Computer Modeling in Engineering & Sciences

Tech Science Press

DOI: 10.32604/cmes.2022.018163

ARTICLE

An Analysis of Integrating Machine Learning in Healthcare for Ensuring Confidentiality of the Electronic Records

Adil Hussain Seh[1], Jehad F. Al-Amri[2], Ahmad F. Subahi[3], Alka Agrawal[1], Nitish Pathak[4], Rajeev Kumar[5,6,*] and Raees Ahmad Khan[1]

[1]Department of Information Technology, Babasaheb Bhimrao Ambedkar University, Lucknow, 226025, India

[2]Department of Information Technology, College of Computers and Information Technology, Taif University, Taif, 21944, Saudi Arabia

[3]Department of Computer Science, Umm Al-Qura University, Makkah, 21421, Saudi Arabia

[4]Department of Information Technology, Guru Gobind Singh Indraprastha University, Delhi, 110078, India

[5]Department of Computer Applications, Shri Ramswaroop Memorial University, Barabanki, 225003, India

[6]Department of Computer Science and Engineering, Babu Banarasi Das University, Lucknow, 226028, India

[*]Corresponding Author: Rajeev Kumar. Email: rs0414@gmail.com

Received: 04 July 2021 Accepted: 14 October 2021

ABSTRACT

The adoption of sustainable electronic healthcare infrastructure has revolutionized healthcare services and ensured that E-health technology caters efficiently and promptly to the needs of the stakeholders associated with healthcare. Despite the phenomenal advancement in the present healthcare services, the major obstacle that mars the success of E-health is the issue of ensuring the confidentiality and privacy of the patients' data. A thorough scan of several research studies reveals that healthcare data continues to be the most sought after entity by cyber invaders. Various approaches and methods have been practiced by researchers to secure healthcare digital services. However, there are very few from the Machine learning (ML) domain even though the technique has the proactive ability to detect suspicious accesses against Electronic Health Records (EHRs). The main aim of this work is to conduct a systematic analysis of the existing research studies that address healthcare data confidentiality issues through ML approaches. B.A. Kitchenham guidelines have been practiced as a manual to conduct this work. Seven well-known digital libraries namely IEEE Xplore, Science Direct, Springer Link, ACM Digital Library, Willey Online Library, PubMed (Medical and Bio-Science), and MDPI have been included to perform an exhaustive search for the existing pertinent studies. Results of this study depict that machine learning provides a more robust security mechanism for sustainable management of the EHR systems in a proactive fashion, yet the specified area has not been fully explored by the researchers. K-nearest neighbor algorithm and KNIEM implementation tools are mostly used to conduct experiments on EHR systems' log data. Accuracy and performance measure of practiced techniques are not sufficiently outlined in the primary studies. This research endeavour depicts that there is a need to analyze the dynamic digital healthcare environment more comprehensively. Greater accuracy and effective implementation of ML-based models are the need of the day for ensuring the confidentiality of EHRs in a proactive fashion.

KEYWORDS

EHRs; healthcare; machine learning; systematic analysis

1 Introduction

The escalating advancements in technology, especially in smart devices, the Internet of Things, internet connectivity, and cloud services play a significant role in customer service enhancement. Technologies are adopted by business entrepreneurs and other organizations throughout the world to provide efficient, reliable, and cost-effective services to their customers [1,2]. The healthcare industry is also one among them [3]. From the last few years, the healthcare sector has shown a significant enhancement in providing digital care services. Today, Electronic health records (EHRs) systems are widely implemented by the healthcare service providers to provide efficient and easily accessible services [4,5]. EHR systems store and manage health records of patients and other relevant data. These systems provide the up-to-date history of patients' health and quick access to patient's data that helps the practitioners to make a better diagnosis and provide effective treatments [5,6]. The primary objective of EHR systems is to deliver comfortable and effective care services [5,7]. As the implementation of EHR systems increases, concerns regarding the healthcare data privacy and confidentiality have also increased [5,8]. The present-day privacy and security of digital data have become a serious issue for both the care providers and patients. Eminent reports and research studies depict that from the last few years, the healthcare sector has been the main target of both the insider as well as outsider intruders. This is evident by the figures that cite that the highest number of data breaches were registered in the healthcare sector [3,9,10]. Different techniques and methods have been practiced and implemented by the experts and the researchers to secure the electronic health record processing. Both cryptographic, as well as non-cryptographic techniques, have been applied in different environments to achieve confidentiality and privacy in healthcare data [11,12]. But there are very few studies that implement machine learning approaches to improve the confidentiality and privacy of healthcare data [13]. The reasons may be different but the important ones are the availability of healthcare smart system log data and the dynamic nature of the healthcare environment. Simulation of a dynamic healthcare environment is a complex task and needs an exhaustive and comprehensive analysis.

Machine learning (ML) as a subdomain of Artificial Intelligence (AI) provides ways for systems or programs to learn from data and past experiences to improve system performance with the minimum human intervention [14,15]. At every correct decision, a computer program improves its performance measure [16] by itself without the involvement of human intervention. ML algorithms have a great ability to learn from historical data and help the computers to solve different complex real life problems in various fields. Recently AI, machine learning, and Data Mining technologies have changed the way people think and played a significant role in different fields of life. Moreover, ML has got significant importance in the field of cyber security [17]. Different machine learning techniques namely, Bayesian Networks [18], Decision Trees [19,20], Artificial Neural Networks [21,22], Association Rule [23,24], Support Vector Machine [25,26], Regression Trees [27], K-nearest neighbors [28,29], Random Forest [30,31] have been implemented to detect and predict cyber intrusions. But there are very few research studies that try to explore machine learning approaches to detect suspicious access or intrusions to EHR systems. Further, there is yet no systematic review available that specifically focuses on ML approaches and the possible significance of these approaches in improving healthcare data confidentiality by detecting suspicious access to the EHRs.

More specifically, in this context, the authors of the present study enlisted the search query ("Machine learning" or "ML") and ("Healthcare" or "Electron Health record" or "Personal health record" or "Patient health record" or "health information systems" or "Electronic Medical Record") and ("Security" or "confidentiality" or "Privacy") and ("Systematic review" or "SLR" or "Survey") on 2nd December 2020 in the well-known libraries specified in Table 1, many times with some other relevant key terms. However, we did not find any systematic review paper that could provide us with satisfactory answers to our specified research questions. The most relevant surveys and systematic reviews found from executing this search string have been discussed in Section 2– as a subsection under the heading "Motivation and Problem Specification". Thus, there is an apparent need to carry out a systematic literature review on those research studies that use ML approaches to improve healthcare data's confidentiality. Such a premise will help the researchers to identify and understand the existing research work in a summarized form on a specified theme in one place. Furthermore, the intended Systematic literature review (SLR) will also cite the unexplored research gaps in the specified domain and deliver a comprehensive summary of the available evidence on the specified theme.

Table 1: Number of retrieved papers from the specified online libraries

Online libraries	Key words used	Number of retrieved results
IEEE Xplore	Suspicious access; Illegitimate access; Malicious access; Anomaly detection; Hacking; Ransomware; Outlier access detection; Fraud access detection; Confidentiality; Privacy; Electronic health Record; Patient Health Record; Electronic patient Records; Electronic Medical Record; Health information systems; Healthcare data; Machine learning; ML; Supervise machine learning; Unsupervised Machine learning; ML.	128
Science Direct		122
Springer Link		159
ACM Digital Library		182
Willey Online Library		26
PubMed		28
MDPI		14
Total		659
Google Scholar		735

Systematic Literature Review is a methodically structured and formal approach to make a review study by using, precise, and explicit methods to identify, select, and critically assess suitable research, and to compile and investigate data from the studies that are included in the review, and is devised as an extensive and productive process [32]. A systematic review can help a researcher to elicit the existing solutions and find a new way to tackle a given problem before addressing the interesting problem. Systematic review minimizes biases in the work and helps the researchers to find knowledge gap, and identify the areas which require further research [33]. Formerly, this methodology was designed for the healthcare reviews and Meta-analysis, but later it was found to be beneficial for various natural and social science fields, including information systems [34] and software engineering [35]. More pertinently, the primary focus of this study is to carry out a systematic process on the existing studies that uses the ML approaches to address the privacy and confidentiality issues in the healthcare data. Moreover, the study also seeks to investigate and analyze their findings and to summarize their results systematically. Thus, the objective is to identify the unexplored gaps and areas that need urgent focus to bridge the research gap. The core effort of this study has also been to identify the existing parameters that were considered by researchers in their pursuits and find out new parameters that would help the researchers to strengthen their future research in a better way.

The rest of this research work has been organized in the following way: Section 2, of this work, briefly discusses the existing related work. Section 3 discusses the systematic methodology adopted in this study and under the subsections of: 1) planning phase; 2) conducting phase, and 3) reporting phase. It includes framing of research questions; evaluation and selection process of primary studies; data extraction and synthesis; results and findings of this work. Thereafter, Section 4, provides discussion and future work, and then Section 5 enlists the validation of the work. Section 6 concludes the study.

2 Existing Related Work

A systematic literature review provides a scientific way to investigate the existing literature on a specified theme. Various systematic reviews and surveys have been carried out by the researchers to sort out the security and privacy issues and challenges faced by the healthcare industry. Some of the pertinent studies are discussed here. In 2020, Newaz et al. [36] proposed a survey that enlists modern healthcare systems and what are the privacy and security issues and challenges faced by them. In this study, the authors have discussed the potential threats and attacks and their impact on the healthcare industry. Qayyum et al. [37] have done a survey on secure machine learning (ML) for healthcare. In their study, the authors enlisted the applications of ML in healthcare sector and security issues and challenges of ML approaches in healthcare. The work also discussed the solutions to use ML for protecting the privacy of the healthcare data. In 2019, Da-Costa et al. [38] proposed a survey study on intrusion detection in the Internet of Things (IoT) through ML and evolutionary computation. This study discussed the security issues and challenges of the Internet of Things (IoT) systems. Ghosal et al. [39] performed an extensive survey on cloud based IoT that used ML for data analysis and cyber security in healthcare. This study proposed a model for data analysis and secure data access. In 2017, Kruse et al. [40] carried out a systematic review on healthcare cyber security to identify the potential threats with their trends and find out the existing solutions for them. Luna et al. [41] proposed a systematic review to enlist and discuss the potential threats that are carried out by cyber-criminals to expose healthcare data from the healthcare industry. Buczak et al. [42] carried out a comprehensive survey on data mining (DM) and ML approaches that have been used for intrusion detection. This study does not focus on a particular sector but includes all the studies that use DM and ML approaches for intrusion detection, dwelling on both misuse and anomaly detection. Rahim et al. [43] carried out a systematic review that describes the privacy issues and challenges faced by the healthcare industry while handling electronic medical records. In 2013, Fernández-Alemán et al. [44] undertook a systematic review to find out the solutions, approaches and standards that are used to address security and privacy issues and challenges of EHRs in the healthcare industry. In 2020, Yeng et al. [13] proposed a systematic review study on data-driven and AI methods used for healthcare security services.

These systematic reviews and survey studies are the most common finds that have been completed on healthcare data security issues and challenges. All these studies have partially or completely adopted the systematic guidelines to make systematic review studies. However, we did not find a complete systematic review study on the machine learning based studies that addressed the healthcare data's privacy issues. We have tried to incorporate all the existing original machine learning studies which are associated with the healthcare data's privacy and confidentiality issues. Henceforth, our main objective in this research endeavour is to make a comprehensive systematic review study on machine learning based studies that address healthcare data's privacy issues.

2.1 Motivation and Problem Specification

In comparison to the above discussed surveys and systematic review studies, the survey article [13] was the most relevant one for our proposed study (identified research questions Q1 to Q4 in Section 3.1 below). However, the stated study did not provide a complete systematic process while conducting the review; for example, there is no specification of research questions, quality assessment, primary study synthesis and validation test, etc. The study was also not conducted with respect to three different phases of an SLR. All these surveys and systematic review studies investigated the original research works and summarized the results to provide better ways for future research in the specified domains. However, we were not able to find a systematic review that satisfied our specified objectives or research questions. Moreover, the ML algorithms, with proactive characteristics, make ML models more efficient in detecting suspicious user accesses against the EHR systems as compared to other data security approaches. Therefore, the implementation of ML models in the digital healthcare environment will help the care service providers to protect data against unauthentic and abnormal user accesses. If not identified at the early stages, this may lead to disastrous data breaches later. Thus, with other data security technologies namely cryptography and blockchain, the adoption of ML in EHR systems as a proactive data security mechanism will bring the much needed change in digital healthcare. It is noticeable that cryptographic techniques namely Secret key cryptography, Public key cryptography, and Hash functions help us to make information and information communication systems secure through different encryption and decryption approaches. Similarly, more effective blockchain technology provides a secure mechanism for storing and processing sensitive EHRs and other data in healthcare environment. However, both the technologies cryptography and blockchain are not able to detect suspicious user accesses against the healthcare information systems. For that, the ML based model plays a vital role by using the EHR system log data. All these different technologies have their own benefits and one cannot replace the other. Henceforth, there integrated implementation in digital health information systems will provide a more secure environment for healthcare data.

In addition to this, the healthcare industry is one of the top industries that stores and carries the most sensitive and confidential data. But improper security measures and design flaws make the healthcare industry more vulnerable to threats and easily accessible to intruders [36]. The last few years have recorded not only growing instances of intrusions by cyber-criminals but also an increase in the number of insider attacks. According to the Data breach research studies and reports, the number of threats and data breach incidents reported by the healthcare service providers has now become a cause of major apprehension. It has been found that from 2010 to 2019, there were 3,033 healthcare data breach incidents that collectively exposed 255.18 million records [3]. Commonly used attacks to expose confidential health data are Hacking/IT incidents, unauthorized access/internal disclosure, Theft/loss, or Improper disposal. Fig. 1, given below, represents different disclosure types that have exposed confidential heath data from 2010 to 2019 [3]. The graph shows a significant growth in Hacking/IT incidents, unauthorized internal access type disclosures in recent years.

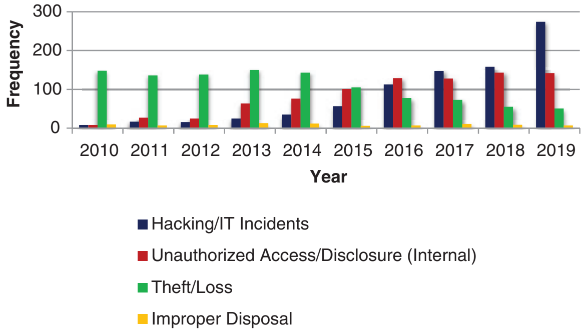

Figure 1: Different types of healthcare data disclosure

Indeed, the above discussed facts and figures depict that there is a need to address these issues faced by the healthcare industry. Furthermore, the literature survey shows that there is no specific SLR that specifically discusses and describes the machine learning approaches that are used to detect suspicious or unauthorized access to the Electronic Health Records or EHR systems in a proactive mode. Our goal is to follow the complete SLR process and find out the research studies that use ML approaches to identify suspicious or unauthorized access to EHRs. From the collated data we intend to draw out the enhancive role of ML in improving healthcare data confidentiality.

3 Adopted Methodology for SLR

In this paper, we have made a systematic review of prior work that comprises the "role of machine learning to preserve or improve the confidentiality of healthcare data or Electronic Health Records (EHRs)". A systematic review [32] commonly consists of the following three steps:

(1) Planning.
(2) Conduction or extraction.
(3) Reporting.

In the ensuing sub-sections, we will discuss our proposed work according to these three steps one by one.

3.1 Planning Phase

Planning is the first phase of the systematic review process in which we identified the needs for a specific systematic literature review. The need and significance of this specific systematic review study are discussed in the Introduction section of this paper. Here, the aim was to identify the machine learning approaches that have been used to preserve the confidentiality of electronic health records (EHRs) on EHR systems. More essentially, for examining how machine learning approaches detect anomalous accesses while accessing the EHR systems, we framed the following research questions:

RQ1. Does the research study use a machine learning approach to ensure confidentiality of sensitive healthcare data in an EHR environment?

RQ2. Does the study discuss healthcare data confidentiality issues?

RQ3. Does the study include experimental results with a theoretical framework?

RQ4. Does the study use real or synthetic data of any healthcare organization to train and test the model?

The period for this systematic review paper was from 2010 to 2020 because we did not find any relevant research study that had been published before 2010, and a study that completely addressed our research questions and fulfilled our inclusion criteria. The authors of this study conducted the systematic review as per the guidelines published by Kitchenham to attain the objective of answering the specified research questions [34]. To make a comprehensive assessment of the systematic review, we undertook the planning, conducting, and reporting steps of a systematic literature review.

3.2 Conducting Phase

Conducting or extraction phase is the second step of the SLR process. It immediately starts after the planning phase. It includes research identification, primary studies selection, quality assessment, data extraction, data synthesis (summarization) [33,35–39]. Research identification ensures the retrieval of all the research studies related to the research questions. For that, we defined a search strategy that includes specific search terms, formation of search strings, and the most common computer science and health science online digital research libraries named as:

• IEEE Xplor	https://ieeexplore.ieee.org/Xplore/home.jsp
• Science Direct	https://www.sciencedirect.com/
• Springer Link	https://link.springer.com/
• ACM Digital Library	https://dl.acm.org/
• Willey Online Library	https://onlinelibrary.wiley.com/
• PubMed	https://pubmed.ncbi.nlm.nih.gov/
• MDPI	https://www.mdpi.com/
• Google Scholar	https://scholar.google.com/

After specifying the online digital libraries, the defined search terms were based on the population (Anomalous or suspicious access detection in EHR systems to preserve the confidentiality of EHRs) and intervention (Existing machine learning approaches used for intrusion detection). Experts were consulted and the search engines were used to identify the most relevant keywords based on specified search terms. The identified keywords concerning the search terms were:

- Suspicious access-Illegitimate access; Fraud access; Anomalous access; abnormal behaviour.
- Electronic health Records-Personal Health Record; Electronic Medical Record; Health information systems; Healthcare data.
- Machine learning-ML; Supervise machine learning; Unsupervised machine learning.
- Confidentiality-Privacy.

The search strings were constructed by using these keywords as groups. Each group is a collection of terms and these terms depict either the common semantic meaning or distinct forms of the same word. According to the search platform, search queries were executed against the title, keywords, abstract, or full text. The total number of results retrieved from each online library is shown in Table 1. These results were found after applying admissible database filters of the respective database and were generated from May to June 2020.

No doubt there are other research databases and individual journals that might have also published eminent and high quality research. But we have included only seven well known and pertinent digital databases in this study. The reason is that in terms of reference, these databases had maximum index the maximum high quality and relevant research studies with respect to our proposed theme throughout the world. We have also included the most well-known and famous medical online digital database "PubMed" to minimize the rate of missing studies that fulfilled our specified study selection criteria. In addition, this study uses Google Scholar as a tool to find out those research studies that are relevant to our proposed theme and meet the selection criteria but are not indexed by our selected databases. This approach was extremely conversant for segregating and incorporating almost all the pertinent research studies, and for minimizing the missing study rate up to a satisfactory limit.

In the initial step, S1 performed a simple search. The number of retrieved results from all specified online libraries was huge in number (N = 659). To narrow the number of retrieved results and make them more precise to achieve our objective, we performed a step-wise (S1 to S5) search process as shown in Fig. 2. At Step S2, a manual search was performed based on titles of papers, and 243 studies were removed as irrelevant studies; then at Step S3, further investigation was carried out by reading the abstracts and keywords of the 416 selected studies to remove duplicates and less relevant papers. Based on the abstracts, keyword, and conclusion filtration, 307 research studies were removed whereas 109 studies were selected for further filtration.

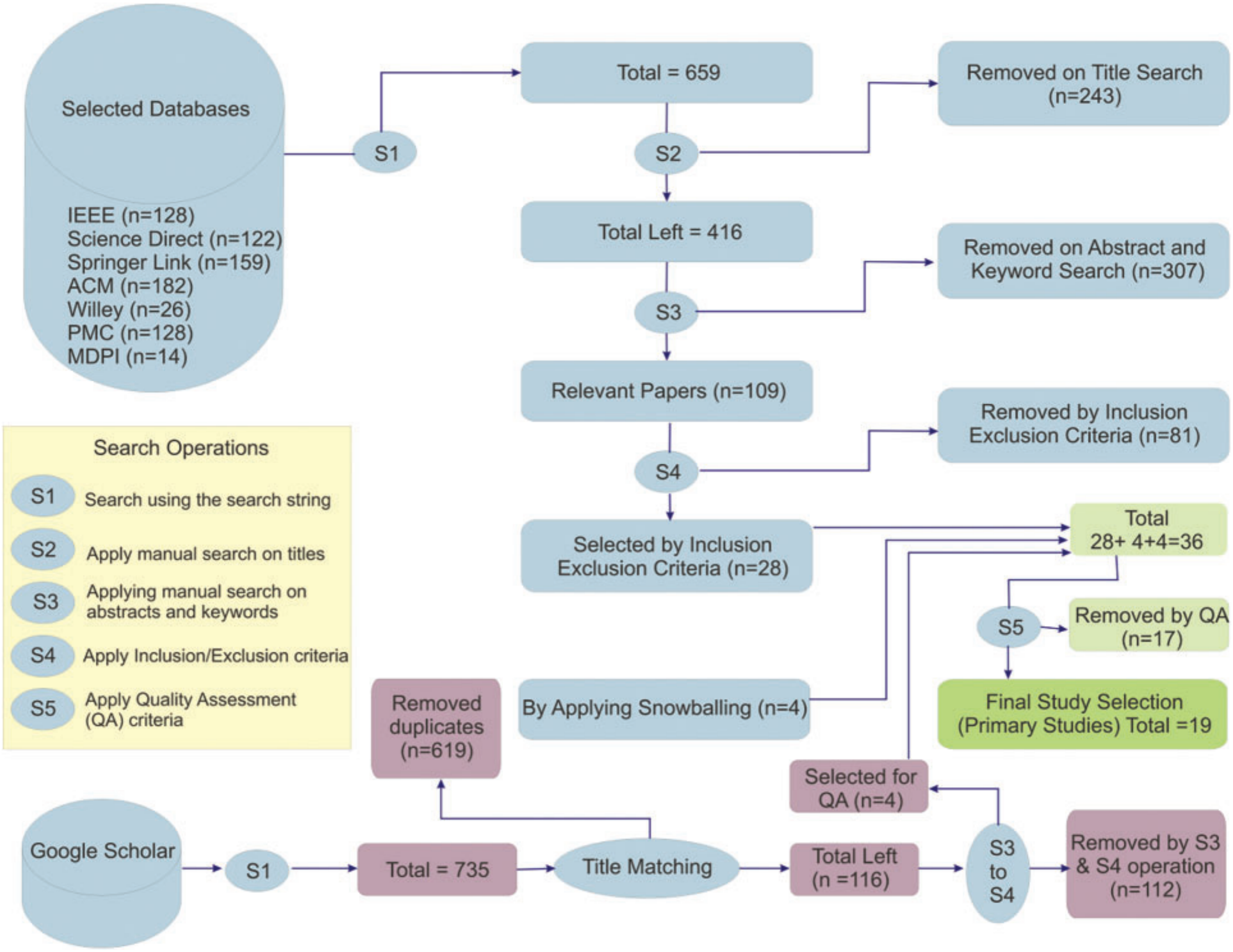
Selected Databases
IEEE (n=128)
Science Direct (n=122)
Springer Link (n=159)
ACM (n=182)
Willey (n=26)
PMC (n=128)
MDPI (n=14)
S1
Total = 659
S2
Removed on Title Search (n=243)
Total Left = 416
S3
Removed on Abstract and Keyword Search (n=307)
Relevant Papers (n=109)
S4
Removed by Inclusion Exclusion Criteria (n=81)
Selected by Inclusion Exclusion Criteria (n=28)
Total 28+ 4+4=36
S5
Removed by QA (n=17)
Final Study Selection (Primary Studies) Total =19
Search Operations
S1 Search using the search string
S2 Apply manual search on titles
S3 Applying manual search on abstracts and keywords
S4 Apply Inclusion/Exclusion criteria
S5 Apply Quality Assessment (QA) criteria
Removed duplicates (n=619)
By Applying Snowballing (n=4)
Selected for QA (n=4)
Google Scholar
S1
Total = 735
Title Matching
Total Left (n =116)
S3 to S4
Removed by S3 & S4 operation (n=112)

Figure 2: Step-by-step process of literature extraction and evaluation

Further, search queries such as ("*Anomalous access" or "Suspicious access" or "misused access" or "abnormal behaviour" or "Fraud access*") and ("*Confidentiality" or "Privacy" or "detection")* *and ("Electronic Health Record" or "Personal Health Record" or "Electronic Medical Record") and ("Machine learning" or "ML"*) were also run in the Google Scholar with the necessary modification. Some other relevant keywords such as *hacking, ransomware, phishing, electronic medical record, patient medical record, and patient health records* have been used in the modified versions of search queries. Retrieved results (research papers) generated by it were compared with the collective results (research papers) extracted from the above specified databases to elicit those research studies that were missing from the above specified databases but were relevant to our specified objectives. Complete graphical representation of this comparative process is also shown in Fig. 2, presented in Subsection 3.2.3 of this study given below.

3.2.1 Selection of Primary Studies

After the completion of Steps S1, S2, and S3 filtration process; the primary study selection process was carried out. In the primary study selection, we applied the inclusion and exclusion criteria to the selected studies. The Inclusion criteria are as:

- The study must be published in the English language.
- The study should answer or be relevant to our specified research questions.
- The study can suggest any approach, method, process, tool, or framework to address data confidentiality issues and challenges.
- The paper should be published in a peer reviewed journal or a conference proceeding.

The Exclusion criteria are as:

- Studies not published in English language and in between specified date.
- Gray literature or duplicate papers.
- Studies irrelevant to our specified research questions.

By applying inclusion-exclusion criteria at Step S4 on 109 selected studies, only 28 studies completely fulfilled the inclusion criteria. By enlisting Snowballing (reference searching), we extracted 4 more relevant research studies which were added to our final selected studies quota that qualified for the QA test. Further, from Google Scholar also, 4 studies were selected for the QA test. Finally, a total of 36 studies were bracketed for quality assessment.

3.2.2 Quality Assessment of the Research Papers

Quality assessment (QA) of the selected studies was performed at Step S5 to obtain the most relevant research studies for literature review. The quality evaluation process of our study is based on the protocol defined by A K-Petersen [33]. Quality evaluation questions have been formulated to assess the overall quality of the selected studies and these questions are defined in Table 2. First, a full text reading of each selected study was carried out then a quality score of each study was independently calculated based on formulated quality questions. The study with a quality score equal to or greater than 50 percent was included in the final list. Q1 and Q2 (quality assessment questions) are mandatory for all the studies to qualify QA test. Based on the independent quality assessment score of each study, only 19 studies were selected for final review as primary studies. Table 3 enlists the quality assessment results of selected primary studies based on quality assessment questions defined in Table 2. Each column from Q1 to Q6 in Table 3, shows that the corresponding primary study is capable of answering the respective quality assessment question or not. If it completely justified the QA question, we gave it 1 as a quality score, if it partially answered the QA question, we gave it a score of 0.5 and if it completely failed to answer the QA question, we marked it as 0 quality scores.

Table 2: Quality assessment questions

Q. No.	Question description
1	Does the research study describe the purpose of our study in a definite way?
2	Does the study propose an approach, method, or framework to achieve healthcare data confidentiality in some way?
3	Does the study present the related work about the main contribution?
4	Does the study highlight the comparative strength of the proposed method with other existing models or mechanisms?
5	Is the proposed method validated on real or synthetic data and the presented research results?
6	Does the study present the conclusions according to the research objectives and future work?

Table 3: Primary study quality assessment results

Ref. of primary studies	Q1	Q2	Q3	Q4	Q5	Q6	Quality score (%)
[45]	1	1	1	0	1	1	83.33
[46]	1	1	1	1	0.5	1	91.66
[47]	1	1	1	0.5	1	1	91.66
[48]	1	1	1	0.5	1	0.5	83.33
[49]	1	1	0.5	0	0.5	0.5	58.33
[50]	1	1	1	0.5	1	0.5	83.33
[51]	1	1	1	0	1	0.5	75.00
[52]	1	1	0	0	1	0.5	58.33
[53]	1	1	1	0.5	1	0.5	83.33
[54]	1	1	0	0.5	1	0.5	66.66
[55]	1	1	0.5	0	0.5	0.5	58.33
[56]	1	1	1	0	1	1	83.33
[57]	1	1	1	0	1	1	83.33
[58]	1	1	1	0.5	1	1	91.66
[59]	1	1	1	0.5	0.5	0.5	75.00
[60]	1	1	1	0	1	0.5	75.00
[61]	1	1	1	0	1	0.5	75.00
[62]	1	1	0.5	0	1	0.5	66.66
[63]	1	1	0.5	0.5	1	0.5	75.00

3.2.3 Data Extraction and Synthesis

Data extraction is defined as the elicitation and recording of the information obtained from the primary studies that have passed the QA test [35]. The information includes context data that provides the purpose of the study, quantitative data includes experimental results and numerical values, and qualitative data that comprises findings and conclusions of the research studies [64]. It also includes additional information about the primary studies such as title, authors, journal, and publication details, etc. [35]. We have extracted the data from the selected set of primary studies and represented that data in tabular and graphic forms in different ways. Moreover, the synthesis of the primary studies includes intervention, population, and context, sample sizes, outcomes, study quality. The SLRs are in IT and software engineering are mostly qualitative in character; this makes them suitable for descriptive synthesis [35].

Table 4, given below depicts the selection process of the selected primary studies from the specified online digital libraries followed by its graphical representation shown in Fig. 2. Tables 5–8 depict the data extraction details of this work and provide an accurate and comprehensive view of primary study data extraction. Figs. 3 and 4, also provide the graphical representation of the year and country-wise research contributions, respectively. This will help in a more systematic and accurate appraisal of our intent.

Table 4: Selection process for primary study

Database name	Total resulted generated	Removed on title search	Removed on abstract and keyword search	Removed on inclusion exclusion criteria	Removed on quality assessment test	Selected as primary study
IEEE Xplore	128	41	61	15	4	7
Science Direct	122	37	64	16	4	1
Springer Link	159	59	77	19	2	2
ACM Digital Library	182	72	84	22	2	2
Willey Online Library	26	15	9	2	0	0
PubMed	28	14	7	4	1	2
MDPI	14	5	5	3	0	1
Total	659	243	307	81	13 + 4 = 17	15 + 4 = 19

Note: +4 studies in column 6 and 7 of Table 4 have been added from Google Scholar and Snowballing.

Table 5: Primary studies with their main idea

Author	Main idea of the study	Ref. No.
Hurst et al.	This study aims to design a machine learning model that can help healthcare organizations to detect suspicious user behaviour against patients' health data and make the auditing process easy, fast and cost-effective in a proactive mode.	[61]
Sicuranza et al.	The goal of the study is to build machine learning models that can detect both known as well as unknown attacks in the healthcare environment. Also, the voting and the expert system are used to sort out the dissension among 3-diifferent classifier outputs and misuse anomaly detection modules respectively.	[63]

(Continued)

Table 5 (continued)

Author	Main idea of the study	Ref. No.
Boddy et al.	To identify outliers or suspicious access to EPRs, a human-in-the-loop ML approach with a density-based outlier detection model has been implemented through ML techniques. Ensemble averaging method is used to generate more efficient and effective results. Ensemble average anomaly scores are used to identify outliers and if the anomaly score value is greater than 2, it is considered as an outlier and taken up for further investigation.	[45]
McGlade et al.	This work is carried out to detect anomalies in EMR systems using ML techniques to ensure the confidentiality of protected health data at the application layer. Time series analysis is used to detect system availability threats for which the exponential moving average method has been practiced.	[46]
Boddy et al.	The main objective of the work is to identify malicious user access to protected healthcare data based on EPR audit logs and density based local outlier model.	[57]
Wesolowski et al.	A host-based intrusion analysis and identification system have been proposed based on user profiling by monitoring real time keystrokes with time dependencies of a user. For that ensemble, the ML algorithms have been used to build the intrusion detection classifiers in a dynamic healthcare environment.	[47]
Costante et al.	To design an integrated ML model based on signature and anomaly detection techniques for the prevention and detection of unauthorized data breaches.	[59]
Tchakoucht et al.	A host-based intrusion detection in an EHR system on the behaviour of the user profile. System log data is used to detect abnormalities whenever the user activity deviates from its profile.	[55]
Menon et al.	An unsupervised ML approach called collaborative filtering is practiced to predict illegitimate accesses to EHR systems. It enhances the previous research work [48] by considering identity involved in accesses as a parameter and latent feature model to develop a more fine grained model for suspicious access detection.	[48]
Ekina et al.	The Bayesian co-clustering approach is used to identify the frauds in dyadic data of healthcare among fraud care providers and customers to detect unusual memberships.	[49]
Zhang et al.	The work proposes a patient-flow-based anomaly detection (PFAD) model based on graph-based analysis that measures the degree of abnormality of a patient's record access based on deviation score.	[53]

(Continued)

Table 5 (continued)

Author	Main idea of the study	Ref. No.
Gupta et al.	This work aims to propose a Random Topic Access model that can control the shortcomings of the Random Access Object model to detect abnormal user behaviour based on topics while accessing health information systems.	[60]
Chen et al.	The authors of this work proposed a SNAD model to detect insider anomalous activities in collaborative information systems such as EHR systems. SNAD on unsupervised ML approach and its experimental results are evaluated on log data of an EHR system and wiki editing log data. Results show that SAND provides better results in comparison to the Spectral model and covers approximately 20% more area under the AUROC curve on an average.	[50]
Chen et al.	The main focus of this work is to analyze the predictive behaviour of users and use it to identify and detect suspicious behaviours concerning a patients' record access. For this objective, the study uses a global and local network relational approach to infer the interactions and relationships of HCO departments and users with health records.	[62]
Boxwala et al.	To evaluate the strength of statistical and ML approaches which are used in digital healthcare infrastructure to detect illegitimate access to EHRs and improve the confidentiality of healthcare protected data proactively.	[51]
Ziemniak	The machine learning J48 algorithm has been used to detect abnormal behaviours performed by healthcare application users while accessing healthcare data. The proposed model will help healthcare admins to detect false behaviour performed by a staff member to access health data and the same can be forwarded for further investigation.	[52]
Kim et al.	The main objective of this study is to provide an extension to [48] through a semi-automated approach to detect illegitimate access. For this, an integrated filtering approach using symbolic clustering and signature detection technique has been developed to identify suspicious activities in an EHR environment.	[54]
Chen et al.	Proposed an anomaly detection framework that on user-community behaviour to identify illegitimate behaviour if it finds community user deviates from usual behaviour of the group.	[58]
Asfaw et al.	This study aims to design an auto-detection model that can analyze the behaviour of current workflow with respect to the normal workflow of user-profiles to detect anomalous workflow in the healthcare system.	[56]

Table 6: Publication details of the primary studies

Conference preceding or journal	Impact factor/Cite score	WoS indexed	Scopus indexed	Quartile	Researchers affiliated country	Ref. No.
5th International Conference on Information Management (ICIM 2019)	✗	✗	✓	✗	England	[45]
Smart Health	✗	✗	✗	Q3	Northern Ireland	[46]
Applied Artificial Intelligence	1.172	✓	✓	Q3	Poland	[47]
Security Informatics	✗	✗	✗	✗	USA	[50]
JAMIA	4.112	✓	✓	Q1	USA	[51]
6th International Conference on IT Security Incident Management and IT Forensics	✗	✗	✗	✗	USA	[52]
ACM Transactions on Management Information Systems	4.2	✓	✓	Q1	USA	[53]
12th International Conference of Computer Systems and Applications (AICCSA)	✗	✗	✗	✗	Morocco, France	[55]
Chemical Engineering Transactions	1.3	✗	✓	Q3	Italy and USA	[49]
Fifth International Conference on Risks and Security of Internet and Systems (CRiSIS)	✗	✗	✗	✗	Ethiopia, Africa, Italy	[56]
IEEE Access	3.745	✓	✓	Q1	England	[57]
First ACM Conference on Data and Application Security and Privacy	✗	✗	✗	✗	USA	[58]

(Continued)

Table 6 (continued)

Conference preceding or journal	Impact factor/Cite score	WoS indexed	Scopus indexed	Quartile	Researchers affiliated country	Ref. No.
AMIA Annual Symposium Proceedings 2011	1.6	✗	✓	Q3	USA	[54]
2016 IEEE Security and Privacy Hops	✗	✗	✗	✗	Netherlands	[59]
Machine learning	2.672	✓	✓	Q1	USA	[48]
The Institute for High Performance Computing and Networking	✗	✗	✗	✗	Italy	[63]
AMIA Annual Symposium Proceedings 2012	1.6	✗	✓	Q3	USA	[62]
2013 IEEE International Conference on Intelligence and Security Informatics	✗	✗	✗	✗	USA	[60]
Future Internet	2.8	✓	✓	Q2	England, UAE	[61]

Table 7: Synthesis of primary studies

Intervention (ML techniques used to secure EHRs)	Population	Sample size	Outcome	Ref. No.
HILML and *K*-nearest neighbor algorithm	Identification of outliers or suspicious access in EPR systems.	100, 7727 Audit log records of Liverpool hospital- England.	Confidentiality and privacy of EPRs.	[45]
Naive Bayes, SVM, Nearest neighbor algorithms, Time series EMA method	Detection of anomalies and availability attacks in EPR systems.	2010 synthetic log records.	Confidentiality and availability of EMR systems.	[46]

(Continued)

Table 7 (continued)

Intervention (ML techniques used to secure EHRs)	Population	Sample size	Outcome	Ref. No.
Random forest; SVM; Bayesian Network; C4.5	Analysis and detection of keyboard based intrusions in HER systems	281 log files of user activities from 15 computers.	Ensemble classifier model that detects unauthorized access based on keystroke monitoring.	[47]
Collaborative filtering approach	Prediction of illegitimate access to EHRs	34.10 million Access records of EHR systems of two hospitals in Boston, US.	Ensuring privacy and confidentiality of EHR systems.	[48]
Bayesian Co-clustering	Identification of unusual membership amounts entities in healthcare to detect frauds.	Gibbs sampler of 10K iterations.	Restrict frauds in the healthcare sector.	[49]
Unsupervised ML approach	Detection of insider threats.	1,327,500 Access records of an EHR system of Vanderbilt University Medical Center.	Securing collaborative information systems from insider attacks.	[50]
Logistic regression and SVM	Role of statistical and ML methods in suspicious access identification.	1,291 labeled event records.	A model that will help healthcare data security officers to detect malicious access to health records.	[51]
J48 Decision tree algorithm	Atypical behaviour identification while accessing confidential health data.	65,000 instances of staff behaviours generated from McKesson Portal data.	An ML-based classifier that detects abnormal behaviours performed by healthcare application users (staff).	[52]

(Continued)

Table 7 (continued)

Intervention (ML techniques used to secure EHRs)	Population	Sample size	Outcome	Ref. No.
Graph-based analysis	Detection of deviated accesses while accessing patient records by clinicians and non-clinicians.	1.14 million Accesses with 16,561 patients are collected from Northwestern Memorial Hospital, US.	A graph-based model that helps to evaluate and detect user access deviations in an EHR system.	[53]
Symbolic clustering and signature detection techniques	Semi-automated identification of skeptical access to EHRs.	8.58 million EHRs accesses for control arm 25.50 million accesses for intervention arm collected from two operational institutes.	Improved [48] suspicious access detection model for EHR systems.	[54]
K-means algorithm	Identification of deviated behaviour of a system used to detect intrusions.	44 profiles of user-profiles for training and 16 additional profiles for testing	Intrusion detection model that helps to identify suspicious behaviour of a user.	[55]
Classification and association rule mining	Analysis and detection of anomalous workflow in the healthcare system.	20 noise-free sample records to design rules and generated data set of 300 association rules.	Anomaly detection model based on classification and association rule.	[56]
Density based local outlier-factor algorithm based on K-nearest neighbor	Proactive monitoring of EPR audit logs to identify suspicious accesses to EPR systems.	10,07,727 Audit log records of Liverpool hospital, England.	Detection model that can prevent patient records from suspicious accesses.	[57]
K-nearest neighbors and Singular value decomposition	Identification of insider threats in healthcare collaborative information systems.	1.5 million Patients' records of an EHR system of Vanderbilt University Medical Center.	An ML-based framework that can detect suspicious user behaviour against usual user-group patterns.	[58]

(Continued)

Table 7 (continued)

Intervention (ML techniques used to secure EHRs)	Population	Sample size	Outcome	Ref. No.
Signature and anomaly detection techniques	Protection of data from illegitimate disclosures.	A synthetic data set of more than 30K queries retrieved from GnuHealth.	A hybrid model to prevent and detect loss of confidential data.	[59]
Latent Dirichlet Allocation and *K*-nearest neighbor	Detection of suspicious user behaviour based on random access of topics.	Data set of 4.9 million audit logs of 7,932 users to 14,606 patients from Northwestern Memorial Hospital.	RTA detection model that is based on ML technique.	[60]
Local Outlier Factor and Density-Based Spatial Clustering of Applications with Noise techniques	How to bridge the gap of healthcare data breaches and complex time consuming auditing investigation.	1,007,727 audit log records with 90,385 different Ids of a UK hospital.	An intelligent model that can help security analysts to detect suspicious user-behaviours and make the auditing process effective.	[61]
Unsupervised relational learning based approach	To find patterns and relationships from authentic user behaviours at the departmental level in HCOs to detect suspicious patient record accesses.	3 months access logs of 1.7 million Patient records of Star Panel EMR system.	Global and local Network relational model that helps HOCs to identify anomalous user behaviours.	[62]
Association rule mining; Decision Tree; Neural Network; K-means algorithms	Proactive intrusion detection in smart health systems.	An Italian EHR system is used to create a data set.	An ML-based Hybrid intrusion detection system.	[63]

Table 8: Qualitative and quantitative results of primary studies

Experimental results	Tool used	Performance measuring scales (Evaluation parameters)	Ref. No.
Each of the three ensemble classifiers has 4 different methods and provides effective results with 98.08% accuracy. In single mode, Random forest provides better results followed by BN, C4.5, and SVM, respectively.	KNIME and WEKA.	Equal error rate and Accuracy.	[47]
The Human-in-the-loop ML approach with ensemble averaging provides effective results and improves the efficiency of unsupervised models. Results show that out of 1.07 million log files, only 145 are identified as suspicious that accounts for 0.014%.	NM*	NM*	[45]
Results of the study show that SVM provides better results with 98.94% accuracy at validation and 100% at testing as compared to KNN (95%) and Naïve Bayes (74%). Among 3, 5, 7, and 10, 7-fold cross validation shows better performance among all the models.	Python Pickle Library; Apache Kafka; Synthea.	Accuracy, Precision, Recall.	[46]
The collaborative filtering model provides effective results with a 0.9900 performance value under the ROC curve in comparison to linear regression (0.9642), logistic regression (0.9688), and support vector machine (0.9658) models in linear mode.	NM*	ROC curve, Precision recall curve, Root mean-square error, and Geometric mean.	[48]

(Continued)

Table 8 (continued)

Experimental results	Tool used	Performance measuring scales (Evaluation parameters)	Ref. No.
The Bayesian co-clustering approach can quantify imprecision in abnormal behaviours and improve decision-making for fraud detection. It has sufficient ability to analyze dyadic data connecting two entities.	NM*	NM*	[49]
The Specialized-NAD model provides effective results in small network environments that depict high similarity among users. Specialized-NAD IDF shows better performance with AUROC value 0.83 on healthcare data set as compared to Binary SNAD (0.79), Spectral IDF (0.69), and Binary Spectral (0.74).	NM*	Sensitivity, ROC curve, Specificity.	[50]
AUROC curve has been used for performance measure that calculates AUC 0.91 for logistic regression and 0.94 for SVM at baseline. SVM model provides better results with 79.3% detection of True positive (sensitivity) while LR detects only 75.8% of a true positive.	NM*	AUROC curve, sensitivity.	[51]
J48 algorithm can handle over-fitting and missing values. Results of the study also show that ad-hoc investigation provides better results as compared to path length, and outlier instance removal.	WEKA; SQL	NM*	[52]

(Continued)

Table 8 (continued)

Experimental results	Tool used	Performance measuring scales (Evaluation parameters)	Ref. No.
The simulated results of the study show that the FFAD model provides effective results under all four service types. However, highest in pediatrics service with 0.916 AUROC curve and lowest in hospital medicine service with 0.83. FFAD provides better efficiency in computations as compared to [47] and shows better results in Obstetrics and pediatrics service types.	R statistical software.	ROC curve, True positive rate, False positive rate.	[53]
Experiments have been conducted under two different scenarios these are the Control arm (CA) and the Intervention arm (IA). MPA1 based logistic regression is applied to CA and MPA2 filtering approach to IA under two different population sets and validated on an independent set of 78 labeled events. The area under AUC for the control arm is 0.960 and 0.990 for the intervention arm. IA shows a higher false-negative rate as compared to CA.	NM*	AUROC curve, False negative rate.	[54]
K-means algorithm has sufficient ability to detect behavioral intrusions, but on small data sets only, and is also lagging in accuracy when user behaviour shows frequent variation. The proposed model provides 91% and 75% accuracy on learned and unlearned groups, respectively.	NM*	True positive rate, False negative rate, False positive rate.	[55]

(Continued)

Table 8 (continued)

Experimental results	Tool used	Performance measuring scales (Evaluation parameters)	Ref. No.
Generated results depict that association rule mining can detect insider intrusions in healthcare systems if data of user activities are preserved. Also, support = 15%, 25% and confidence = 50%, 100% has been considered to generate rules.	J2ME; Java Servlets; MySQL Server.	Support, Confidence.	[56]
Study Results show that out of 1.07 million log files, only 144 are identified as suspicious that accounts for 0.014% of the total. ML techniques have a significant ability to detect abnormal behaviours of users while accessing the EPR data.	NM*	NM*	[57]
Community-based anomaly detection (CAD) model works more efficiently in complex environments as compared to KNN, PCA, and HVU models. In various experiments, the simulated user rate varies from 0.5% to 5%, and at this rate, CAD has got the highest AUC score as compared to the other 3 models.	NM*	ROC curve, True positive rate, True negative rate	[58]

(Continued)

Table 8 (continued)

Experimental results	Tool used	Performance measuring scales (Evaluation parameters)	Ref. No.
Experimental results depict that the proposed framework has high efficiency to prevent and detect data breaches as compared to existing models. 3:7 ratio of data is used for training and validation purposes. On small data set, the approximate reduction of the prevention module was 35%, and on a large set, it was 10% to produce anomalous alerts. This shows that with timely detection and prevention, the module achieves accuracy.	Rapid Miner	True positive rate, False positive rate	[59]
The proposed model can detect 3 types of anomalous users. Masquerade (Direct) user where $\alpha < 1$, Random user where $\alpha = 1$, and Indirect user where $\alpha > 1$. In experimental results area under ROC varies from 0.7 to 0.99 for different roles at different α settings 0.01, 0.1, 1, 10, 100.	NM*	AUROC curve.	[60]
DBSCAN detects 102 and LOF identifies 385 anomalies out of 90,385 different Ids under the Human-in-the-loop approach. From the results, the recommended technique is LOF as compared to DBSCAN.	R software package	NM*	[61]

(Continued)

Table 8 (continued)

Experimental results	Tool used	Performance measuring scales (Evaluation parameters)	Ref. No.
Results of the work show that the global network is partially stable for exhibiting the relations among HOC departments and shows a small certainty change. They also depict a small average variation in the reciprocity of 0.2770; implying that departmental relations are unbalanced. Local network show a small variation in both the network scores and reciprocity. Extracted relations from the networks are used to detect suspicious record accesses.	NM*	Certainty and reciprocity	[62]
Experimental results depict that misused detection system integrated with voting-based anomaly detection model provide highly effective results. The sensitivity score of DT, ANN, and K-means on both the normal and attack events are (1, 0.792), (0.983, 0.683), and (0.987, 0.815), respectively. And overall accuracy of model after voting is (0.999, 0.878).	KNIME and WEKA	Support, Confidence, Specificity, sensitivity, precision, recall.	[63]

Note: NM* (Not mentioned in the original study).

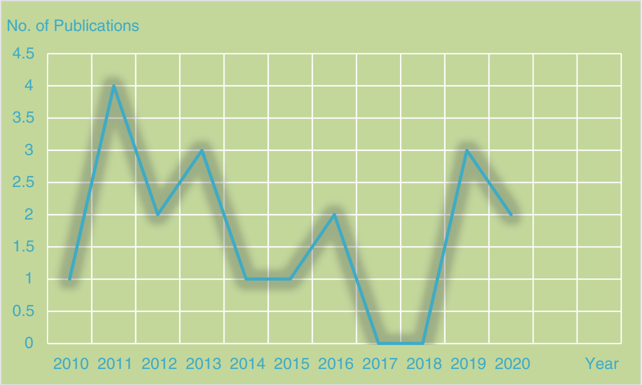

Figure 3: Year-wise research contribution

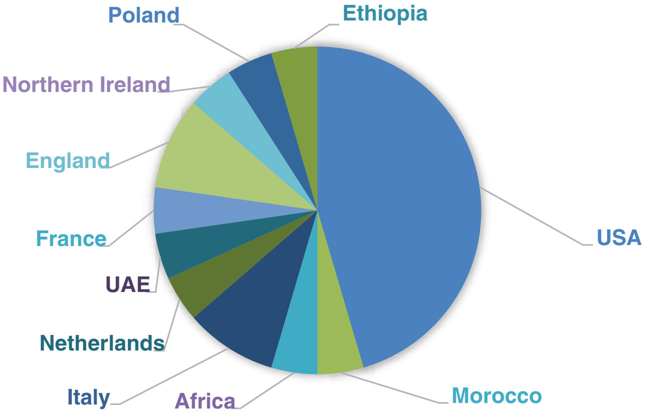

Figure 4: Country wise research contribution by researchers

3.3 Reporting Phase

Reporting phase is the final step of the systematic literature review process [65]. It defines how to write down the results and disseminating these results among the interested parties [35,65]. It ensures an effective way to communicate the results of SLR. In this phase, we will present the results of this study and provide answers to the research questions that are mentioned above in the Planning Phase.

3.3.1 Results and Findings

The ubiquitous technologies such as Artificial Intelligence, Cloud Computing, machine learning, the Internet of Things, and other web technologies in the present scenario have a significant role in every field of life. These technologies have made human life smarter by providing cost-effective and easily accessible services [1,2]. These technologies have also had a dramatic impact on the healthcare industry, making the healthcare services prompt and extending their outreach [13,39,66]. But the digital revolution is also the harbinger of serious issues and challenges. Security and privacy of healthcare data (EHRs) is one among them and has become a pressing issue for both the healthcare service providers and patients. This is because the transformation of conventional paper based system to smart health information systems makes the healthcare data more vulnerable to cyber-attacks [3,44]. Various approaches and techniques have been practiced by the researchers and experts to address the security and privacy issues and challenges faced by digital healthcare industry [7,11,44].

Addressing the identified research gap, this study aims to attempt an SLR that collated the stated research endeavours on ML approaches. In this context, we followed a systematic process to investigate the research studies thoroughly and based upon it, found 19 research studies as primary studies that satisfied the eligibility criteria framed in this study. Here we will discuss the results extracted from these primary studies with the above-specified research questions (RQ1-RQ4).

RQ1. Does the research study use a machine learning approach to ensure confidentiality of sensitive healthcare data in an EHR environment?

The final selection of the primary studies based on the formal process of the systematic literature review is outlined and summarized in Table 5. It provides sufficient details regarding the primary studies that include the authors' names and the publishing year of the study, the main ideal of the study that provides the purpose of that study, and references of these studies. Further, Table 7 also provides the answer to RQ1 as it separately presents the machine learning techniques used in these studies under column intervention. It also provides population (what is the problem), sample size, and outcome of the study to provide a satisfactory answer to QR1 defined in the Planning phase above.

RQ2. Does the study discuss healthcare data confidentiality issues?

The primary concern of this study was to investigate the original research studies that used ML to ensure healthcare data confidentiality. For that, RQ2 was framed to assess that the selected studies discussed healthcare data confidentiality issues. Data confidentiality ensures that protected data is accessible to only those who have got privileges to access it. Healthcare data confidentiality and privacy nowadays have become a compelling concern for both the service providers and patients [3]. The implementation of technology in the healthcare domain increases the probability of exposing sensitive health data [48]. It is one of the top three sectors that face the highest number of breached incidents annually [61]. Boddy et al. discuss healthcare data confidentiality and specify that both the insider as well as outsider threats violate healthcare data confidentiality and results in patients' loss of trust in the healthcare service providers [45,57]. Statistics regarding healthcare data confidentiality [46] specify that there is a 63% increment in cyber-attacks from 2015 to 2016 against hospitals of the USA. Moreover, health data records have ten to twenty times higher selling value than credit card data in the online market. Sensitive data such as EHRs are managed by collaborative information systems where their misuse can lead to detrimental issues [50,58]. It is reported that in the year 2010, out of the total annual healthcare expenditure, more than $60,000 million had been lost in healthcare frauds in the USA [49]. Healthcare data breaches led to reputation and organizational competitive advantage loss [59], patients' trust loss, and its cost is estimated to be higher as compared to other industries [51].

The inherent complex and dynamic nature of the healthcare industry make EHR systems potentially vulnerable to insider attacks that adversely affect the confidentiality of data [53]. Healthcare data is exploited by intruders to compromise the privacy of patients [63]. The pervasive character of EHR systems imposes more security risks to health data [56]. The open-access nature of the smart healthcare environment for its users makes it a more serious issue concerning the privacy of data [52]. Insider threats are prevalent in hospital management systems [60,62]. It has been found that 60 to 70 percent of the attacks come from insiders (legitimate users) and one of its main reasons is the role based access security mechanisms [54]. These facts, figures, and concepts depict that selected primary studies justify the RQ2 and provide details about healthcare data confidentiality.

RQ3. Does the study include experimental results with a theoretical framework?

Theoretical framework with experimental results have enhanced the study quality and provided a practical view of the research work concerning the proposed research problem. Experimental results of a research study specify the degree of practicability in real-world environment and help the readers to understand the theoretical framework discussed in a concrete perspective and to find insights from the drawn results for further investigation to make future research efforts. Table 8, given below provides information to address the research question RQ3 satisfactorily from different primary studies.

From the analysis of the selected primary study set, we have conclusively determined that the ML based procedures used by the researchers to identify suspicious user accesses against the EHR systems have achieved desirable comparative benefits in healthcare domain. Before the implementation of ML in healthcare cybersecurity, the traditional procedures used for data security were the cryptographic approaches to encrypt and decrypt the patients' data for secure communication of the data among all the stakeholders at all levels. However, cryptographic approaches don't have the capability to identify suspicious user accesses against the EHR systems. Further, detecting any type of illegitimate user access at any given time needs manual investigation by the domain experts. This takes a lot of time and other resources because the system log entries are very huge in number and their manual human investigation is very complex and error prone. To overcome this burden and make digital healthcare environment more secure, the researchers have practiced ML approaches. This approach provides effective results in case of investigating large amount of system log data in relatively less time. Further, the approach enhances the manual identification of suspicions user accesses and provides very effective results on real time basis. Implementing of ML-based models in digital health environment is also more efficacious because the models examine all new log entries of users periodically and alert the system admins whenever a suspicious user access is detected. This helps the system's admin to make a real time analysis of the processing system and take urgent steps against the suspicious user accesses. It protects the healthcare service providers and patients from catastrophic data breaches which could otherwise lead to loss of the organization's brand image, and disclosure of patients' sensitive data. Thus, compromising the patients' privacy and also jeopardizing their treatment as the data can be tampered by the intruders. Hence, the, integration of ML-based models in healthcare domain is comparatively very effective, productive, and a reliable approach.

The comparative statistical results of the commonly proposed ML-based models for privacy and confidentiality preservation of EHRs by different researchers are as: In one experiment SVM provides better results with 98.94% accuracy as compared to KNN 95% and Naïve Bayes 74%. In another experimental work, the Collaborative filtering model provides effective results with 99% accuracy in comparison to linear regression with 96.42%, logistic regression with 96.88%, and support vector machine with 96.58% accuracy in linear mode. Specialized-NAD IDF shows better performance 83% as compared to Binary SNAD 79%, Spectral IDF 69%, and Binary Spectral 74% in a separate experimental work. Random forest gives 98.08% accuracy in another separate experiment. One more experimental work reveals K-means model provides 91% and 75% accuracy on learned and unlearned groups, respectively. The sensitivity score of Decision Tree, and K-mean on both the normal and attack events are 100%, 79.2%, and 98.7%, 81.5%, respectively. And overall accuracy of the model after voting is 99.9%, and 0.87.8% in another work. The variation in results generated from different experimental works is because of the variation in parameters for selection of dataset, variation in training data, stochastic learning algorithms, stochastic evaluation procedures, and differences in the experimental platforms.

RQ4. Does the study use real or synthetic data of any healthcare organization to train and test the model?

In the given scenario, data is produced in big volumes by different data industries across the world. Analytics of this data provides immense insights for the industries in this domain, mainly for those in the healthcare sector [67]. ML models need an adequate amount of this data to make proactive predictions, categorizations, and estimations [68]. There is a direct impact of input data quality on the quality of training an ML model. Identification of relevant data with the characteristic of completeness and accuracy provides a strong base for a high quality ML model [69]. But because of data privacy policies, it is not always possible to find real data for research experiments [46]. Therefore the researchers use different tools such as Synthea [46] and CA technologies Datamaker [70] to generate synthetic data sets to train and validate ML models [71]. Similarly, for the present SLR, the researchers have selected either the data sets generated from the real healthcare environment as real data sets or produced synthetic data sets to conduct the experiments. Table 7 of this study under the column "Sample size" provides sufficient information regarding the data sets used by the researchers of these studies to enlist the answers for research question RQ4. Further, Table 9 demonstrates the most common features (attributes) considered by researchers in their studies to design data sets and conduct experiments. This table only includes the attributes that are practiced by most of the researchers in their experiments. However, there are few studies [45,47] that do not mention the data set attributes used in their original studies. [47] practiced ensemble classification of keystroke dynamics approach to secure unauthorized accesses against EHR systems. [45] practiced the Bayesian co-clustering method to detect frauds in electronic healthcare systems, and involved only the user id attribute among the commonly used attributes to carry out the experiment.

4 Discussion and Future Work

The sole aim of this work was to conduct a systematic literature review on those research studies that practice machine learning approaches to ensure confidentiality of EHRs and EHR systems against intrusions. From the systematic analysis of various research studies, we found that the healthcare sector is highly susceptible to intrusions. As it is evident from several surveys in this context, the industry has been continuously targeted by cyber intrusions ever since conventional healthcare systems gave way to smart digital systems. Various security mechanisms have been practiced and adopted by the researchers and organizations to secure sensitive health data. Systematic review studies to analyze and summarize the collective results of the existing research endeavors in the stated context have been done by several researchers. But analysis of the academic research work shows that ensuring healthcare data confidentiality through ML approaches has not been the prime focus of these researchers as compared to other security mechanisms.

Table 9: Commonly used attributes

Attributes												Ref. No. of the Study
Date and time	User action/ Operation	Device ID	Login flow	User ID	No. of logins	Patient ID	Location	Routine/ Encounter	Duration	Access fails	Break glass	
■		■		■		■		■	■			[45]
								■			■	[46]
				■								[47]
				■		■						[48]
												[49]
■	■			■		■						[50]
■				■		■	■		■			[51]
	■	■		■	■	■			■		■	[52]
■	■			■		■	■					[53]
■				■		■	■		■			[54]
			■		■					■		[55]
■						■	■		■			[56]
■		■		■		■		■	■			[57]
■	■			■		■						[58]
■				■		■						[59]
■	■	■		■		■	■					[60]
■		■		■		■		■	■			[61]
■				■		■						[62]
■	■			■		■	■					[63]

Therefore, the authors of the present study undertook a systematic review on the specified domain. The foremost findings of these studies are represented in Table 10. Analysis of the primary studies depicts that most of the researchers have focused on the insider attacks to detect privileged abuse. Misuse and anomaly detection remain a special concern of these studies to detect known and unknown attacks. Signature detection techniques are used to detect known attacks but with static nature. Moreover these rule based techniques are incompetent in addressing the unknown attacks. On the other side, anomaly detection approaches are competent in addressing this issue but with a sufficient false positive rate. All these studies have shortcomings at their levels, which are discussed in these studies, but here we will discuss some common and important limitations that will help the researchers in future pursuits. These are as follows:

- On the basis of the normal user-behaviour profile analysis, patterns (rules) have been generated and these patterns are then used to detect abnormal user behaviours. These rules are environment specific, lacking in generalization, and inefficient to justify the completeness and accuracy of the rule.

- Anomaly detection techniques need human intervention (expert decision) to finalize that the detected anomaly is a threat or not.
- Unavailability of sufficient labeled data for supervised ML techniques is also a challenge and using Boolean (binary) logic to predict the class label for an event or access in an HER system is always doubtful. This is because binary logic does not address imprecise information and uncertainties about a decision.
- High false positive rate in anomaly detect models is also a matter of concern and needs future research for improvement.
- Deviation and irregularity scores generated by models are used to identify suspicious user accesses. But these scores still need improvement for producing accurate results and for enhancing the system's performance.
- There are very few ML techniques that are practiced in our specified domain mostly KNN, SVM, and Naïve Bayes with less comparative analysis. So, other new and advanced ML techniques can be utilized for more effective results.
- Synthetic data sets with complete EHR environmental scenarios can also help to improve ML model's accuracy and performance up to a satisfactory level in the EHR environment. But there are too few studies that focus on an incomplete EHR environmental scenario.
- Integration of misuse and anomaly detection techniques to build a complete auto-detection system with feedback and maximum suspicious scenario consideration is also a challenging issue in a dynamic healthcare environment.
- Comparative analyses of models and performance measure have not been discussed sufficiently and need good attention of the researchers to improve it with respect to healthcare environment.
- As a limitation of this work, deep learning as an advanced and complex sub-domain of ML have not been included in this research endeavor and will be the premise for our future research work as a separate study for systematic review.
- All the above discussed issues need the immediate focus of researchers to improve the confidentiality of Electronic Health Records in the dynamic healthcare environment.

Table 10: Foremost findings of primary studies

Classification	Commonly practiced	Mostly practiced
Techniques	K-nearest neighbor; Support Vector Machine; Naive Bayes or Bayes Network.	K-nearest neighbor.
Performance measuring scales	Area under ROC curve; True positive rate; False Positive rate; Precision; Recall; Sensitivity; Specificity.	ROC curve.
Attributes	Patient Id; User Id; Date and Time; Duration; User action; Location.	Patient Id; User Id; Date and Time.

(Continued)

Table 10 (continued)

Classification	Commonly practiced	Mostly practiced
Main focus	User behaviour analysis; work flow analysis; misuse detection; anomaly detection.	User behaviour analysis and anomaly detection.
Implementation tools	WEKA; KNIME; R statistical software package; Python Pickle Library; Rapid Miner.	WEKA and KNIME.
Source of data	EHR system log data; EMR system log data; Synthetic data.	EHR system log data.

5 Validation

A systematic literature review provides a complete as well as a methodical procedure to carry out an unbiased and error-free literature review [71,72]. Conducting an error-free survey and minimizing the biasness up to the highest level was also the core priority of this study. However, there are some threats that do exist and while conducting a systematic review, they may harm the validity of the study [73]. These include a selection of limited digital databases and search engines, imperfection in the retrieval of research papers, inaccuracy in the framing of inclusion-exclusion parameters, biasness in setting QA questions. To justify these threats, the following perusal of the types of validities has been enlisted:

Internal Validity: It defines the strength of evaluation to perform the study and ensures efficiency in the process of literature survey [72]. For this, we developed a strong searching mechanism discussed comprehensively in "Conducting phase" of this study and presented in Fig. 2. It defines five search operations from S1 to S5 to make the searching process more efficient. Experts were also consulted and search engines were used to identify the most relevant keywords based on the specified search terms for better results. Still framing inclusion-exclusion parameters, designing quality assessment questions, and assigning QA scores to primary studies are solely based on the knowledge, experience and decisions of authors of the work that may have led to some level of biasness.

External Validity: It ensures the generalization of the results, or we can say that it defines the applicability of the results. Here, our generated results are dependent on the search engines of specified digital libraries [73]. Every search engine and the digital repository have some kind of limitations in generating accurate results. Moreover, the search strings framed by us to run on different digital databases by using hit and trial means may have added some biasness. Despite this, we have applied different kinds of filters that are provided by the respective databases to enhance the efficiency of the results.

6 Conclusions

This research work depicts that ML has got significant importance in the healthcare sector. ML-based models can help the security experts to implement intrusion detection systems in the EHR environment and identify suspicious accesses against EHRs in a proactive mode. There are very few research endeavours that adopt ML approaches to address EHR confidentiality and privacy issues. These studies address some of the issues but not all. This research gap became the main objective of our study with the intent to highlight the instrumental role of ML approaches which would be highly effective in overcoming the rapidly growing rate of healthcare data breaches. The identified research areas depict the importance and need of this research endeavour. Furthermore, the cooperation of healthcare service providers and security analysts is of utmost importance for understanding the dynamic environment of the healthcare industry and different scenarios of normal user profiles. This understanding will provide a generalized view of the same and will be helpful in identifying all abnormal user behaviours in a better way. In the existing research studies, the commonly used technique is K-nearest neighbor and SVM, and the implementation tool is WEKA and KNIME. Mostly, the EHR system log data is practiced to conduct experiments. The commonly used attributes in these experiments are Patient Id, User Id, Date and Time, Duration, User action, and Location. Most of the studies do not provide comparative and performance results of the ML models, nor do they provide a complete dynamic view of normal as well as abnormal user profiles. The systematic investigation carried out for this study shows that among the specified online digital libraries, PubMed and IEEE Explore have better ability to generate the relevant results.

Funding Statement: This research was supported by Taif University Researchers Supporting Project under the Grant No. TURSP-2020/211, Taif University, Taif, Saudi Arabia.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Gurbaxani, V., Whang, S. (1991). The impact of information systems on organizations and markets. *Communications of the ACM, 34(1),* 59–73. DOI 10.1145/99977.99990.
2. Henderson, J. C., Venkatraman, H. (1999). Strategic alignment: Leveraging information technology for transforming organizations. *IBM Systems Journal, 38(2,3),* 472–484. DOI 10.1147/SJ.1999.5387096.
3. Seh, A. H., Zarour, M., Alenezi, M., Sarkar, A. K., Agrawal, A. et al. (2020). Healthcare data breaches: Insights and implications. *Healthcare, 8(2),* 133. DOI 10.3390/healthcare8020133.
4. Bhatia, T., Verma, A. K., Sharma, G. (2019). Towards a secure incremental proxy re-encryption for e-healthcare data sharing in mobile cloud computing. *Concurrency and Computation: Practice and Experience, 32(5),* 668. DOI 10.1002/cpe.5520.
5. Subahi, A. F. (2019). Edge-based IoT medical record system: Requirements, recommendations and conceptual design. *IEEE Access, 7,* 94150–94159. DOI 10.1109/ACCESS.2019.2927958.
6. Shahzad, F., Iqbal, W., Bokhari, F. S. (2015). On the use of CryptDB for securing Electronic Health data in the cloud: A performance study. *17th International Conference on E-Health Networking, Application & Services,* pp. 120–125. Boston, USA.
7. Shrestha, N. M., Alsadoon, A., Prasad, P. W. C., Hourany, L., Elchouemi, A. (2016). Enhanced e-health framework for security and privacy in healthcare system. *Sixth International Conference on Digital Information Processing and Communications,* pp. 75–79. Beirut, Lebanon.
8. Vinaykumar, S., Zhang, C., Shahriar, H. (2019). Security and privacy of electronic medical records. *SAIS 2019 Proceedings,* pp. 1–6. Georgia, USA: St. Simon's, Island.
9. Chernyshev, M., Zeadally, S., Baig, Z. (2019). Healthcare data breaches: Implications for digital forensic readiness. *Journal of Medical Systems, 43(1),* 50. DOI 10.1007/s10916-018-1123-2.
10. Rathee, A. (2020). Data breaches in healthcare: A case study. *Cybernomics, 2(2),* 25–29.
11. Chenthara, S., Ahmed, K., Wang, H., Whittaker, F. (2019). Security and privacy-preserving challenges of e-health solutions in cloud computing. *IEEE Access, 7,* 74361–74382. DOI 10.1109/ACCESS.2019.2919982.

12. Prathima, S., Priya, C. (2020). Privacy preserving and security management in cloud-based electronic health records—A survey. In: *Intelligent computing and innovation on data science,* vol. 118*,* pp. 21–29. Springer Nature Singapore Pte, Ltd.
13. Yeng, P., Nweke, L. O., Woldaregay, A. Z., Yang, B., Snekkenes, E. A. (2020). Data-Driven and Artificial Intelligence (AI) approach for modelling and analyzing healthcare security practice: A systematic review. *Proceedings of SAI Intelligent Systems Conference,* pp. 1–18. London, UK.
14. Seh, A. H., Chaurasia, P. K. (2019). A review on heart disease prediction using machine learning techniques. *International Journal of Management, IT and Engineering, 9(4),* 208–224.
15. Alpaydin, E. (2020). *Introduction to machine learning*. USA: MIT press.
16. Mitchell, T. M. (1997). *Machine learning*. New York: McGraw-Hill.
17. Apruzzese, G., Colajanni, M., Ferretti, L., Guido, A., Marchetti, M. (2018). On the effectiveness of machine and deep learning for cyber security. *10th International Conference on Cyber Conflict,* pp. 371–390. Tallinn, Estonia.
18. Wu, J., Yin, L., Guo, Y. (2012). Cyber-attacks prediction model based on Bayesian network. *IEEE 18th International Conference on Parallel and Distributed Systems,* pp. 730–731. Singapore.
19. Kumar, M., Hanumanthappa, M., Kumar, T. S. (2012). Intrusion Detection System using decision tree algorithm. *IEEE 14th International Conference on Communication Technology,* pp. 629–634. Chengdu, China.
20. Sahu, S., Mehtre, B. M. (2015). Network intrusion detection system using J48 Decision Tree. *International Conference on Advances in Computing, Communications and Informatics,* pp. 2023–2026. Kochi, India.
21. D, A., Kumar, A. V. K., Visu, P. (2019). Malware traffic classification using principal component analysis and artificial neural network for extreme surveillance. *Computer Communications, 147(2),* 50–57. DOI 10.1016/j.comcom.2019.08.003.
22. Kosek, A. M. (2016). Contextual anomaly detection for cyber-physical security in smart grids based on an artificial neural network model. *Joint Workshop on Cyber-Physical Security and Resilience in Smart Grids,* pp. 1–6. Vienna, Austria.
23. Adebayo, O. S., Abdul Aziz, N. (2019). Improved malware detection model with apriori association rule and particle swarm optimization. *Security and Communication Networks, 2019(6),* 1–13. DOI 10.1155/2019/2850932.
24. Douzi, S., Benchaji, I., ElOuahidi, B. (2018). Hybrid approach for intrusion detection using fuzzy association rules. *2nd Cyber Security in Networking Conference,* pp. 1–3. Paris, France.
25. Ghanem, K., Aparicio-Navarro, F. J., Kyriakopoulos, K. G., Lambotharan, S., Chambers, J. A. (2017). Support vector machine for network intrusion and cyber-attack detection. *Sensor Signal Processing for Defence Conference,* pp. 1–5. London, UK.
26. Anton, S. D. D., Sinha, S., Schotten, H. D. (2019). Anomaly-based intrusion detection in industrial data with SVM and random forests. *International Conference on Software, Telecommunications and Computer Networks,* pp. 1–6. Split, Croatia.
27. Evangelou, M., Adams, N. M. (2020). An anomaly detection framework for cyber-security data. *Computers & Security, 97(10),* 101941. DOI 10.1016/j.cose.2020.101941.
28. Rao, B. B., Swathi, K. (2017). Fast kNN classifiers for network intrusion detection system. *Indian Journal of Science and Technology, 10(14),* 1–10. DOI 10.17485/ijst/2017/v10i14/93690.
29. Senthilnayaki, B., Venkatalakshmi, K., Kannan, A. (2019). Intrusion detection system using fuzzy rough set feature selection and modified KNN classifier. *International Arab Journal of Information Technology, 16(4),* 746–753.
30. Tan, X., Su, S., Huang, Z., Guo, X., Zuo, Z. et al. (2019). Wireless sensor networks intrusion detection based on SMOTE and the random forest algorithm. *Sensors, 19(1),* 203. DOI 10.3390/s19010203.
31. Masarat, S., Sharifian, S., Taheri, H. (2016). Modified parallel random forest for intrusion detection systems. *Journal of Supercomputing, 72(6),* 2235–2258. DOI 10.1007/s11227-016-1727-6.
32. Siddaway, D. A. (2014). What is a systematic literature review and how do I do one. *University of Stirling, 1(1),* 1–13.

33. Kofod, P. A. (2012). How to do a Structured Literature Review in computer science. https://www.researchgate.net/publication/265158913_How_to_do_a_Structured_Literature_Review_in_computer_science.
34. Shah, R., Chircu, A. (2018). IoT and AI in Healthcare: A systematic literature review. *Issues in Information Systems, 19(3),* pp. 33–41. DOI 10.48009/3_iis_2018_33-41.
35. Kitchenham, B., Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering. http://citeseerx.ist.psu.edu/viewdoc/summary?doi= 10.1.1.117.471.
36. Newaz, A. I., Sikder, A. K., Rahman, M. A., Uluagac, A. S. (2020). A Survey on Security and Privacy Issues in Modern Healthcare Systems: Attacks and Defenses. http://arxiv.org/abs/2005.07359
37. Qayyum, A., Qadir, J., Bilal, M., Al-Fuqaha, A. (2020). Secure and robust machine learning for healthcare: A survey. https://arxiv.org/abs/2001.08103.
38. da-Costa, K. A. P., Papa, J. P., Lisboa, C. O., Munoz, R., de Albuquerque, V. H. C. (2019). Internet of Things: A survey on machine learning-based intrusion detection approaches. *Computer Networks, 151(22),* 147–157. DOI 10.1016/j.comnet.2019.01.023.
39. Ghosal, P., Das, D., Das, I. (2018). Extensive survey on cloud-based IoT-healthcare and security using machine learning. *Fourth International Conference on Research in Computational Intelligence and Communication Networks,* pp. 1–5. Kolkata, India.
40. Kruse, C. S., Frederick, B., Jacobson, T., Monticone, D. K. (2017). Cybersecurity in healthcare: A systematic review of modern threats and trends. *Technology and Health Care, 25(1),* 1–10. DOI 10.3233/THC-161263.
41. Luna, R., Rhine, E., Myhra, M., Sullivan, R., Kruse, C. S. (2016). Cyber threats to health information systems: A systematic review. *Technology and Health Care, 24(1),* 1–9. DOI 10.3233/THC-151102.
42. Buczak, A. L., Guven, E. (2016). A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications Surveys & Tutorials, 18(2),* 1153–1176. DOI 10.1109/COMST.2015.2494502.
43. Rahim, F., Ismail, Z., Samy, G. (2014). Privacy challenges in electronic medical records: A systematic review. *Proceedings of the Knowledge Management International Conference,* pp. 12–15. Malaysia.
44. Fernández-Alemán, J. L., Señor, I. C., Lozoya, P. Á. O., Toval, A. (2013). Security and privacy in electronic health records: A systematic literature review. *Journal of Biomedical Informatics, 46(3),* 541–562. DOI 10.1016/j.jbi.2012.12.003.
45. Boddy, A., Hurst, W., Mackay, M., El Rhalibi, A. (2019). A hybrid density-based outlier detection model for privacy in electronic patient record system. *5th International Conference on Information Management,* pp. 92–96. Cambridge, UK.
46. McGlade, D., Scott-Hayward, S. (2019). ML-based cyber incident detection for Electronic Medical Record (EMR) systems. *Smart Health, 12(2),* 3–23. DOI 10.1016/j.smhl.2018.05.001.
47. Wesołowski, T. E., Porwik, P., Doroz, R. (2016). Electronic health record security based on ensemble classification of keystroke dynamics. *Applied Artificial Intelligence, 30(6),* 521–540. DOI 10.1080/08839514.2016.1193715.
48. Menon, A. K., Jiang, X., Kim, J., Vaidya, J., Ohno-Machado, L. (2014). Detecting inappropriate access to electronic health records using collaborative filtering. *Machine Learning, 95(1),* 87–101. DOI 10.1007/s10994-013-5376-1.
49. Ekina, T., Leva, F., Ruggeri, F., Soyer, R. (2013). Application of bayesian methods in detection of healthcare fraud. *Chemical Engineering Transaction, 33,* pp. 151–156. DOI 10.3303/CET1333026.
50. Chen, Y., Nyemba, S., Zhang, W., Malin, B. (2012). Specializing network analysis to detect anomalous insider actions. *Security Informatics, 1(1),* 4. DOI 10.1186/2190-8532-1-5.
51. Boxwala, A. A., Kim, J., Grillo, J. M., Ohno-Machado, L. (2011). Using statistical and machine learning to help institutions detect suspicious access to electronic health records. *Journal of the American Medical Informatics Association, 18(4),* 498–505. DOI 10.1136/amiajnl-2011-000217.
52. Ziemniak, T. (2011). Use of machine learning classification techniques to detect atypical behavior in medical applications. *Sixth International Conference on IT Security Incident Management and IT Forensics,* pp. 149–162. Stuttgart, Germany.

53. Zhang, H., Mehotra, S., Liebovitz, D., Gunter, C. A., Malin, B. (2013). Mining deviations from patient care pathways via electronic medical record system audits. *ACM Transactions on Management Information Systems, 4(4),* 1–20. DOI 10.1145/2544102.
54. Kim, J., Grillo, J. M., Boxwala, A. A., Jiang, X., Mandelbaum, R. B. et al. (2011). Anomaly and signature filtering improve classifier performance for detection of suspicious access To EHRs. *AMIA Annual Symposium Proceedings,* vol. 2011, pp. 723–731. LA, USA.
55. Tchakoucht, T. A., Ezziyyani, M., Jbilou, M., Salaun, M. (2015). Behavioral approach for intrusion detection. *IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA),* pp. 1–5. Marrakech, Morocco.
56. Asfaw, B., Bekele, D., Eshete, B., Villafiorita, A., Weldemariam, K. (2010). Host-based anomaly detection for pervasive medical systems. *Fifth International Conference on Risks and Security of Internet and Systems,* pp. 1–8. Montreal, Canada.
57. Boddy, A. J., Hurst, W., Mackay, M., El Rhalibi, A. (2019). Density-based outlier detection for safeguarding electronic patient record systems. *IEEE Access, 7,* 40285–40294. DOI 10.1109/ACCESS.2019.2906503.
58. Chen, Y., Malin, B. (2011). Detection of anomalous insiders in collaborative environments via relational analysis of access logs. *Proceedings of the First ACM Conference on Data and Application Security and Privacy,* pp. 63–74. San Antonio, USA.
59. Costante, E., Fauri, D., Etalle, S., Den Hartog, J., Zannone, N. (2016). A hybrid framework for data loss prevention and detection. *IEEE Security and Privacy Workshops,* pp. 324–333. San Jose, USA.
60. Gupta, S., Hanson, C., Gunter, C. A., Frank, M., Liebovitz, D. et al. (2013). Modeling and detecting anomalous topic access. *IEEE International Conference on Intelligence and Security Informatics,* pp. 100–105. Seattle, WA, USA.
61. Hurst, W., Boddy, A., Merabti, M., Shone, N. (2020). Patient privacy violation detection in healthcare critical infrastructures: An investigation using density-based benchmarking. *Future Internet, 12(6),* 100. DOI 10.3390/fi12060100.
62. Chen, Y., Nyemba, S., Malin, B. (2012). Auditing medical records accesses via healthcare interaction networks. *AMIA Annual Symposium Proceedings*, vol. 2012, pp. 93–102. NY, USA.
63. Sicuranza, M., Paragliola, G. (2020). Ensuring electronic health record cyber-security through an hybrid intrusion detection system. https://intranet.icar.cnr.it/wp-content/uploads/2020/05/RT-ICAR-NA-2020-01.pdf.
64. Taylor, P. J., Dargahi, T., Dehghantanha, A., Parizi, R. M., Choo, K. K. R. (2020). A systematic literature review of blockchain cyber security. *Digital Communications and Networks, 6(2),* 147–156. DOI 10.1016/j.dcan.2019.01.005.
65. Andročec, D., Novak, M., Oreški, D. (2018). Using semantic web for Internet of Things interoperability: A systematic review. *International Journal on Semantic Web and Information Systems, 14(4),* 147–171. DOI 10.4018/IJSWIS.
66. Gagnon, M. P., Desmartis, M., Labrecque, M., Car, J., Pagliari, C. et al. (2012). Systematic review of factors influencing the adoption of information and communication technologies by healthcare professionals. *Journal of Medical Systems, 36(1),* 241–277. DOI 10.1007/s10916-010-9473-4.
67. Reddy, A. R., Kumar, P. S. (2016). Predictive big data analytics in healthcare. *Second International Conference on Computational Intelligence Communication Technology,* pp. 623–626. Ghaziabad, India.
68. Dahl, Ø. (2019). The future role of big data and machine learning for health and safety inspection efficiency. https://osha.europa.eu/en/tools-and-resources/seminars.
69. Team, E. (2017). The importance of machine learning and of building data sets. InsideBIGDATA. https://insidebigdata.com/2017/10/11/importance-machine-learning-building-data-sets/.
70. Dilmegani, C. (2018). The ultimate guide to synthetic data: Uses, benefits and tools. https://research.aimultiple.com/synthetic-data/.
71. Sarkar, T. (2019). Synthetic data generation—A must-have skill for new data scientists. *Medium, 4(2).* https://towardsdatascience.com/synthetic-data-generation-a-must-have-skill-for-new-data-scientists-915896c0c1ae.

72. Kaur, J., Agrawal, A., Khan, R. A. (2020). Security issues in fog environment: A systematic literature review. *International Journal of Wireless Information Networks, 27(3),* 467–483. DOI 10.1007/s10776-020-00491-7.

73. Mohammad, M. N. A., Nazir, M., Mustafa, K. (2019). A systematic review and analytical evaluation of security requirements engineering approaches. *Arabian Journal for Science and Engineering, 44(11),* 8963–8987. DOI 10.1007/s13369-019-04067-3.