**Machine Learning Based Framework for Maintaining Privacy of Healthcare Data**

**Abstract**

The Adoption of Artificial Intelligence (AI), Machine Learning (ML), Internet of Things (IoT), cloud services, web-based software systems, and other wireless sensor devices in the healthcare infrastructure have led to phenomenal improvements and benefits in the healthcare sector. Digital healthcare has ensured early diagnosis of the diseases, greater accessibility, and mass outreach in terms of treatment. Despite this unprecedented success, the privacy and confidentiality of the healthcare data have become a major concern for all the stakeholders. Data breach reports reveal that the healthcare data industry is one of the key targets of cyber invaders. In fact the last few years have registered an unprecedented rise in healthcare data breaches. Hacking incidents and privilege abuse are the most common threats and have exposed sensitive and protected health data. Experts and researchers are working on various techniques, tools, and methods to address the security issues related to healthcare data. In this article, the main focus is on evaluating the impact of research studies done in the context of healthcare data breach reports to identify the contemporary privacy and confidentiality issues of sensitive healthcare data. Analysis of the research studies depicts that there is a need for proactive security mechanisms that will help the healthcare organizations to identify abnormal user behavior while accessing healthcare data. Moreover, studies also suggest that ML techniques would be highly effective in securing the privacy and confidentiality of the healthcare data. Working further on this premise, the present study also proposes a conceptual framework that will secure the privacy and confidentiality of healthcare data proactively. The proposed framework is based on ML techniques to detect deviated user access against Electronic Health Records. Further, fuzzy-based Analytical Network Process (ANP), a multi-criteria decision-making approach, is used to assess the accuracy of the supervised and unsupervised ML approaches for achieving a dynamic digital healthcare data security environment

**Introduction:**

The Electronic health (e-health) is an advanced form of conventional paper-based health infrastructure that practices advanced healthcare information technology to enhance the care services. To provide efficient, reliable, and cost-effective service to the patients, various ubiquitous technologies namely the Internet of Medical Things, smart devices, web-based applications, wireless sensor devices, telemedicine devices, AI, ML, and cloud services have been adopted by the healthcare service providers. These technologies have significantly improved the dissemination of healthcare data among various interested entities and facilitated in providing online care services such as online patient monitoring and telemedicine, etc. Implementation of Electronic Health Record (EHR) systems is one of the major outcomes of health information technology [1,2]. The EHR systems store the medical and treatment histories of patients as records and make them accessible to the authorized users all the time. However, these health records are frequently breached by the hackers. The EHR automates access to information and has the potential to streamline the clinician’s workflow. It improves the decision-making ability of the healthcare professionals and helps them in reducing the error rate in diagnosis. Nevertheless, several potential hazards threaten the privacy and confidentiality of protected healthcare data. 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. The healthcare sector is one of the top three sectors that are currently facing the highest number of breached incidents [3]. Authentic research studies and data breach reports reveal that the healthcare sector is highly susceptible to both internal and external intrusions. There have been 225 healthcare data breach episodes in the first half of 2020 itself. 130 of these breach instances were because of hacking/IT incidents which; i.e., 57.77% of the total number of breach cases [4], and 59 breaches were because of internal unauthorized access. Reports also cite that 3033 healthcare data breach incidents engineered during 2010 to 2019 exposed 255.18 million patients’ records in the USA [5]. Out of the 3033 breach cases, 850 were disclosed because of hacking/IT incidents, and 843 were exposed because of internal unauthorized access. The healthcare data industry is being victimized by both internal as well as external threats. Healthcare data breaches also harm the reputation of the organizations and service providers, resulting in the loss of patients’ trust and, consequently, loss of revenue [6]. One of the main reasons for the unprecedented rise in data breach cases is the selling value of health data records which is estimated to be ten to twenty times higher than the credit card data in the online market [7]. Facts and figures reveal that due to its’ high sensitivity and valuable character, healthcare data has become the most sought-after entity for intruders. Such a scenario calls for immediate intervention mechanisms to eliminate the possibilities of healthcare data pilferage. To address the privacy and confidentiality issues of healthcare data, the authors of this study have provided a theoretical Machine Learning framework that will detect suspicious user access to an EHR system that holds the patients’ health records. Machine Learning, as a co-domain of Artificial Intelligence, ensures that historical data programs (intelligent models) can be enlisted to learn, achieve experience, and improve system’s performance to classify, predict future trends and make decisions without human involvement [8,9]. Every correct future prediction or decision enhances the performance measure of the intelligent program. Artificial intelligence and machine learning have changed the way people think and play a significant role in different fields of life. ML-based models have been practiced in different areas to address real-life problems, and fortunately, in most of the areas, they have achieved the desired targets. Weather forecasting, disease analysis, and diagnosis, defense, sentimental analysis, marketing, traffic prediction, Fraud detection are some of the prominent examples of ML-based mechanisms. ML also has a significant role in cyber security because of its proactive character to detect misuse and anomalous behaviors such as intrusion, fraud, and email spam detection. Here we aim to address healthcare data’s confidentiality and privacy issues through the ML approach. To achieve the stated objective, the study proposes a theoretical ML-based framework that will describe the normal users’ behaviors of dynamic healthcare environment and detection of suspicious or abnormal user behaviors against the normal user profiles. This model will help the healthcare service providers to secure sensitive health data from suspicious accesses that result in healthcare data breaches. The rest of the work has been organized as: the second section of this research endeavor presents the analysis of some important existing research studies; the third section discusses the healthcare data privacy and confidentiality issue in the contemporary scenario; part 3.1-the sub-section discusses the arrival rate and inter-arrival time of different data disclosures. The fourth section describes the ML-based theoretical framework for suspicious user access detection. Section 4.1 provides Fuzzy-ANP based idealness assessment of ML approaches. The final section concludes the proposed study.

**LITERATURE REVIEW:** **“Security and privacy in ehealth: Is it possible?,” Authors: T. Sahama, L. Simpson and B. Lane**

Advances in Information and Communication Technologies have the potential to improve many facets of modern healthcare service delivery. The implementation of electronic health records systems is a critical part of an eHealth system. Despite the potential gains, there are several obstacles that limit the wider development of electronic health record systems. Among these are the perceived threats to the security and privacy of patients' health data, and a widely held belief that these cannot be adequately addressed. We hypothesize that the major concerns regarding eHealth security and privacy cannot be overcome through the implementation of technology alone. Human dimensions must be considered when analyzing the provision of the three fundamental information security goals: confidentiality, integrity and availability. A sociotechnical analysis to establish the information security and privacy requirements when designing and developing a given eHealth system is important and timely. A framework that accommodates consideration of the legislative requirements and human perspectives in addition to the technological measures is useful in developing a measurable and accountable eHealth system. Successful implementation of this approach would enable the possibilities, practicalities and sustainabilities of proposed eHealth systems to be realised.

**Edge-based IoT medical record system: Requirements, recommendations and conceptual design. Authors: A. F. Subahi**

This paper presents a conceptual architecture, design, and recommendation for the IoT Edge-based healthcare management system. The suggested architecture aims at distributing the workload of system performance (electronic healthcare services), including monitoring, diagnosis, prediction, as well as managing and archiving medical data of patients across different points of the system (at the edge and on the cloud). The proposed system design consists of two main subsystems (one for monitoring tasks and another one for medical record management activities). Both subsystems interact with multiple kinds of database systems (SQL and NoSQL). Transformational-based system for data migration is presented as a contribution of this paper. Two styles of transformation compositions are considered in the architectural design of transformation agents.

**Patient privacy violation detection in healthcare critical infrastructures: An investigation using density-based benchmarking, Authors: W. Hurst, A. Boddy, M. Merabti and N. Shone.**

Hospital critical infrastructures have a distinct threat vector, due to (i) a dependence on legacy software; (ii) the vast levels of interconnected medical devices; (iii) the use of multiple bespoke software and that (iv) electronic devices (e.g., laptops and PCs) are often shared by multiple users. In the UK, hospitals are currently upgrading towards the use of electronic patient record (EPR) systems. EPR systems and their data are replacing traditional paper records, providing access to patients’ test results and details of their overall care more efficiently. Paper records are no-longer stored at patients’ bedsides, but instead are accessible via electronic devices for the direct insertion of data. With over 83% of hospitals in the UK moving towards EPRs, access to this healthcare data needs to be monitored proactively for malicious activity. It is paramount that hospitals maintain patient trust and ensure that the information security principles of integrity, availability and confidentiality are upheld when deploying EPR systems. In this paper, an investigation methodology is presented towards the identification of anomalous behaviours within EPR datasets. Many security solutions focus on a perimeter-based approach; however, this approach alone is not enough to guarantee security, as can be seen from the many examples of breaches. Our proposed system can be complementary to existing security perimeter solutions. The system outlined in this research employs an internal-focused methodology for anomaly detection by using the Local Outlier Factor (LOF) and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithms for benchmarking behaviour, for assisting healthcare data analysts. Out of 90,385 unique IDs, DBSCAN finds 102 anomalies, whereas 358 are detected using LOF.

**ML-based cyber incident detection for Electronic Medical Record (EMR) systems Authors : D. McGlade and S. S. Hayward,**

An upward trend in cyber incidents across both U.K. and U.S. hospitals has been observed since 2015. Attacks range from identity theft to insurance fraud and extortion/blackmail. The [Electronic Medical Record](https://www.sciencedirect.com/topics/computer-science/electronic-medical-record) (EMR) systems used in hospitals are targeted due to the sensitivity of data within a healthcare setting. This work is motivated by the necessity to protect patient information and to ensure the availability of such EMR systems. A failure in either case can have grave implications for patients being treated and practitioners using the system. In this research, we propose the application of [Machine Learning](https://www.sciencedirect.com/topics/computer-science/machine-learning) (ML) and Time Series (TS) [anomaly detection](https://www.sciencedirect.com/topics/computer-science/anomaly-detection) to the problem of confidentiality and availability attacks on EMR systems. The results presented in this paper indicate that confidentiality incident detection is fully achievable using ML, with [Support Vector Machines](https://www.sciencedirect.com/topics/computer-science/support-vector-machine) obtaining the highest accuracy, precision and recall of a number of models tested. Results from the availability prototype show that the detection of a message surge is possible within 10 seconds, by using an Exponential Moving Average implementation to identify anomalies in message flow. This finding paves the way for an automated surge defence to be developed, presenting a significant advance over the manual method used today. The feasibility and practicality of implementing these detection systems in a clinical setting are also discussed with consideration of parameter tuning, skill-sets, and data protection.

**EXISTING SYSTEM:**

**PROPOSED SYSTEM:**

**CONCLUSION:**

The healthcare sector is one of the top data industries of the world that holds and processes sensitive and valuable data. Healthcare sector has become the most vulnerable target of intruders and the number of attacks against healthcare data is increasing rapidly. Healthcare data is being breached and exposed through Hacking (ransomware, malware, and fishing), and internal unauthorized access. The soaring cases of data theft call for more effective security mechanisms. The present research endeavor’s theoretical framework for ensuring healthcare data confidentiality based on supervised and unsupervised machine learning is an attempt in this direction. Implementation of these models will help the healthcare organisations to identify suspicious user accesses against the protected healthcare data in a more robust manner. Moreover, the effective implementation of the model will also reduce the time spent in investigation and the cost incurred in the process. This work also depicts that the unsupervised machine learning approach is an ideal option for maintaining healthcare data security as compared to the supervised approach. Research is a dynamic process, thus we cannot claim that our identified attribute set is optimal but it is also an ideal choice. Moreover, the proposed assessment approach- fuzzy-based ANP is an effective MCDM approach but not optimal. So, researchers can practice other techniques for better results if possible. In our future work, we will focus on the implementation of the proposed theoretical framework.

**REFERENCES:**

[1] T. Sahama, L. Simpson and B. Lane, “Security and privacy in ehealth: Is it possible?,” in Proc. 2013 IEEE 15th Int. Conf. on e-Health Networking, Applications and Services, Lisbon, Portugal, pp. 249–253, 2013. 710 IASC, 2021, vol.29, no.3 [2] A. F. Subahi, “Edge-based IoT medical record system: Requirements, recommendations and conceptual design,” IEEE Access, vol. 7, no. 5, pp. 94150–94159, 2019. [3] W. Hurst, A. Boddy, M. Merabti and N. Shone, “Patient privacy violation detection in healthcare critical infrastructures: An investigation using density-based benchmarking,” Future Internet, vol. 12, no. 6, pp. 100– 105, 2020. [4] June 2020 Healthcare Data Breach Report, HIPAA Journal, Jul. 24, 2020. 2021. [Online]. Available: https://www. hipaajournal.com/june-2020-healthcare-data-breach-report/. [5] A. H. Seh, M. Zarour, M. Alenezi, A. K. Sarkar, A. Agrawal et al., “Healthcare data breaches: Insights and implications,” Healthcare, vol. 8, no. 2, pp. 133–148, 2020. [6] A. A. Boxwala, J. Kim, J. M. Grillo and L. Ohno-Machado, “Using statistical and machine learning to help institutions detect suspicious access to electronic health records,” Journal of the American Medical Informatics Association, vol. 18, no. 4, pp. 498–505, 2011. [7] D. McGlade and S. S. Hayward, “ML-based cyber incident detection for Electronic Medical Record (EMR) systems,” Smart Health, vol. 12, no. 2, pp. 3–23, 2019. [8] A. H. Seh and P. K. Chaurasia, “A review on heart disease prediction using machine learning techniques,” International Journal of Management, IT and Engineering, vol. 9, no. 4, pp. 208–224, 2019. [9] E. Alpaydin, “Introduction to machine learning,” MIT press, vol. 9, no. 2, pp. 1–42, 2020. [10] R. D. Stachel and M. DeLaHaye, “Security breaches in healthcare data: An application of the actor-network theory,” Issues in Information Systems, vol. 16, no. 2, pp. 1–14, 2015. [11] J. Kwon and M. E. Johnson, “The market effect of healthcare security: Do patients care about data breaches?,” Vol. 16, no. 2, pp. 1–14, 2015. [12] M. Meisner, “Financial consequences of cyber attacks leading to data breaches in healthcare sector,” Copernican Journal of Finance & Accounting, vol. 6, no. 3, pp. 63–73, 2018. [13] M. Chernyshev, S. Zeadally and Z. Baig, “Healthcare data breaches: Implications for digital forensic readiness,” Journal of Medical Systems, vol. 43, no. 1, pp. 50, 2019. [14] A. H. Seh and P. K. Chaurasia, “Digging deeper into data breaches: An exploratory data analysis of hacking breaches over time,” Procedia Computer Science, vol. 151, pp. 1004–1009, 2019. [15] S. B. Wikina, “What caused the breach? An examination of use of information technology and health data breaches,” Perspectives in health Information Management, vol. 11, no. 6, pp. 1–15, 2014. [16] A. F. Subahi, “A model transformation approach for detecting distancing violations in weighted graphs,” Computer Systems Science and Engineering, vol. 36, no. 1, pp. 13–39, 2021. [17] How Much Data Is Created Every Day? [27 Powerful Stats], Seed Scientific, Jan. 30, 2020. 2021. [Online]. Available: https://seedscientific.com/how-much-data-is-created-every-day/. [18] How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read, 2021. [Online]. Available at: https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-daythe-mind-blowing-stats-everyone-should-read/#318c2c6d60ba. [19] S. Bajrić, “Data security and privacy issues in healthcare,” Applied Medical Informatics, vol. 42, no. 1, pp. 19–27, 2020. [20] Security and Privacy in the Medical Internet of Things: A Review. 2021. [Online]. Available at: https://www. hindawi.com/journals/scn/2018/5978636/. [21] September 2020 Healthcare Data Breach Report: 9.7 Million Records Compromised, HIPAA Journal, Oct. 22, 2020. 2021. [Online]. Available at: https://www.hipaajournal.com/september-2020-healthcare-data-breachreport-9-7-million-records-compromised/. [22] December 2019 Healthcare Data Breach Report. 2021. [Online]. Available at: https://www.hipaajournal.com/ december-2019-healthcare-data-breach-report/. [23] Largest Healthcare Data Breaches of 2016, HIPAA Journal, Jan. 04, 2017. 2021. [Online]. Available at: https:// www.hipaajournal.com/largest-healthcare-data-breaches-of-2016-8631/. IASC, 2021, vol.29, no.3 711 [24] January 2019 Healthcare Data Breach Report, HIPAA Journal, Feb. 25, 2019. 2021. [Online]. Available at: https:// www.hipaajournal.com/january-2019-healthcare-data-breach-report/. [25] January 2020 Healthcare Data Breach Report, HIPAA Journal, Feb. 21, 2020. 2021. [Online]. Available at: https:// www.hipaajournal.com/january-2020-healthcare-data-breach-report/. [26] U. Lechtenberg, “Research guides: Organizing academic research papers, theoretical framework,” 2020. [Online]. Available at: https://library.sacredheart.edu/c.php?g=29803&p=185919. [27] A. Agrawal, A. H. Seh, A. Baz, H. Alhakami, W. Alhakami et al., “Software security estimation using the hybrid fuzzy ANP-TOPSIS approach: Design tactics perspective,” Symmetry, vol. 12, no. 4, pp. 598–613, 2020. [28] J. W. Lee and S. H. Kim, “Using analytic network process and goal programming for interdependent information system project selection,” Computers and Operations Research, vol. 27, no. 6, pp. 367–382, 2000. [29] A. Solangi, “An integrated Delphi-AHP and fuzzy TOPSIS approach toward ranking and selection of renewable energy resources in Pakistan,” Processes, vol. 7, no. 118, pp. 1–18, 2019