for Cloud-Assisted Electronic Health Record System Using Blockchain, MDPI Sensors, vol. 20, nr. 10, p. 2913, 2020.
40. L. Hang, E. Choi and D.-H. Kim, A Novel EMR Integrity Management Based on a Medical Blockchain Platform in Hospital, MDPi Electronics, vol. 8, nr. 4, p. 467, 2019.
41. B. Shen, J. Guo and Y. Yang, MedChain: Efficient Healthcare Data Sharing via Blockchain, MDPi Applied Sciences, vol. 9, nr. 6, p. 1207, 2019.
42. W. Hurst, A. Boddy, M. Merabti and N. Shone, Patient Privacy Violation Detection in Healthcare Critical Infrastructures: An Investigation Using Density-Based Benchmarking, MDPI Future Internet, vol. 12, nr. 6, p. 100, 2020.
43. G. James, D. Witten, T. Hastie and R. Tibshirani, An Introduction to Statistical Learning (8th Edition), New York : Springer, 2017.
44. M. J. Vidya and K. V. Padmaja, Affirmation of electronic patient record through bio-electric signal for medical data encryption authenticity, in 10 Annual International Conference, Proceedings/TENCON, 2017.
1. \* Corresponding author. Tel.: +0-000-000-0000; fax: +0-000-000-0000; e-mail: author@university.countryacronym [↑](#footnote-ref-1)