Comparative analysis of machine learning methods for analyzing security practice in electronic health records' logs.

Prosper K Yeng
Department of Information Security
and Communication Technology
NTNU
Gjøvik, Norway
prosper.yeng@ntnu.no

2nd Muhammad Ali Fauzi
Department of Information Security
and Communication Technology
NTNU
Gjøvik, Norway
muhammad.a.fauzi@ntnu.no

3rd Bian Yang
Department of Information Security
and Communication Technology
NTNU
Gjøvik, Norway
bian.yang@ntnu.no

Abstract—**Electronic health records (EHR) consists of broad, numerous and erratic accesses through self-authorizations and "brake the glass" scenarios. This is to fulfil the availability aspect of the the CIA (confidentiality, integrity) due to the time sensitive nature in healthcare especially during health emergency situations. Adversaries can use this as opportunity to illegitimately access patients records, thereby, compromising the entire EHR system.**

To avert this, a comparative analysis of machine learning classification methods was conducted with simulated EHR logs. The methods which were compared are Multinomial Naive Bayes(multnb), Bernoulli Naive Bayes (bernnb), Support Vector Machine (svm), Neural Network (nn), K-Nearest Neighbours(knn), Logistic Regression (lr), Random Forest (rf), and Decision Tree (dt).

The experiment results show that all of the machine learning models used in this work performed very well for the role classification task but, Decision Tree (dt) and Random Forrest (rf) obtained the best result among all of the methods with the same accuracy value of 0.889 on all three datasets. For the anomaly detection task, generally, our proposed approach obtained a high recall and accuracy but low precision and F1-score. Soft Classification approach performed better than the Hard Classification approach. The best performance was achieved with Bernoulli Naive Bayes with none normalised data, with an F1-score of 0.893.

Index Terms—**Electronic Health Records,logs,healthcare professionals,Machine learning,Security practice**

I. Introduction

Undermining required information security practice is in fact, a paradox to healthcare's objective. Healthcare professionals and major stakeholders (governments, non-governments, cares, and love ones) do put in all their efforts to save the lives of their subjects of care. In that vein, information systems are being relied on in recent times by hospitals to obtain better efficiency. This demands for the adoption of appropriate security measures (otherwise called required security practice) by the healthcare staff. Intentional or unintentional negligence in observing these required security practices tend to reverse the efforts of healthcare on patients' care since the sensitive patient records can be compromised. For instance, in a recent ransomware attack at Duesseldorf University Clinic in Germany, the medical records of a patient were not timely available during emergency and this resulted in the death of that patient [1].

Sound security practice involves all categories of the information systems' users who form the healthcare staff (including the healthcare professionals who provide therapeutic care and paramedical staffs such as health administrators, IT administrators, human resource personnel and finance) to follow laid down standards, policies, procedures, guidelines and code of conduct in the usage of the information systems in order to avoid compromising the confidentiality, integrity and availability (CIA) of the systems.

Good security practice is so much needed in healthcare because the healthcare data is classified as one of the most sensitive personal information [2] which is faced with multi-faceted threats. Such threats are masquerades (insiders, service providers, outsiders), communication interference, repudiation, misuse of system resources, system failures or errors, theft, damaging of resources and unauthorised access. Meanwhile, the healthcare systems are exposed to many users including their subjects of care, the healthcare professionals, contracted IT staff and locum personnel who are the weakest link in the security chain.

The critical importance in healthcare requires the sector to collect detailed patients information to enable them to correctly identify each patient and correctly map each patient to their medical records. This results in a collection of huge sensitive personal data which is of great importance to cyber criminals who can use it to commit multiple harm including identity theft [2].

Therefore, technological measures have since been the default and traditional approach in protecting these records. But these technical measures are being circumvented by the adversaries through the frequent manipulation of the healthcare workers to compromise these records. Due to the difficulties for cybercriminals to directly overcome the perimeters of technical security solutions, the healthcare workers are often masqueraded through social engineering attacks and other human related means of attack to gain unauthorised access. Insider intentional or unintentional security malpractice also tend to cause data breaches which can course serious harm to the patients.

In 2017, the healthcare sector in the United Kingdom had a bad experience with the wannacry ransomware which affected critical care [3], [4], spread to about 150 countries and affected about 230,000 computers in different sectors. Subsequently, about 3 million healthcare records were compromised in Norway in 2018 [5], [6] of which an insider aid was involved. Additionally, there was another phishing attack which resulted in compromising about 38,000 patient records in Portland, Oregon-based Legacy Health in the United States in 2019 [7]. The personal data which was comprised includes patients' email accounts, demographic information, dates of birth, health insurance data, billing details, medical data, Social Security numbers, and driver's licenses numbers. Healthcare data breaches continue to surge, with the passage of time. Globally, about 5 million healthcare records were compromised in 2017, followed by 15 million records in 2018 and 25 million records as at the middle of 2019 [8]. Quite recently, Universal Health Services (UHS), which is operating about 400 health facilities, was massively attacked with a ransomware and this was believed to possibly be the largest security incident in healthcare in the US [9]. The impact has led to a multi-day offline IT network across UHS facilities throughout the country. Information security incidents are threatening the quality of healthcare [10] delivery of which the information technology was rather to improve. Many bad actors use malicious emails, malspam and social engineering to make their way into the network, while some rely on exploiting vulnerabilities on Internet-facing devices.

As the saying goes that "an unexamined life is not worthy living", there is hence the need to assess the way of life of security practice of the human elements towards controlling these data breaches in healthcare. Good security practices have been defined in regulations, policies, standards, guidelines and code of conducts which are required to be implemented with both technical and non-technical measures. Technical security measures including firewalls, intrusion detection, and prevention have been fortified over time because they have since been the default and traditional security countermeasures. The challenging part is the human elements in the security chain who are the weakest link of which hackers mostly use in recent time to complete their attack.

In contributing to the fight against cyber attacks in healthcare, there is the need to understand the extend of users' compliance with the established security policies.For instance What are the challenges often faced by the healthcare workers in their effort to comply with these required security practices while doing their work? Are these security measures conflicting with the healthcare worker? How can the required security practices be improved for effective compliance while improving security effectiveness? How can the healthcare workers be incentivized to better comply with required security practice amidst their work? Or which required security measures need to be modified to enhance effective compliance?

In efforts towards answering these questions effectively, there is the need to analyze healthcare information security practice in the human context by looking for the gaps that exist between current healthcare workers' security practice and their required security practice which are defined in the legislation, regulations, standards, policies, guidelines, procedures, best practices and code of conducts [6].

There exists various ways in which security practices can be analysed. One of them is the modeling and analysing of the psychological, social, cultural and demographic perspectives of the healthcare workers' security practice [11]. This can be achieved by gathering and analyzing data on knowledge, attitude, opinion, behaviour, facts, etc., on the healthcare workers', objects, and events in a research survey [6], [10]–[16]. Another dimension involves analysing the social engineering behavior of the healthcare workers to determine their ability to identify and avoid such related attacks. [17]–[19]. Additionally, since the healthcare workers often access various assets and resources (eg healthcare records, physical access, networks etc) while leaving traces of their accesses which that can be reconstructed into their unique profile, there is the need to model and analyze the access logs of healthcare workers to understand their security practice in the aspect of big data [20]–[22].

Identified dilemmas can then be resolved with appropriate measures by devising means of resolving the challenges and providing incentivization methods for enhanced security practice. While all these approaches are important, the focus of this paper is to analyze healthcare security behaviour in the context of big data in relation to logs of electronic healthcare records.

Yeng et al identified network logs, EHR logs, keystroke dynamics and host based logs as some of the data sources which are being used for modeling and analyzing healthcare security practices [23]. Among these data sources, EHR logs was mostly used in the context of data-driven and AI approach. According to Boddy et al, EHR is one of the cardinal assets in the healthcare infrastructure which should be proactively monitored to detect both internal and external threats. To detect anomaly activities such as medical record snooping, social engineering threats to acquire healthcare professionals' logon credentials, erratic or unusual activities, there is the need to consider the modeling and analyzing of EHR [21], [24].

Due to emergency situations and the time sensitive nature of healthcare, there is usually the provision of a broad access to patients' healthcare records by the healthcare professionals in a typical hospital. In a role-based access control scenario, the healthcare workers in their assigned roles need to have similar pattern of access to patient records. For instance, the behaviour of users with nursing role should be similar in their accesses. However, if a nurse accesses within a period, tend to deviate from nursing role, then an abnormally can be quarried. Similarly, if an IT officers' role tend to act like a medical doctor within a given time, then an anomaly flag need to be raised. Inference can therefore be conducted into the anomalies to determine their maliciousness.

A. *Related work*

Various related studies have been conducted to safeguard electronic health records through the detection of anomalies in electron health records.For instance, [21] employed density-based local outlier detection model to profile users activities and their respective interactions with devices to detect and visualized abnormal security practices.A local outlier detection factor (LOF) assessed the local deviation of density by measuring the isolated distance of a data point to its k-nearest neighbours. Out of an unlabeled data set of 1,007,727 audit logs, the algorithm detected 144 anomalous behaviours. Also, a prediction method, dyadic prediction, [25] with collaborative filtering techniques was adopted by [26]. This method was used to predict the interaction of entity pairs just like how friends are recommended in social network, click-through rate prediction in computational advertising [27] and the prediction of the performance of students' test scores. The collaborative principle is about the assumption that if a person A and a person B share the same opinion on an issue, it is highly probable that the pair will have the same opinion in a different issue either than a randomly selected different person [27]. Additionally, Ziemniak et al employed C4.5 decision tree to detect abnormal security practice in a healthcare application. Ad-hoc analysis was used to determine atypical behaviour by visually looking for interesting nodes such as path-length investigation [24]. Furthermore, Gupta et al used K- Nearest Nieghbor (KNN) algorithm for the detection of outliers with the goal to detect anomalous users. Random topic access model (RTA) was targeted to identify users with illegitimate accesses with focus on common semantic themes [28]. Latent Dirichlet Allocation (LDA) was adopted in this study as a feature extraction technique. All these studies [21], [25], [27], [28] adopted various machine learning methods in their work however, a comparative analysis was not conducted to guide in the selection of the methods.

Healthcare data logs consists of various roles in which different roles can have close similarities in their operations. Additionally, there are erratic accesses due to uncertainties in healthcare such as emergency situations [23], [29]. For instance, healthcare systems have an emergency access mechanism known as "break the glass" or self-authorization which enables healthcare workers to access patients records without following the conventional authorization process [23], [29]. This opens up the system for numerous accesses in which various difference accesses can be similar [23], [29]. For instance, how will nurse A activities be distinguished from doctor A's activities in which both provided diagnosis and prescription to patients? Therefore, in analysing security practice in healthcare, it is necessary to compare the algorithms to determine the method that is fit for the purpose.

In that light, [30] compared Hidden markov model(HMM) and Distance-based model towards detecting anomalous user behaviours based on the sequence of their accesses within web sessions of electronic health application. The web sessions of users were converted into their respective workflows based on their respective access targets. So the anomalous workflows of users were being detected as their respective abnormal behaviours. Additionally, [22] compared community-based anomaly detection system wtih K-nerarset neighbors(KNN) and principal component analysis towards detecting threats in EHR based on the access logs of the healthcare staff. In the study, CAD performed better than KNN and PCA in Area under the ROC curve (AUC). Two methods each were compared in these studies([22], [30] to enable the selection of a better algorithm. However, there a other classification methods such as decision trees and rules, Bayesian classifiers, nearest neighbor classifiers, discriminant functions, support vector machines and neural networks [31] which were not considered in their studies.

B. *Scope and contribution*

Based on the gaps in the related works and review [23], we simulated electronic health records logs to perform the comparative analysis of the machine learning classification algorithms towards analyzing healthcare security practice. Aside the comparative analysis, various approaches called hard and soft classifications were performed and compared. The hard classification computed for the probabilities of each daily accumulated activities and classified the most probable into the respective role. But, the soft classification adopted a thresh holding mechanism. So if the probability of accumulated daily activities of a user meets a given threshold, that activity is then assigned into the given role. Furthermore, we compared the performance of z-score and Min-max normalization methods to access the performance of the algorithms in that aspect.

II. Our Method

A. *Health record logs data simulation*

We simulated a one-year access log data of the hospital information system from 01 January 2019 until 31 December 2019. We simulated five main modules in the hospital information system: Report, Finance, Patient Management, Laboratory Management, and Pharmacy Management. In the data simulation setting, we use 19 departments and 12 roles as displayed in Table I and Table II. There are two kinds of shifts used: the regular shift and the three 8 hours shift. The regular shift is from Monday to Friday 08.00-16.00 while the three 8 hours shift contains three shifts every day: a) Shift 1: 06.00-14.00, b) Shift 2: 14.00-22.00, and c) Shift 3: 22.00-06.00 (next day). The number of roles and employees in a regular shift can be seen in Table III while that in three 8-hours shifts can be seen in Table IV.

This simulation was built following some rule in the Norwegian code of conduct for healthcare security practices [6], [32] such as accessing patients records is only allowed for therapeutic purposes and is given to only those with an official need to use, self-authorization or "break the glass" scenarios is allowed but the necessary measures should be provided, and all of the activities related to the personal health data

TABLE I: List of Departments

ID	Name
0	IT
1	Finance
2	Administration
3	Laboratory
4	Pharmacy
5	Out Patients Ear-Nose-Throat
6	Out Patients Eyes
7	Out Patients Tooth
8	Out Patients Child
9	Out Patients Orthopedic
10	Out Patients Neurological
11	Out Patients Gynecological
12	Out Patients Diabetes
13	Out Patients Rheumatology
14	Out Patients Cancer
15	Emergency
16	In Patients Ward1
17	In Patients Ward2
18	In Patients Ward3

TABLE II: List of Roles

ID	Name	Code
0	Head of IT	HIT
1	Technical Support	TS
2	Head of Finance	HF
3	Finance Staff	FS
4	Head of Administration	HA
5	Staff of Administration	SA
6	Head of Lab	HL
7	Lab Assistant	LA
8	Head of Pharmachy	HP
9	Pharmacy Assistant	PA
10	Doctor	DO
11	Nurse	NU

TABLE III: Regular Shift

ID	Department	Roles (number of employees)
0	IT	HIT(1), TS(2)
1	Finance	HF(1), FS(4)
2	Administration	HA(1), SA(2)
3	Laboratory	HL(1), LA(5)
4	Pharmacy	HP(1), PA(2)
5	Out Patients Ear-Nose-Throat	DO(1), NU(2)
6	Out Patients Eyes	DO(1), NU(2)
7	Out Patients Tooth	DO(1), NU(2)
8	Out Patients Child	DO(1), NU(2)
9	Out Patients Orthopedic	DO(1), NU(2)
10	Out Patients Neurological	DO(1), NU(2)
11	Out Patients Gynecological	DO(1), NU(2)
12	Out Patients Diabetes	DO(1), NU(2)
13	Out Patients Rheumatology	DO(1), NU(2)
14	Out Patients Cancer	DO(1), NU(2)
16	In Patients Ward1	DO(1)
17	In Patients Ward2	DO(1)
18	In Patients Ward3	DO(1)

and personal data must be recorded. In this simulation, the flow of patients in the inpatients, outpatients, and emergency department are displayed in Fig. 1, 2, and 3 respectively. Based on the flows, we simulated the data and recorded the logs. The logs is considered as normal data (non anomaly). Besides, we also simulate some abnormal data. The abnormal data are generated by simulating attackers that are assumed have compromised some users credential and use it to access patients records (e.g. identity theft). The attacker will access more data than legitimate users and sometimes not follow the flows. From this data simulation, 283.678 logs were created with 274.983 of them are legitimate access while 8.695 of them are fraudulent. There are 21 fields recorded in this data simulation like displayed in Table V.

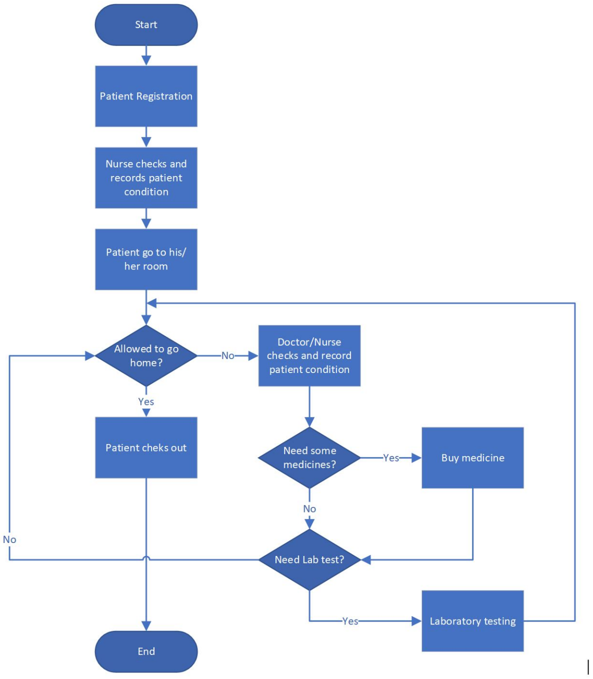


Fig. 1: The Inpatients Department Flow

TABLE IV: Three 8-hours shift

ID	Department	Roles (number of employees)
15	Emergency	DO(2), NU(7)
16	In Patients Ward1	NU(2)
17	In Patients Ward2	NU(2)
18	In Patients Ward3	NU(2)

B. Proposed method for anomaly detection

The anomaly detection method used is based on the fact that people with the same role have similar activities and people with different roles tend to have different activities. For example, Doctor A and Doctor B tend to have similar activities but Doctor A and IT staff C are unlikely to have similar activities. If Doctor A's activity on a particular day has a low similarity with doctor's activity but has a high similarity

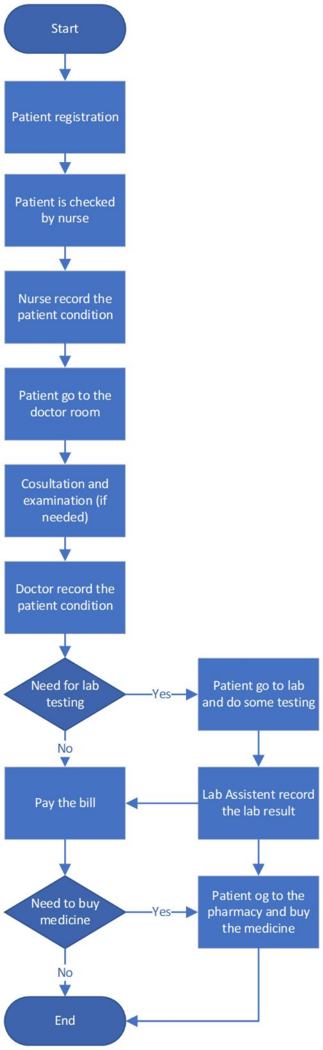

Start
Patient registration
Patient is checked by nurse
Nurse record the patient condition
Patient go to the doctor room
Cosultation and examination (if needed)
Doctor record the patient condition
Need for lab testing
Yes
Patient go to lab and do some testing
No
Pay the bill
Lab Assistent record the lab result
Need to buy medicine
Yes
Patient og to the pharmacy and buy the medicine
No
End

Fig. 2: The Outpatients Department Flow

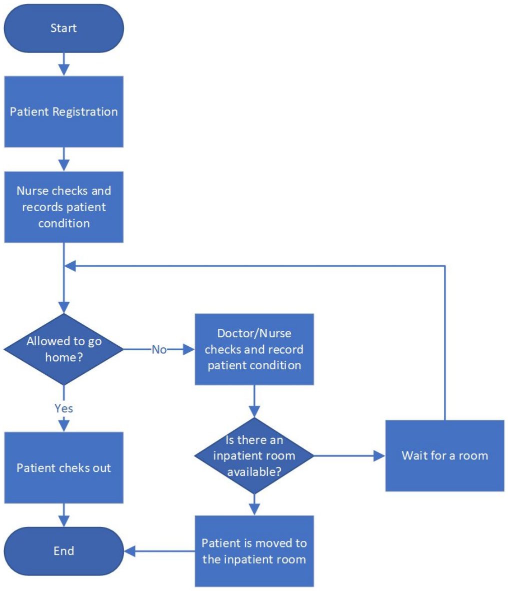

Start
Patient Registration
Nurse checks and records patient condition
Allowed to go home?
No
Doctor/Nurse checks and record patient condition
Yes
Is there an inpatient room available?
Wait for a room
Patient cheks out
End
Patient is moved to the inpatient room

Fig. 3: The Emergency Department Flow

with the IT staff's activity, then the Doctor A's activity on that particular day can be abnormal.

The method proposed in this work aims to identify the anomaly by comparing the user's activity to their role's normal activity, such as the type of actions being taken and the number of patients they are viewing. First, a model for role classification is trained. Then, using the model, the activity of each user is classified. If the activity is classified into the real role of the user, the activity is considered normal. Otherwise, the activity is considered an anomaly. In this way, potentially illegitimate access to patient records can be highlighted and investigated.

1) Feature Extraction: To develop the anomaly detection model, including the role classification model, some features were extracted. Each log entry represents a single transaction for a user. To analyze the user activity, the logs from each user are consolidated into a particular period. Every single activity of Doctor A is a poor data point that will be hard to analyze separately. However, by observing several activities of Doctor A for a particular period, it will be easier to do the anomaly detection task. In this work, we process the logs data into 24-hour blocks so that an instance represents the cumulative activity of a user in a single day. As the results, 24,648 instances are extracted from the raw logs with 24,286 of them are considered normal and 362 of them are considered an anomaly. The definition of anomaly data here is all the instances had at least one fraudulent log access in a single day while the normal data are all the instances whose all access logs are in line with the roles. Afterward, these instances are then transformed into features for role classification and anomaly detection processes. Table VI shows the features extracted from the dataset. In this work, we also use two normalization methods: Z-score normalization, and Min-Max normalization.

2) Role Classification Model: Since the anomaly detection method in this work is based on the role classification, we need to build the role classification model first. The goal of role classification is to classify the cumulative user activity in a single day into one of the 12 categories as shown in II . The model is trained using only normal data because the anomaly data are data from the attacker that tend to behave differently from the real users. Then, the model is used to classify the cumulative activity of a user in a single day. Eight machine learning methods were used as classifiers for the role classification model including Multinomial Naive Bayes(multnb), Bernoulli Naive Bayes (bernnb), Support Vector Machine (svm), Neural Network (nn), K-Nearest Neighbours(knn), Logistic Regression (lr), Random Forest (rf), and Decision Tree (dt). To evaluate the model, we conducted 5-folds cross-validation on the normal data. The number of normal data is 24,286 instances. The evaluation method for this task is accuracy.

TABLE V: Record Fields

Number	Field Name	Description
1	startAccessTime	The time employee start to acces the patient record. format = ’dd/mm/yyy HH:mm tt’
2	endAccessTime	The time employee end the patient record access. format = ’dd/mm/yyy HH:mm tt’
3	employeeID	The ID of the employee who access the patient record
4	roleID	The role of the employee who access the patient record
5	patientID	The ID of the patient whose record is being accessed by employee
6	activityID	The ID of the activity (1: Create, 2: Read, 3:Update, 4: Delete)
7	employeeDepartmentID	The department of the employee who access the patient record
8	employeeOrganizationID	The organization of the employee who access the patient record
9	osID	The OS of the computer used by the employee to access patient record
10	deviceID	The ID of the computer used by the employee to access patient record
11	browserID	The browser used by the employee to access patient record
12	ipAddress	The IP Address of the computer used by the employee to access patient record
13	ReasonID	The reason of employee access the patient record (optional)
14	shiftID	The ID of shift the employee belong to on the day of patient access record
15	siftStartDateTime	The start time of shift the employee belong to on the day of patient access record
16	siftEndDateTime	The end time of shift the employee belong to on the day of patient access record
17	CRUD	The ID of the activity (C: Create, R: Read, U:Update, D: Delete)
18	AccessControlStatus	Access Control Status
19	SessionID	The ID of the session access
20	AccessPatient_Warnings	Warning for not usual access
21	ModuleUsed	The module acessed by the employee

TABLE VI: Dataset feature names and descriptions

Feature Name	Description
number of create	Number of ’create’ transactions conducted in a single day
number of read	Number of ’read’ transactions conducted in a single day
number of update	Number of ’update’ transactions conducted in a single day
number of delete	Number of ’delete’ transactions conducted in a single day
number of patient record	Number of access to the patient records in a single day
number of unique patient	Number of unique patients whose records has been accessed in a single day
number of modules	Number of kind of modules in the information system accessed in a single day
number of report module	Number of transactions conducted in the report module in a single day
number of finance module	Number of transactions conducted in the finance module in a single day
number of patient module	Number of transactions conducted in the patient management module in a single day
number of lab module	Number of transactions conducted in the laboratory module in a single day
number of pharmacy module	Number of transactions conducted in the pharmacy module in a single day
number of outside access	Number of transactions conducted from outside hospital network in a single day
number of browser	Number of browser type used in a single day
number of chrome	Number of chrome browser used in a single day
number of ie	Number of Internet Explorer browser used in a single day
number of safari	Number of Safari browser used in a single day
number of firefox	Number of Firefox browser used in a single day
number of other browser	Number of other browser used in a single day

3) Anomaly Detection: The anomaly detection method used in this work is based on the role classification model. There are two different approaches employed as follows:

- Hard Classification: In this approach, we classify each instance (cumulative user activity in a single day) into one category. Like mentioned before, the categories used are the list of roles in the hospitals. Since there are 12 roles in the simulated hospital, the number of categories is also 12. If the user’s cumulative activity in a single day is classified into her/his actual role, then the instance is considered normal. Otherwise, if the user’s cumulative activity in a single day is not classified into her/his actual role, then the instance is considered an anomaly. For example, if Doctor A’s cumulative activity in a single day is classified into the Doctor category, then it is considered normal. Otherwise, if Doctor A’s cumulative activity in a single day is classified into other categories than Doctor (e.g. Nurse, Technical Support, etc.), then it is considered an anomaly.
- Soft Classification: This approach is similar to the hard classification approach but in a softer way. It gives tolerance for the user to act like users from other roles because some roles have quite similar activities. In this approach, the classifier computed the probability of the user’s instance belong to their role class. If the probability is above a particular threshold, then it is considered normal. Otherwise, it will be considered an anomaly. For example, the classifier will compute the probability of Doctor A’s cumulative activity in a single day into the Doctor category because his actual role is Doctor. Then, if the probability is above a particular threshold, then it

is considered normal. Otherwise, it will be considered an anomaly.

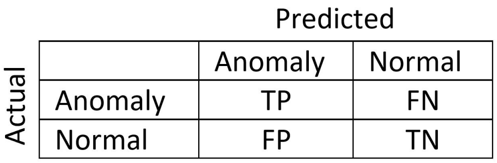
To evaluate this anomaly detection, we use the logs from January until August as training data while data from September until December is used for testing data. The training data is used to train the role classification model. Then, this model is used to detect anomaly based on the two proposed approaches. For this task, precision, recall, and f1-measure are used to evaluate the method.

C. Performance Evaluation

For the role classification task, accuracy is used for evaluation. The following is the formula to calculate the accuracy:

$$Accuracy = \frac{NumberOfCorrectPrediction}{NumberOfData} \tag{1}$$

where $NumberOfCorrectPrediction$ is the number of instances that are correctly classified into their actual role while $NumberOfData$ is the total number of instances in the dataset.



		Predicted	
		Anomaly	Normal
Actual	Anomaly	TP	FN
	Normal	FP	TN

Fig. 4: Confusion Matrix

For the anomaly detection, several measurements including Accuracy (Acc), Precision (P), Recall (R), and F_1-score (F_1) were used to evaluate the performance. All measurements were calculated based on the confusion matrix displayed in Fig. 4. True Positive (TP) and True Negative (TN) are the numbers of features that were correctly predicted. TP represents the number of anomaly data that were correctly predicted as an anomaly while TN represents the number of normal data or users that were correctly predicted as normal. Meanwhile, False Positive (FP), or often called Type I Error is the number of normal data that were incorrectly predicted as anomaly ones and False Negative (FN) or Type II Error represents the number of anomaly data that were incorrectly predicted as normal ones. The followings are the formulas for each measurement:

$$Acc = \frac{TP + TN}{TP + FP + FN + TN} \tag{2}$$

$$P = \frac{TP}{TP + FP} \tag{3}$$

$$R = \frac{TP}{TP + FN} \tag{4}$$

$$F_1 = 2\frac{P \cdot R}{P + R} \tag{5}$$

III. Result

A. Role Classification Model Result

The experiment result of the role classification model is depicted in Table VII. Overall, all of the machine learning method employed shows a good performance with an accuracy of more than 0.7. Decision Tree (dt) and Random Forrest (rf) obtained the best result among all of the methods with the same accuracy value of 0.889 on all three datasets. Meanwhile, Multinomial Naive Bayes (multnb) achieved the lowest accuracy on the Min-Max based normalized data with an accuracy value of 0.716.

The use of normalization does not make any significant improvement in this case. Only SVM and KNN that have a slight increase in accuracy by using normalization on the dataset. Decision Tree (dt) and Random Forrest (rf) obtained the same result on all three dataset types while Bernoulli Naive Bayes achieved the same accuracy on None Normalised data and normalized data using Min-Max. To be noted, Multinomial Naive Bayes cannot classify normalized data using Z-score because this classifier cannot get negative value as the input. On the None Normalised dataset, there is no feature with a negative value. After normalized using Z-score, there are several negative values so that it does not suitable with the Multinomial Naive Bayes requirement.

TABLE VII: Role Classification Model Accuracy

Method	None Normalised data	Normalized data (Z-score)	Normalized data (Min-Max)
multnb	**0.881**	-	0.715
bernnb	**0.774**	0.733	**0.774**
nn	**0.886**	0.868	0.878
knn	0.858	0.865	**0.888**
lr	**0.882**	0.879	0.852
rf	**0.889**	**0.889**	**0.889**
dt	**0.889**	**0.889**	**0.889**
svm	0.871	**0.875**	0.862

B. Anomaly Detection result

TABLE VIII: Anomaly Detection Result using Hard Classification Approach on None Normalised Data

Method	Acc	Prec	Rec	F1
multnb	0.880	0.037	0.698	0.071
bernnb	0.776	0.025	0.868	0.048
nn	0.909	0.045	0.642	0.084
knn	0.873	0.030	0.585	0.057
lr	0.891	0.046	0.792	0.087
rf	0.913	0.041	0.547	0.076
dt	0.913	0.050	0.679	0.093
svm	0.909	0.046	0.660	0.086

The anomaly detection results using Hard Classification approach are displayed in Table VIII, IX, X. In terms of accuracy, generally Random Forrest (rf), Decision Tree (dt), and neural network (nn) have the best result. In terms of precision, Decision Tree (dt) has the best result but it is still very low (0.050). Decision Tree (dt) also has the best result

TABLE IX: Anomaly Detection Result using Hard Classification Approach on on Normalized Data (Z-Score)

Method	Acc	Prec	Rec	F1
multnb	-	-	-	-
bernnb	0.728	0.020	0.868	0.040
nn	0.914	0.049	0.660	0.091
knn	0.893	0.032	0.528	0.060
lr	0.879	0.025	0.472	0.048
rf	0.913	0.041	0.547	0.076
dt	0.914	0.050	0.679	0.093
svm	0.889	0.023	0.396	0.044

TABLE X: Anomaly Detection Result using Hard Classification Approach on Normalized Data (Min-Max)

Method	Acc	Prec	Rec	F1
multnb	0.715	0.019	0.849	0.037
bernnb	0.776	0.025	0.868	0.048
nn	0.910	0.039	0.547	0.073
knn	0.913	0.041	0.547	0.075
lr	0.847	0.009	0.208	0.017
rf	0.913	0.041	0.547	0.076
dt	0.913	0.050	0.680	0.093
svm	0.857	0.007	0.151	0.014

for F1-score. Meanwhile, Naive Bayes methods (multnb and bernnb) have the best result in terms of recall. It can also be seen from the results that the use of normalization does not have any improvement for anomaly detection using the Hard Classification approach.

Overall, using this approach, the anomaly detection methods achieved very good accuracy and adequate recall but low precision and F1-score. Despite all of the machine learning methods used to have good accuracy, we cannot conclude that all of the methods are good to detect an anomaly. It is important to note that the dataset is unbalanced. The number of normal data is far higher than the number of anomaly data. A method could have a good accuracy even though the TP is very low as long as the TN is high. In other words, a method could still have good accuracy even though it cannot detect the anomaly. The good accuracy does not always mean that a method is good enough for detection for this case. In an extreme case, because the number of normal data is far more than the anomaly data, the accuracy would still remains very good even though a method predicts all of the data as normal. Therefore, accuracy alone is not suitable for the anomaly detection task evaluation in this work and we need to see the other measurements such as precision, recall, and F1. Based on the fact that recall of all of the methods is quite good but the precision is very low, it can be agreed that in all of the methods the number of FP is high but the number of FN is low. It means that there are many normal data that are wrongly classified as an anomaly but there only a few anomaly data that are wrongly classified as normal. The high recall is actually good if the data that are predicted anomaly will be investigated again so that most of the actual anomaly data will not be missed.

Meanwhile, using the Soft Classification approach, the threshold become a significant factor for the performance as shown in Fig. 5, 6, 7, 8 . As expected, generally, the higher threshold, the higher recall, and the lower the precision. It happens because a lower threshold will give more tolerance for the activity to be called normal. The consequences of a low threshold is that there are more data classified as normal and fewer data classified as anomaly so that the precision of the method to detect anomaly become higher but the recall becomes lower. Otherwise, a higher threshold provides a high qualification for the data in order to be classified as normal. As a result, there are fewer data classified as normal and more data classified as anomaly so that the precision of the method to detect anomaly becomes lower but the recall becomes higher. Generally, the best result is achieved when the threshold used is 0.1. Table XI shows the F1-score of anomaly detection result using Soft Classification with a threshold value of 0.1. Bernoulli Naive Bayes unexpectedly achieved the best F1-Score on the None Normalised data with a quite high score (0.893). The use of binary features employed by Bernoulli Naive Bayes has become very effective for this task.

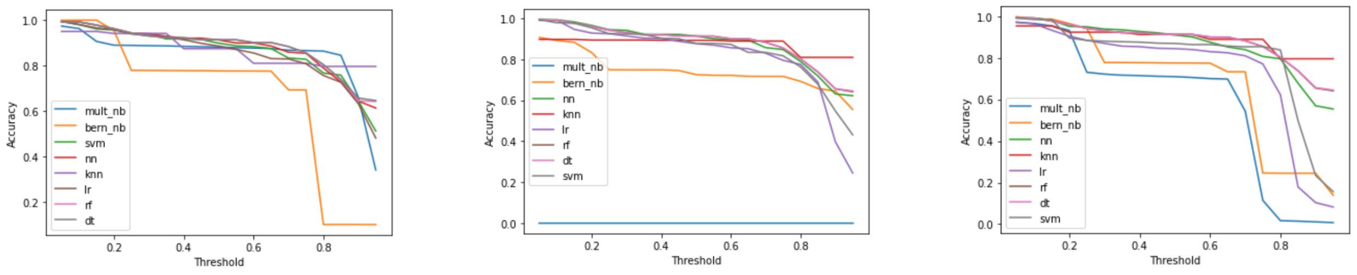
The experiment results also show that generally, the Soft Classification approach obtained better performance than the Hard Classification approach. It happens because the activity of different roles can be very similar so that giving a tolerance can improve the performance. However, apart from Soft Classification based Bernoulli Naive Bayes method, the performance of the proposed method is still low.

TABLE XI: F1-Score of Anomaly Detection Result using Soft Classification Approach with Threshold = 0.1

Method	None Normalised data	Normalized data (Z-score)	Normalized data (Min-Max)
multnb	0.152	-	0.243
bernnb	0.893	0.091	0.457
nn	0.208	0.548	0.214
knn	0.375	0.046	0.095
lr	0.115	0.206	0.032
rf	0.264	0.377	0.355
dt	0.383	0.482	0.485
svm	0.507	0.184	0.075

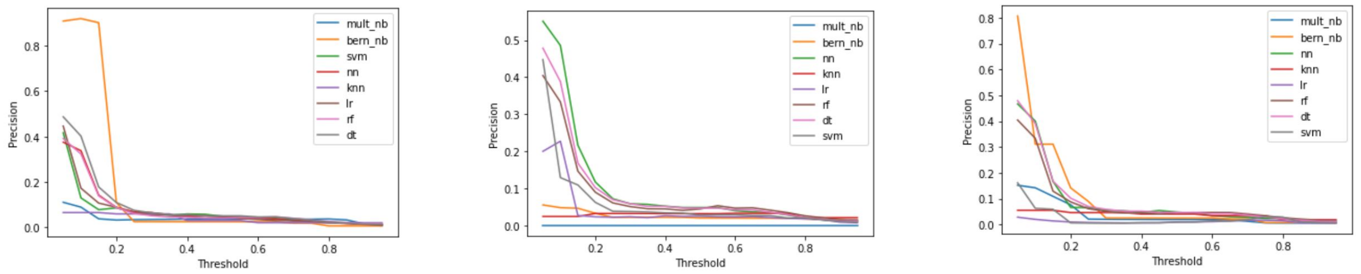
IV. Discussion

Following the surge in data breaches within healthcare in recent years [3]–[6], [8], [10] and their related life-threatening consequences [1], there is the need to analyse healthcare security practices in various ways. One of the ways is the analysis of EHR logs in the context of big data [6], [11], [23]. According to [6] the accesses of healthcare staff can be reconstructed to form their unique profiles. As healthcare personnel frequently access electronic healthcare records for therapeutic and other functions, the logs can be analysed with the suitable machine learning methods to detect anomalies and if possible to determine maliciousness. The healthcare staff's access can be broad in self-authorization or ”break the glass” scenarios and this can make it complex for the IT personnel



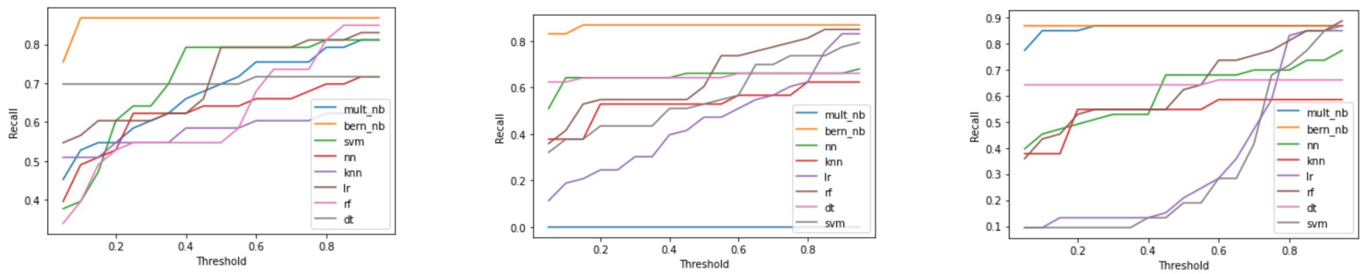
(a) Result on None Normalised Data (b) Result on Normalized Data (Z-Score) (c) Result on Normalized Data (Min-Max)

Fig. 5: Accuracy of Anomaly Detection using Soft Classification



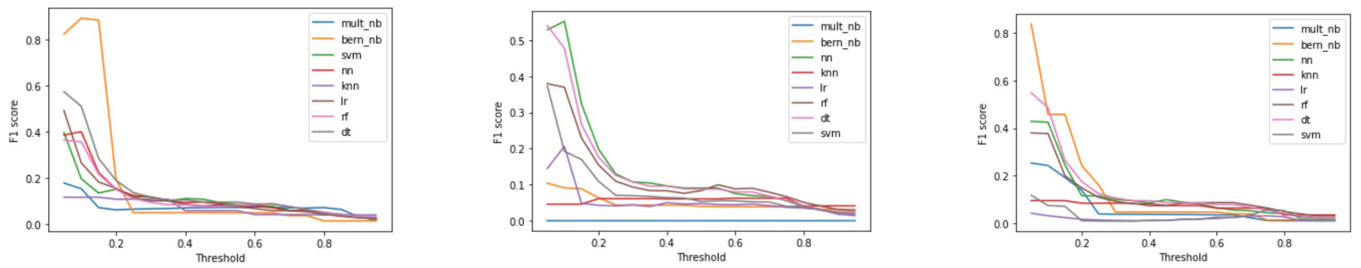
(a) Result on None Normalised Data (b) Result on Normalized Data (Z-Score) (c) Result on Normalized Data (Min-Max)

Fig. 6: Precision of Anomaly Detection using Soft Classification



(a) Result on None Normalised Data (b) Result on Normalized Data (Z-Score) (c) Result on Normalized Data (Min-Max)

Fig. 7: Recall of Anomaly Detection using Soft Classification



(a) Result on None Normalised Data (b) Result on Normalized Data (Z-Score) (c) Result on Normalized Data (Min-Max)

Fig. 8: F1-score of Anomaly Detection using Soft Classification

in the hospital to manually determine unauthorised accesses such as insider or outsider masquerades.

To this end, we focused on comparing machine learning classification methods using simulated logs of EHR. The simulated data of EHR logs in this study was quite useful since the different types of machine learning algorithms needed to be evaluated to assess the performance of the methods [33] prior to usage in real applications. Health record logs data is confidential and most hospitals do not want to take the risk in sharing their logs. Clearly, real EHR logs or semi-synthetic data could be the better option in the assessment. However, EHR logs is very sensitive and there are regulatory hurdles and stringent privacy laws across the globe [34] that are protecting the sensitive healthcare data. So hospitals are not willing to risk in giving out such data. To succeed in accessing the performance of the machine learning algorithms amidst these challenges, a simulated logs data was a clear choice to serve as a playground or a test range for comparing the suitability of these algorithms for analysing healthcare security practice without violating security and privacy laws [35]. Yeng et al adopted similar approach in testing algorithms towards detecting disease outbreak [34]. Therefore, simulated electronic health record logs data was used in this work.

We used a role classification based anomaly detection method because users with the same roles tend to have similar activity while users with different roles tend to have different activities [21], [22], [24], [27]. The experiment results show that all of the machine learning method employed shows a good accuracy for roles classification as shown in figure 5. However, despite the good accuracy and recall, the methods still have a low performance in detecting anomaly in terms of precision and F1-score as shown in figure 6, figure 7 and figure 8. The high recall is actually good for the data administrators if they undergo further investigation. That way, most of the actual anomaly data will not be missed. Usually, in the hospital, broad access is given to healthcare staff through self-authorisation but this require the the IT staffs to manually evaluate the anomaly and malicious access [29]. Therefore, the result from this work can be used by the hospital to narrow down the data for the manual investigation work.

The experiment results also show that generally, the Soft Classification approach achieved better performance than the Hard Classification approach as shwon in figure 7. It happens because the activity of different roles can be very similar so that giving a tolerance can improve the performance. The use of normalization also did not give any improvement to the performance. The best performance is obtained using Bernoulli Naive Bayes on the None Normalised data with an F1-score of 0.893.

V. Conclusion

Due to the recent increases in data breaches within healthcare, we compared various machine learning classification methods using simulated EHR logs towards determining anomalies. The experiment results show that all of the methods used achieved quite a good accuracy for role classification. For the anomaly detection, generally, all of the methods obtained a high recall and accuracy but low precision and F1-score. This high recall means that the method from this work can be a good tool to narrow down the data for further manual investigation. Since the activity of different role can be very similar, Soft Classification approach performed better than the Hard Classification approach because the former provides some tolerances. The best performance is obtained using Bernoulli Naive Bayes on the None Normalised data with an F1-score of 0.893.

In fact since anomaly detection does not entirely means maliciousness, there is the need for future works on further processing the anomalies to detect malicious activities. Besides, since real EHR logs data have not been used for such a comparison, the use of real data instead of simulated one can give a better insight. Additional, as labeled real data can be difficult to get, it is also important to compare unsupervised methods for the detection of anomalies and maliciousness in the context of big data.

References

[1] AssociatedPress, "German hospital hacked, patient taken to another city dies." [Online]. Available: "https://www.securityweek.com/german-hospital-hacked-patient-taken-another-city-dies"

[2] ISO, "Health informatics information security management in health using iso/iec 27002," 2016. [Online]. Available: "https://www.iso.org/obp/ui/iso:std:iso:27799:ed-2:v1:en"

[3] J. Ehrenfeld, "cybersecurity and health information technology: A time to act," 2017.

[4] HealthITSecurity, "Incentivize cybersecurity best practices for data security," 2017. [Online]. Available: https://healthitsecurity.com/news/incentivize-cybersecurity-best-practices-for-data-security

[5] P. K. Yeng, B. Yang, and E. A. Snekkenes, "Healthcare staffs' information security practices towards mitigating data breaches: A literature survey." in *pHealth*, 2019, pp. 239–245.

[6] P. Yeng, B. Yang, and E. Snekkenes, "Observational measures for effective profiling of healthcare staffs' security practices," in *2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)*, vol. 2. IEEE, 2019, pp. 397–404.

[7] SearchHealthIT, "Hospital takes aim at patient health data security with ai tools," 2019. [Online]. Available: hhttps://searchhealthit.techtarget.com/feature/Hospital-takes-aim-at-patient-health-data-security-with-AI-tools.

[8] Verison, "Data breaches report. 2019," 2019. [Online]. Available: https://www.nist.gov/system/files/documents/2019/10/16/1-2-dbir-widup.pdf

[9] HIMSS, "So you've been hit with a ransomware attack. what now?" [Online]. Available: "https://www.healthcareitnews.com/news/so-youve-been-hit-ransomware-attack-what-now"

[10] S. D. Cannoy and A. Salam, "A framework for health care information assurance policy and compliance," *Communications of the ACM*, vol. 53, no. 3, pp. 126–131, 2010.

[11] P. K. Yeng, B. Yang, and E. A. Snekkenes, "Framework for healthcare security practice analysis, modeling and incentivization," in *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, 2019, pp. 3242–3251.

[12] E. H. Rosch, "On the internal structure of perceptual and semantic categories," in *Cognitive development and acquisition of language*. Elsevier, 1973, pp. 111–144.

[13] R. B. Radhakrishna, "Tips for developing and testing questionnaires/instruments," *Journal of extension*, vol. 45, no. 1, pp. 1–4, 2007.

[14] N. Humaidi and V. Balakrishnan, "The influence of security awareness and security technology on users' behavior towards the implementation of health information system: A conceptual framework," in *2nd International Conference on Management and Artificial Intelligence IPEDR*, vol. 35, 2012, pp. 1–6.

[15] N. S. Safa, M. Sookhak, R. Von Solms, S. Furnell, N. A. Ghani, and T. Herawan, “Information security conscious care behaviour formation in organizations,” *Computers & Security*, vol. 53, pp. 65–78, 2015.

[16] J. L. Fernández-Alemán, A. Sánchez-Henarejos, A. Toval, A. B. Sánchez-García, I. Hernández-Hernández, and L. Fernandez-Luque, “Analysis of health professional security behaviors in a real clinical setting: An empirical study,” *International journal of medical informatics*, vol. 84, no. 6, pp. 454–467, 2015.

[17] A. Wright, S. Aaron, and D. W. Bates, “The big phish: cyberattacks against us healthcare systems,” 2016.

[18] W. J. Gordon, A. Wright, R. Aiyagari, L. Corbo, R. J. Glynn, J. Kadakia, J. Kufahl, C. Mazzone, J. Noga, M. Parkulo *et al.*, “Assessment of employee susceptibility to phishing attacks at us health care institutions,” *JAMA network open*, vol. 2, no. 3, pp. e190 393–e190 393, 2019.

[19] W. J. Gordon, A. Wright, R. J. Glynn, J. Kadakia, C. Mazzone, E. Leinbach, and A. Landman, “Evaluation of a mandatory phishing training program for high-risk employees at a us healthcare system,” *Journal of the American Medical Informatics Association*, vol. 26, no. 6, pp. 547–552, 2019.

[20] A. Boddy, W. Hurst, M. Mackay, and A. El Rhalibi, “A study into detecting anomalous behaviours within healthcare infrastructures,” in *2016 9th International Conference on Developments in eSystems Engineering (DeSE)*. IEEE, 2016, pp. 111–117.

[21] A. J. Boddy, W. Hurst, M. Mackay, and A. El Rhalibi, “Density-based outlier detection for safeguarding electronic patient record systems,” *IEEE Access*, vol. 7, pp. 40 285–40 294, 2019.

[22] Y. Chen, S. Nyemba, W. Zhang, and B. Malin, “Specializing network analysis to detect anomalous insider actions,” *Security informatics*, vol. 1, no. 1, p. 5, 2012.

[23] P. Yeng, L. O. Nweke, A. Z. Woldaregay, B. Yang, and E. A. Snekkenes, “Data-driven and artificial intelligence (ai) approach for modelling and analyzing healthcare security practice: A systematic review,” 2020.

[24] T. Ziemniak, “Use of machine learning classification techniques to detect atypical behavior in medical applications,” in *2011 Sixth International Conference on IT Security Incident Management and IT Forensics*. IEEE, 2011, pp. 149–162.

[25] T. Hofmann, J. Puzicha, and M. I. Jordan, “Learning from dyadic data,” in *Advances in neural information processing systems*, 1999, pp. 466–472.

[26] 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, no. 1, pp. 87–101, 2014.

[27] A. K. Menon, K.-P. Chitrapura, S. Garg, D. Agarwal, and N. Kota, “Response prediction using collaborative filtering with hierarchies and side-information,” in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2011, pp. 141–149.

[28] S. Gupta, C. Hanson, C. A. Gunter, M. Frank, D. Liebovitz, and B. Malin, “Modeling and detecting anomalous topic access,” in *2013 IEEE International Conference on Intelligence and Security Informatics*. IEEE, 2013, pp. 100–105.

[29] L. Rostad and O. Edsberg, “A study of access control requirements for healthcare systems based on audit trails from access logs,” in *2006 22nd Annual Computer Security Applications Conference (ACSAC’06)*. IEEE, 2006, pp. 175–186.

[30] X. Li, Y. Xue, and B. Malin, “Detecting anomalous user behaviors in workflow-driven web applications,” in *2012 IEEE 31st Symposium on Reliable Distributed Systems*. IEEE, 2012, pp. 1–10.

[31] I. Kononenko and M. Kukar, *Machine learning and data mining*. Horwood Publishing, 2007.

[32] D. for e Health, “Code of conduct for information security and data protection in the healthcare and care services sector,” 2018. [Online]. Available: https://ehelse.no/normen/documents-in-english

[33] N. Jafarpour Khameneh, “Machine learning for disease outbreak detection using probabilistic models,” Ph.D. dissertation, École Polytechnique de Montréal, 2014.

[34] P. Yeng, A. Z. Woldaregay, and G. Hartvigsen, “K-cusum: Cluster detection mechanism in edmon,” 2019.

[35] J. P. Burgard, J.-P. Kolb, H. Merkle, and R. Münnich, “Synthetic data for open and reproducible methodological research in social sciences and official statistics,” *AStA Wirtschafts-und Sozialstatistisches Archiv*, vol. 11, no. 3-4, pp. 233–244, 2017.