Workflow-based anomaly detection using machine learning on electronic health records' logs: A Comparative Study

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

2$^{\text{nd}}$ Muhammad Ali Fauzi
Department of Information Security and Communication Technology
NTNU
Gjøvik, Norway
muhammad.a.fauzi@ntnu.no

3$^{\text{rd}}$ Bian Yang
Department of Information Security and Communication Technology
NTNU
Gjøvik, Norway
bian.yang@ntnu.no

***Abstract*—Timely access to patients' healthcare records is very essential. As a result, broad access to EHR is mostly provided to users in efforts towards complying with the availability trait of the CIA. However, this opens up the system for abuse and misuse. This paper, therefore, analyzed the workflows of healthcare staff's security practices in electronic health records (EHR) logs to determine anomalous security practices. Different classification types of machine learning algorithms were used. The EHR logs were simulated based on healthcare workflow scenarios. A number of machine learning algorithms were used to analyze the logs for deviations of accesses from the workflow. Based on the analysis results, all of the machine learning methods generally obtained a very good performance. The best performance on the non-normalized dataset is achieved by the Logistic Regression method with accuracy, precision, recall, and F1 value of 0.998, 0.849, 0.978, and0.909 respectively while Random Forest obtained the best result on Normalized data with accuracy, precision, recall, and F1 value of 0.998, 0.867, 0.836, and 0.851 respectively. It however remains challenging to detect malicious security practice if a malicious actor follows the workflow to access healthcare records with legitimate access right.**

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

I. Introduction

The adverse impacts in data breaches within healthcare are no longer hypothetical statements but realities that are causing harm to innocent data subjects, healthcare providers, and other stakeholders. For instance, in Finland, there was a data breach in 2018 and 2019 on Vastaamo, a center for psychotherapy which has 20 clinics located in different places [1]. The breach resulted in compromising about 300 patients records of which several of the affected data subjects have been demanded by the hackers to pay a ransom of about 500 Euros each in bitcoin or else, their data will be released to the public. In a related incident in September 2020, there was a ransomware attack at Duesseldorf University Clinic in Germany. As a result, the medical records of a patient were not timely available during an emergency and this resulted in the death of that patient [2].

Though the cause of these breaches was not disclosed, most of the data breaches in recent times have been attributed to human factors [3], [4] through the usage of social engineering tricks and other techniques [5], [6]. Technological countermeasures (eg firewalls, intrusion detection and prevention systems, antivirus, etc) have been the default and traditional methods [7]. These countermeasures have therefore been strengthened, making it more difficult for the circumvention of hackers. So, the hackers tend to exploit the human elements who are the weakest link in the security chain [8]–[10]. Some legitimate users with access rights can also tend to sell their access credentials to hackers thus complicating the security measures. The data breaches continue to increase in healthcare based on human factors According to the IBM report, the healthcare sector has recorded the costliest data breaches among various sectors in the aspect of the cost of loss of business, ex-post response, notification, and detection, and escalation [11].

To this end, various efforts are being adopted to minimize the data breaches in healthcare which includes a comprehensive approach involving modeling and analyzing healthcare security practice in the context of big data [10].

The general objective of this paper is to therefore analyze simulated electronic healthcare records logs with the aim to compare various classification methods. The most effective and efficient algorithms are to be adopted towards analyzing real EHR logs to determine anomalies towards developing security countermeasures.

A. Security requirement in EHR

Security measures in EHR need to conform with the confidentiality, integrity, and availability (CIA) traits requirements. For instance, in complying with the Norwegian code of conduct for healthcare security practices [12], [13]:

- Patients records need to be accessed for therapeutic purposes
- Access must be granted following a specific decision based on the completed or planned implementation of measures for the medical treatment of the patient. So access must be provided to comply with the confidentiality rules. This means access to personal health data

and personal data is given to only those with an official need to use.

- In the case of data exchange between organizations, the hospital needs to have the technical and organizational solutions to prevent access to health data as specified within the CIA traits, including authorized access with adequate authentication and least privileges.
- In self-authorization or ”break the glass” scenarios, the necessary measures should be provided to enable access to patient information when necessary. However, misuse of self-authorization needs to be handled as a breach [12]–[14].
- EHR with permission functions must record rights to read, register, correct, erase, and/or block personal health data and personal data in the access management.

The logs must contain the following [12]–[14]:

- unique identifier for the authorized user
- The role of the authorized user at the time of access
- Organisational affiliation
- Organisational affiliation of the authorized person
- Type of data to which access has been gained
- The user who disclosed health data that is linked to the name or national ID number of the patient or health care user
- Basis for the access
- Time and duration of access.

Additionally, Confidentiality measures need to include the following [13]:

- Persons outside the organization must not gain unauthorized access to EHRs
- Persons within the organization must be given access in accordance with established principles for access control especially based on the need to use basis.
- Details of persons who gained access (entered records, changes, corrections, and deletions) need to be registered in the logs to ensure audit trail to the origin.
- Persons or technologies, within or outside the organization, must not be able to access healthcare data without appropriate authorization.
- Personal health data and personal data must be linked to an identifiable person and must be accurate.

In the context of availability, personal health data and personal data must be accessible when there is an official need to access such data amidst the confidentiality framework. Self-authorization or ”break the glass” may be established to provide access to authorized users to access EHR without following the conventional authorization procedures on the basis of the need to use. But established procedures need to be established, the reason for and self-authorization must be documented. All misuse of self-authorization must be followed up as a breach.

B. Problem statement, scope and contribution

In this work, we hypothesized that anomaly behavior can be detected effectively if we model and analyze the normal healthcare staff behavior of a hospital in general. This can be achieved by considering all other activities that deviate from the normal workflow (as specified in section I-A) as an anomaly. An establishment of normal behavior at the hospital level, in general, is considered in this work, having consolidated all activities to establish the normal behavior pattern. For this purpose, we developed a health information system workflow to simulate health records logs data. The simulation is based on the general workflow in the hospital. This data is then used for the anomaly detection task. We extract several features from this data for the anomaly detection task. Furthermore, a comparative analysis is conducted by applying several machine learning classifiers to do the detection. The feature selection and normalization scenarios are also performed in this work.

C. Related work

Timely access to patients healthcare records is mostly essential. As a result, broad access to EHR is mostly provided to users in efforts towards complying with the availability trait of the CIA [14]. However, this opens up the system for abuse and misuse [14]. Based on that, Zhang et al., hypothesized and model the patient care pathways as a progression of a patient through the healthcare system [15]. So the patient flow was modeled as a sequence of accesses of the users defined in the form of a graph and further modeled the trend of patients records accesses which showed deviations from patients care pathways. Graph-based approach was used to model the accesses of patients records in EHR in the context of patient care. A three-month EHR logs was used to evaluate the framework. The framework detected some outliers and deviations of accesses which were different from various types of medical accesses. There was also a high deviation of normal access patterns of nonclinical healthcare workers from clinical users. The ROC curve for the prediction showed 92%, suggesting the performance of the approach was efficient. Additionally, Ziemniak et al used C4.5 decision tree to detect abnormal security behaviour in a healthcare application. Ad-hoc analyis was used to determine atypical behaviour by visually looking for interesting nodes such as path-length investigation [16]. These studies , [15], [16], adopted the general work-flow in accessing patients records which yielded a good results but each of them only adopted to graph-based approach in their work without comparing other methods (such as Nearest neighbor, Bayesian probability methods, support vector machine (SVM) etc) to exhibit their performance measure. Due to the nature of data (eg noisy or not noisy), the performance of algorithms differ. Therefore, in effort to analyse security practice in EHR logs, it is important to compare the performance of algorithms to select the most efficient and effective method for the purpose.

In a related work, Boddy et al tried to avert challenges in restrictions of access controls, often faced by both patients and healthcare workers in EHRs. So human-in-the-loop model was developed by Boddy et al using logs of EHR [17].The model was assessed for anomaly using the human-in-the-loop model with local outlier factor (LOF). A weighted average was applied to each audit log and their respective anomaly scores were computed. The computed average score of the ensemble were plotted against the date and time stamp. The output was visualized for the analysis. The model was able to detect 145 anomalous activities using unlabelled dataset.Additionally, Boddy et al adopted density-based local outlier detection model to profile users' behaviour in relation to their security practice [18]. A local outlier detection factor (LOF) assesses the local deviations from similar group of users (eg doctors, nurses etc) by measuring the isolated distance of a data point to its k-nearest neighbours.

Additionally, Chen et al., developed a framework for anomalous insiders from access logs called community-based anomaly detection system [19]. The detection is based on the behaviour of users and their relation with each other. The model is based on the hypothesis that typical users mostly form and function as communities. So the access logs of users were mined and modeled for the relation of users and their behavioural profiles. Based on the accesses of users, a typical user's accesses of objects should be similar to the peers. For instance, in EHR, a typical user such as a nurse should access similar set of patients records like other nurses due to commonalities in patient care path-ways. Principal Component Analysis (PCA) was relied on to develop an intrusion detection model. The PCA was applied on training features to determine the major and minor principal components for the model. k-Nearest neighbor was was found for each user and calculated the deviation of each of the users from their nearest neighbours. The experimental result showed that, the CAD was able to distinguish anomalous users in a real data log.

While the studies by [17]–[19] showed good performances, a comparative analysis was not also considered. To this end we analyzed simulated EHR logs to compare neural network (nn), LogisticRegression (lr), Random Forrest (rf), decision tree (dt), Support Vector Machine (svm), K-nearest neighbor (knn), Naive Bayes multi-nomial (mulnb) and Naive Bayes binomial(bennb) methods by analysing for deviations from the workflows.

II. Our Method

A. Health record logs data simulation

We developed a hospital information system work-flows to simulate healthcare record log data for this work. This system has five main modules including Patient Management, Pharmacy Management, Laboratory Management, Finance, and Reporting. In this simulation, the hospital is assumed to have 19 departments including IT, Finance, Administration, Laboratory, Pharmacy, Emergency, 10 Out Patients departments (Ear-Nose-Throat, Eyes, Tooth, Child, Orthopedic, Neurological, Gynecological, Diabetes, Rheumatology, and Cancer), with 3 In Patients Departments. Meanwhile, the professions employed by the hospital in this simulation include Head of IT (HIT), Technical Support (TS), Head of Finance (HF), Finance Staff (FS), Head of Administration (HA), Staff of Administration (SA), Head of Lab (HL), Lab Assistant, (LA), Head of Pharmacy (HP), Pharmacy Assistant (PA), Doctor (DO), and Nurse (NU).

Two types of shifts are applied in this simulation: the daily shift and the three 8-hour shift. The daily shift is from 08.00-16.00 Monday to Friday, while three 8-hour shifts involve three shifts each 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 details of the shift and schedule including the number of staffs and professions in each shift are displayed in Table I and II.

TABLE I: Daily Shift

Shift ID	Department	Profession (number of staffs)
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)

The inpatient, outpatient, and emergency department patient flow is shown in Fig. 1, 2, and 3, respectively. Under the described simulation setting and this patient flow, a one-year health record log data are simulated starting from 1 January 2019 until 31 December 2019. The logs are considered as normal data (non-anomaly). In addition, we also generate some abnormal data by simulating attackers who are assumed to have stolen the passwords of certain users and use it to access records of patients (e.g. identity theft). The attackers are simulated to frequently do not follow the hospital flows and tend to deviate from expected behaviors (e.g. make more transactions than the actual users) [20]. In our simulation system, 21 fields are recorded as depicted in Table III. As the result, 283,678 logs were produced with 274,983 of them are legitimate access while 8,695 of them are anomalous accesses.

TABLE II: Three 8-hour shift

Shift ID	Department	Profession (number of staffs)
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 used in this work is using a data-driven method by applying machine learning classifiers. Data-driven is a popular approach in anomaly detection and proven

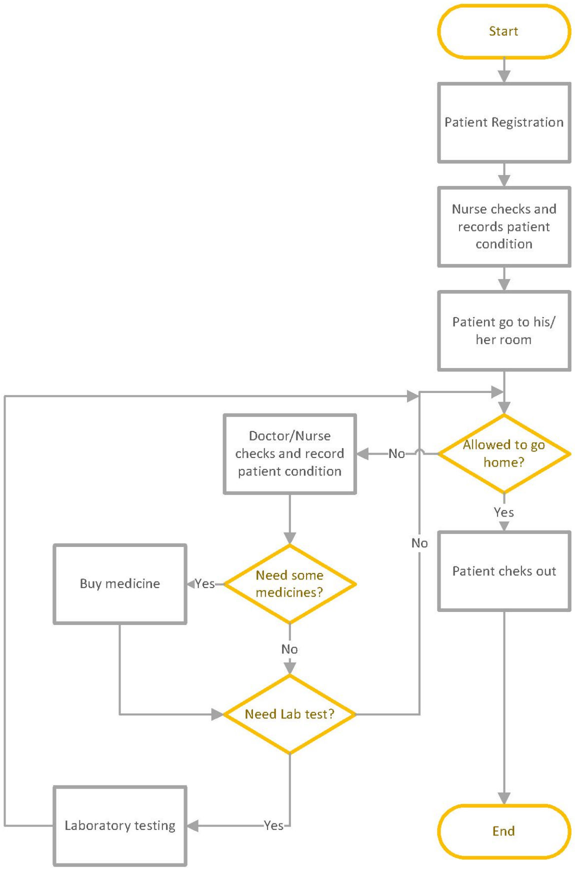


Fig. 1: The Inpatients Department Flow

to get promising performance [21], [22]. The classifier learns from the labeled dataset to determine normal and abnormal activities. This model is then used to detect deviations or anomaly from normal behavior.

1) Data Preparation: Each log record in the dataset portrays a single activity of a user. In order to get a good understanding of the user's behavior, we need to combine several activities from a particular period. Hence, in this work, the raw log data were processed into 24-hour blocks such that an instance reflects a user's cumulative activity in a single day. 24,648 instances were extracted from the raw logs as the outcome with 24,286 of them are labeled as normal data and 362 of them are labeled as anomaly data. The instances are labeled as an anomaly when they had at least one abnormal log access in a single day while the labeled normal instances are all the instances whose all logs are legitimate access.

2) Feature Extraction: After the log combination process, several features are extracted from each instance. The list of features used in this work is displayed in Table IV. In addition, we also applied the Min-Max normalization method and Chi-Square based feature selection to the feature data.

3) Anomaly Detection: After the features are extracted, we use 9 machine learning methods including Multinomial Naive Bayes (multnb), Bernoulli Naive Bayes (bernnb), Gaussian Naive Bayes (gaussnb), Support Vector Machine (svm), Neural Network (nn), K-Nearest Neighbours (knn), Logistic Regression (lr), Random Forest (rf), and Decision Tree (dt). To evaluate the methods, the data is divided into training and testing. Logs from January until August are employed as training data while logs from September until December are utilized as testing data. Using the extracted features and the training data, each machine learning classifier is trained. Furthermore, this trained classifier is employed to do the anomaly detection task. Accuracy, precision, recall, and f1-measure are the metric used for evaluation.

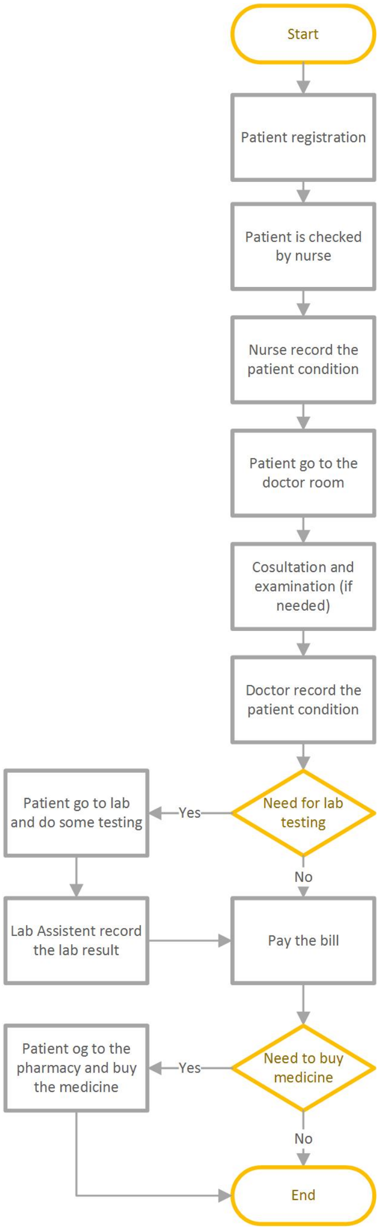

Start
Patient registration
Patient is checked by nurse
Nurse record the patient condition
Patient go to the doctor room
Cosultation and examination (if needed)
Doctor record the patient condition
Need for lab testing
Yes
Patient go to lab and do some testing
No
Lab Assistent record the lab result
Pay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

Fig. 2: The Outpatients Department Flow

C. Performance Evaluation

Several assessments were employed to evaluate the result including Accuracy (Acc), Precision (Prec), Recall (Rec), and

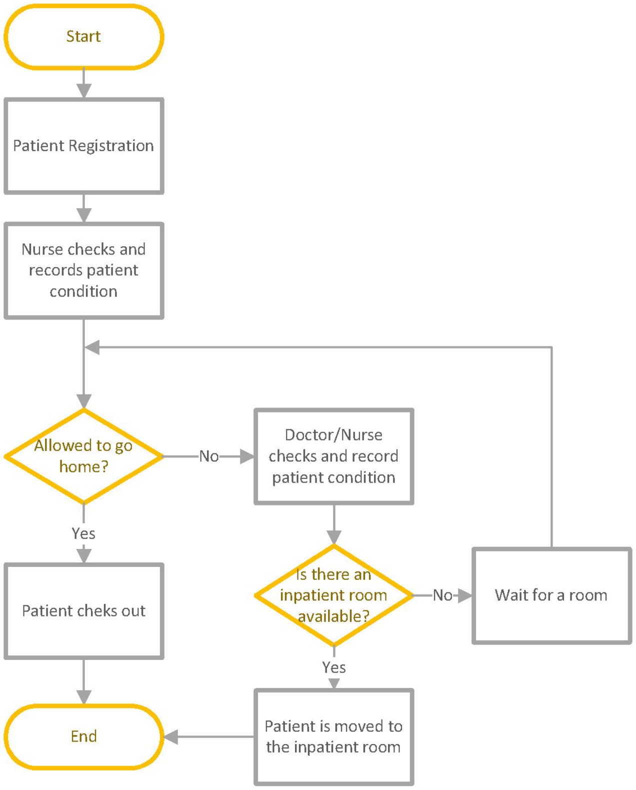
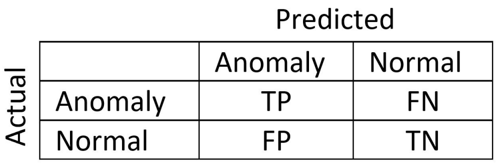

Start
Patient Registration
Nurse checks and records patient condition
Allowed to go home?
No
Doctor/Nurse checks and record patient condition
Yes
Is there an inpatient room available?
No
Wait for a room
Patient cheks out
Yes
End
Patient is moved to the inpatient room

Fig. 3: The Emergency Department Flow

TABLE III: Record Fields

Number	Field Name	Description
1	start_access_time	The starting time the staff accesses the patient record.
2	end_access_time	The end time the staff accesses the patient record access.
3	staff_ID	The ID of the staffs who do the activity
4	role_ID	The role of the staff who access the patient record
5	patient_ID	The ID of the patient whose record is being accessed
6	activity_ID	The ID of the activity (1: Create, 2: Read, 3:Update, 4: Delete)
7	staff_department_ID	The department of the staff who do activity
8	staff_organization_ID	The organization of the staff who access the patient record
9	device_ID	The ID of the computer used by the staff to access patient record
10	browser_ID	The browser used by the staff to access patient record
11	ip_address	The IP Address of the computer used by the staff to access patient record
12	reason_ID	The reason of staff access the patient record (optional)
13	shift_ID	The ID of shift the staff belong to on the day of patient access record
14	sift_start_time	The start time of shift the staff belong to on the day of patient access record
15	sift_end_time	The end time of shift the staff belong to on the day of patient access record
16	module	The module acessed by the staff



Predicted

Actual

	Anomaly	Normal
Anomaly	TP	FN
Normal	FP	TN

Fig. 4: Confusion Matrix

F$_1$-score (F$_1$). We calculated all of the evaluation measures based on the confusion matrix shown in Fig. 4. True Positive (TP) and True Negative (TN) are the numbers of data that were correctly classified. TP is the number of anomaly data correctly classified into the anomaly category, while TN is the number of normal data correctly classified into the normal category. Meanwhile, False Positive (FP), or also referred to as Type I Error, is the number of data that actually belongs to the normal category but incorrectly classified into the anomaly category. On the other hand, False Negative (FN) or Type II Error is the number of anomaly data incorrectly classified as normal data. The formulas for the measurement are as follows:

$$Acc = \frac{TP + TN}{TP + FP + FN + TN} \tag{1}$$

$$Prec = \frac{TP}{TP + FP} \tag{2}$$

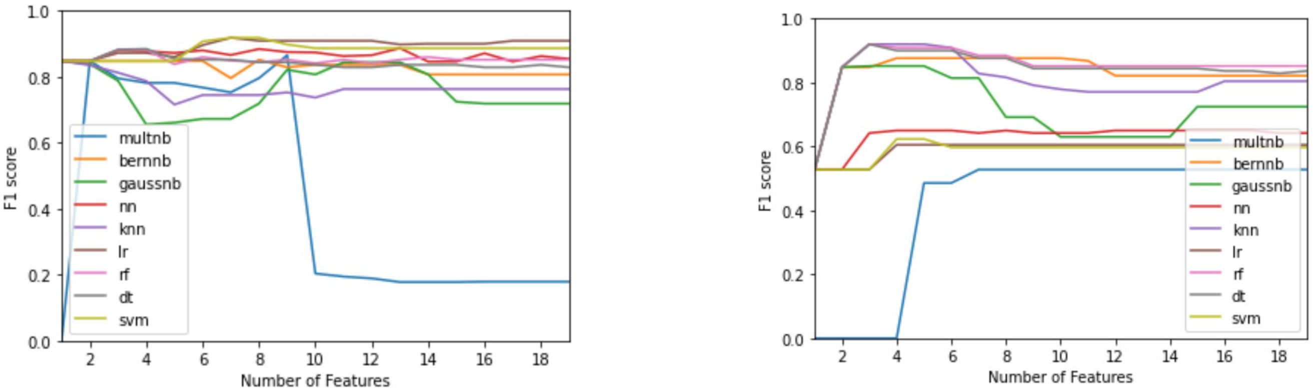
$$Rec = \frac{TP}{TP + FN} \tag{3}$$

$$F_1 = 2\frac{P \cdot R}{P + R} \tag{4}$$

III. Experiment Results and Discussion

Due to the sensitive nature of healthcare records, it is sometimes challenging for hospitals to overcome the legal and regulatory hurdles in order to provide such logs for experimental purposes. Meanwhile, the importance to assess algorithms to select the most effective and efficient ones for real implementation can not be downplayed. So the better option is to simulate such related data logs for the data analysis [23]. Therefore, we simulated EHR logs to analyze the security practices of healthcare professionals.

The anomaly detection results on the non-normalized data are displayed in Table V. In terms of accuracy, generally, Neural Network, Logistic Regression, Random Forrest, and Support Vector Machine have the best result with 0.998. However, the accuracy value difference between the best performing methods and the other methods is very insignificant with Multinomial Naive Bayes and Gaussian Naive Bayes as the worst methods still have an accuracy value of 0.955. In terms of precision, generally, Bernoulli Naive Bayes, Gaussian Naive Bayes, Random Forrest, and Decision Tree have the best



(a) Result on Non-Normalised Data

(b) Result on Normalized Data (Min-Max)

Fig. 5: F1-Score of Anomaly Detection with Number of Feature Variance.

TABLE IV: Dataset Feature Name and Description

Feature Name	Description
create_activity	Number of 'create' transactions conducted in a single day
rea_activity	Number of 'read' transactions conducted in a single day
update_activity	Number of 'update' transactions conducted in a single day
delete_activity	Number of 'delete' transactions conducted in a single day
all_patient_record	Number of access to the patient records in a single day
unique_patient	Number of unique patients whose records has been accessed in a single day
modules	Number of kind of modules in the information system accessed in a single day
report_module_access	Number of transactions conducted in the report module in a single day
finance_module_access	Number of transactions conducted in the finance module in a single day
patient_module_access	Number of transactions conducted in the patient management module in a single day
lab_module_access	Number of transactions conducted in the laboratory module in a single day
pharmacy_module_access	Number of transactions conducted in the pharmacy module in a single day
outside_access	Number of transactions conducted from outside hospital network in a single day
browsers	Number of browser type used in a single day
number of chrome	Number of chrome browser used in a single day
ie_access	Number of Internet Explorer browser used in a single day
safari_access	Number of Safari browser used in a single day
firefox_access	Number of Firefox browser used in a single day
otherbrowser_access	Number of other browser used in a single day

TABLE V: Anomaly Detection Result on Non-normalized Data

Method	Acc	Prec	Rec	F1
multnb	0.955	0.735	0.101	0.178
bernnb	0.997	**0.867**	0.754	0.807
gaussnb	0.995	**0.867**	0.613	0.718
knn	0.997	0.698	0.840	0.762
nn	**0.998**	0.830	0.880	0.854
lr	**0.998**	0.849	**0.978**	**0.909**
rf	**0.998**	**0.867**	0.836	0.851
dt	0.997	**0.867**	0.807	0.836
svm	**0.998**	0.811	0.977	0.886

TABLE VI: Anomaly Detection Result on Min-Max based Normalized Data

Method	Acc	Prec	Rec	F1
multnb	0.995	0.358	**1.000**	0.527
bernnb	0.997	**0.867**	0.779	0.821
gaussnb	0.995	**0.867**	0.621	0.724
knn	0.997	0.698	0.948	0.804
nn	0.996	0.490	0.928	0.641
lr	0.996	0.433	**1.000**	0.605
rf	**0.998**	**0.867**	0.836	**0.851**
dt	0.997	**0.867**	0.793	0.828
svm	0.996	0.433	0.958	0.597

result with 0.867. KNN has the lowest precision value with 0.698 and it is quite far below the other methods. Meanwhile, Logistic Regression obtained the best recall and F1 values with 0.978 and 0.909 respectively. Multinomial Naive Bayes, on the other hand, achieved the worst recall and F1 values with 0.101 and 0.178 respectively.

Concerning the accuracy performance, it is important to note that the dataset is unbalanced. Since anomaly data is very rare, most of the dataset for anomaly detection is unbalanced with normal data far more than the anomaly data (e.g. [24], [24]). In this case, it is not effective enough to determine the performance of the methods based on accuracy alone. In the anomaly detection task, since we want to detect an anomaly, we consider the anomaly data as positive data and normal data as negative data. Since the number of negative data is far higher than positive data, the number of TN, in this case, tends to be very high as well. A method with a low TP could still have very good accuracy because the TN is very high. In other words, even if it can not detect the anomaly, a method may still have good accuracy. Even in an extreme case, when the data is highly unbalanced, the accuracy of a method would still very good even though the method predicts all of the data as normal. Hence, accuracy alone is not suitable for the anomaly detection task evaluation if the dataset is unbalanced. Other evaluation methods such as precision, recall, and F1 are needed.

One of the examples of the previously described case can be seen in Table V where Multinomial Naive Bayes has very good accuracy but a very low F1-score. This method only classifies very little data as anomaly so that the recall is very low. Almost all of the data, including a lot of anomaly data, are classified into normal category. As the consequence, the F1-score becomes very low as well. This method cannot be considered good because it misses a lot of anomaly data and considers them as normal. It could be dangerous because there would be many attacker's actions that would be considered normal if we use this method.

Overall, almost all of the machine learning methods employed for non-normalized data in this work achieved good results except Multinomial Naive Bayes. The explanation about this case can be seen in Fig. 5a. There are many irrelevant or noisy features in the dataset. We can see from the figure that almost all of the methods have a very good F1-score even though only use a few features. It means that only some features that can distinguish the anomaly data from normal data well. Some other features are irrelevant or noisy because it can separate the two categories well. In some cases, the noisy features can decrease the performance. Multinomial Naive Bayes suffer a lot from the noisy features. The noisy features give a big influence on this method so that the performance decrease significantly.

The use of normalization in this work generally cannot increase the performance. In fact, it decreases the performance of some methods such as Neural Network, Logistic Regression, and SVM. The explanation about this case can be found in Fig. 5b. As described before, there are some noisy features in the dataset and some features are more determinant than others. The important features such as outside_access have a high difference value between that in normal and anomaly data. Normalizing the feature value makes all features appear on similar scales and are all treated as equally important. This condition can decrease the model performance. Therefore, as displayed in Fig. 5b, the result of normalized data is generally equal to the result of non-normalized data. After 6 or more features, the performance on the normalized data starts to decline because the method starts to deal with noisy data.

The use of feature selection is proven to give an improvement in the anomaly detection task in this work because of the existence of some noisy data. Chi-square is able to select the best feature used to detect an anomaly. The use of Chi-square based feature selection can improve the performance of all machine learning methods employed by removing several noisy features. Based on the results in Fig. 5, the optimal number of features is between 3-6. The other benefit of the use of feature selection is reducing the time complexity to make the method perform faster [25].

IV. Conclusion and future work

Due to the surge in data breaches within healthcare, there is the need to determine the causes in various ways including big data context [10]. This is because users often leave their traces of accesses in access logs which when analyzed can provide knowledge of access deviations from workflows [13]. To this end, this paper analyzed the workflows of healthcare staff's security practices in simulated electronic health records (EHR) logs to determine anomalous security practices with different classification types of machine learning algorithms. The EHR logs were simulated based on healthcare workflow scenarios. A number of machine learning algorithms were used to analyze the logs for deviations of accesses from the workflow, which is termed as anomaly accesses. The algorithms used were then compared in terms of their performance including accuracy (Acc), precision (Prec), recall (Rec), and F-Measure (F1).

Overall, based on the experiment results, generally, all of the machine learning methods employed in this work perform very well to detect an anomaly. The best performance on the non-normalized dataset is achieved by the Logistic Regression method with accuracy, precision, recall, and F1 value of 0.998, 0.849, 0.978, and 0.909 respectively. Meanwhile, on Normalized data, Random Forest obtained the best result with accuracy, precision, recall, and F1 value of 0.998, 0.867, 0.836, and 0.851 respectively. However, The use of normalization generally could not increase the performance of the used methods. The results also reveal that there are several noisy features that can affect performance. Therefore, the use of Chi-square as feature selection is very important in this work. This method can select the best feature and remove the bad features so that the performance can be improved. Based on the experiment results, the optimal number of features is between 3-6. The additional advantage of reducing several features is we can also reduce the time complexity to make the method run faster. Therefore, it can be concluded that the proposed workflow based approach can be adopted with the well-performed algorithms for analyzing security practice in real logs of EHR.

It however remains challenging to detect malicious security practice if a malicious actor follows the workflow to access healthcare records with legitimate access rights. Additionally, future works need to further assess the detected anomalies for maliciousness. This will provide more knowledge for the appropriate security measures to be taken. Another aspect that needs to be considered in future works includes assessing the logs with unsupervised methods to compare their performance to be used in scenarios where the EHR logs are unlabeled.

References

[1] B. Hjellen, “Hacking scandal shakes finland - patients pressured for money.” [Online]. Available: ” https://www.nrk.no/urix/hacking-skandale-ryster-finland—pasienter-presset-for-penger-1.15214710

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

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

[4] M. E. Whitman, P. Fendler, J. Caylor, and D. Baker, “Rebuilding the human firewall,” in *Proceedings of the 2nd annual conference on Information security curriculum development*, 2005, pp. 104–106.

[5] HIMSS, “Health share of oregon: 654,000 patients.” [Online]. Available: ”https://healthitsecurity.com/news/the-10-biggest-healthcare-data-breaches-of-2020-so-far”

[6] M. Butavicius, K. Parsons, M. Pattinson, and A. McCormac, “Breaching the human firewall: Social engineering in phishing and spear-phishing emails,” *arXiv preprint arXiv:1606.00887*, 2016.

[7] 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.

[8] A. Martins and J. Elofe, “Information security culture,” in *Security in the information society*. Springer, 2002, pp. 203–214.

[9] M. A. Sasse, S. Brostoff, and D. Weirich, “Transforming the ‘weakest link’—a human/computer interaction approach to usable and effective security,” *BT technology journal*, vol. 19, no. 3, pp. 122–131, 2001.

[10] 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.

[11] IBM, “The 2020 cost of a data breach report explores financial impacts and security measures that can help your organization mitigate costs.” [Online]. Available: ” https://www.ibm.com/security/data-breach”

[12] 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

[13] 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.

[14] 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.

[15] H. Zhang, S. Mehotra, D. Liebovitz, C. A. Gunter, and B. Malin, “Mining deviations from patient care pathways via electronic medical record system audits,” *ACM Transactions on Management Information Systems (TMIS)*, vol. 4, no. 4, pp. 1–20, 2013.

[16] 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.

[17] 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.

[18] 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.

[19] Y. Chen and B. Malin, “Detection of anomalous insiders in collaborative environments via relational analysis of access logs,” in *Proceedings of the first ACM conference on Data and application security and privacy*, 2011, pp. 63–74.

[20] A. Nandi, A. Mandal, S. Atreja, G. B. Dasgupta, and S. Bhattacharya, “Anomaly detection using program control flow graph mining from execution logs,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 215–224.

[21] M. Du, F. Li, G. Zheng, and V. Srikumar, “Deeplog: Anomaly detection and diagnosis from system logs through deep learning,” in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017, pp. 1285–1298.

[22] S. Lu and R. Lysecky, “Data-driven anomaly detection with timing features for embedded systems,” *ACM Transactions on Design Automation of Electronic Systems (TODAES)*, vol. 24, no. 3, pp. 1–27, 2019.

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

[24] H. Studiawan and F. Sohel, “Performance evaluation of anomaly detection in imbalanced system log data,” in *2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4)*. IEEE, 2020, pp. 239–246.

[25] M. A. Fauzi, A. Z. Arifin, S. C. Gosaria, and I. S. Prabowo, “Indonesian news classification using naïve bayes and two-phase feature selection model,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 2, no. 3, pp. 401–408, 2016.