ADDRESSING CYBER-SECURITY OF HEALTHCARE INFORMATION SYSTEM USING PROACTIVE SECURITY MECHANISM

A Abstract submitted to the
Babasaheb Bhimrao Ambedkar University, Lucknow
in fulfilment of Requirement for the Award of Degree of

Doctor of Philosophy

in Information Technology

BY

Adil Hussain Seh

Enrollment No. - 1670/19

Under the Supervision of

Prof. Raees Ahmad Khan

Under the Co-Supervision of

Dr. Pawan Kumar Chaurasia

DEPARTMENT OF INFORMATION TECHNOLOGY
SCHOOL OF INFORMATION SCIENCE AND TECHNOLOGY
BABASAHEB BHIMRAO AMBEDKAR UNIVERSITY
(A CENTRAL UNIVERSITY)
LUCKNOW, UTTAR PRADESH-226025

2022

ABSTRACT

The modern world is shrinking, more connected, and has become more volatile. Advancements in science and technology especially in information and communication technology (ICT) made the world a global village. Person to person, organization to organization, state to state, and country to country, communication, data generation, and information sharing, becomes more easy, comfortable, and more cost-effective due to ICT. Both the public and private sector organizations have swiftly switched their conventional infrastructure to advanced digital infrastructure. The digital revolution in modern-day organizations enhances the potential and significance of virtual communication. Data generation in huge quantities is one of the significant outcome of ICT and the digital revolution. It is estimated that by 2025 the amount of data produced each day throughout the world will be 463 Exabytes'. It reveals how swiftly data is generated by organizations and concerned stakeholders. And one of the prominent organization throughout the world that generates a huge amount of sensitive data is healthcare industry.

However, with the digital transformation data privacy and confidentiality remain a serious issue from the beginning for those organizations including the healthcare sector. In the last few years, the healthcare sector is one of the top sectors that has faced the highest number of cyber-attacks and lost a huge amount of sensitive electronic health records of patients in these attacks. Since 2005-19, 2072 data breaches have been faced by seven well-known organizations namely MED compromises all Healthcare Service Providers, EDU represents Educational Organizations, BSF includes Businesses-Financial, Insurance Institutes, and Organizations, BSO represents Businesses-Other, BSR represents Business-Retail Includes Online Retail, GOV represents Government and

Defense Institutes, and NGO represents Non-Governmental Organizations. Out of these 2072, data breaches healthcare sector alone faced 1587 data breaches which make 76.49% of the total. Even only in the year 2020, 656 data breaches have been recorded related to healthcare organizations that lead to the loss of more than 37 million patient health records.

These facts and figures depict the sensitivity and value of healthcare data and demand the immediate attention of experts and researchers to address healthcare cyber-security. Thus, we have proposed this research study to investigate the healthcare data breach issues and their impact and to provide a proactive security mechanism for a healthcare information system. For that, we have proposed a three-layer methodology that compromises comprehensive analysis of healthcare data breach issues, and their impacts on healthcare service providers and patients; identification and prioritization of cyber-security attributes to identify the most indispensable intrusion detection model using fuzzy based Analytical Network Process (ANP); and ensemble machine learning to build a hybrid suspicious user access detection model for a healthcare information system. ANP is a multi-criteria decision making approach that is well-known to address decision making problems. Incorporation of fuzzy logic with ANP makes it more suitable and ideal to generate reliable and effective results. Ensemble learning is a meta approach in the domain of Machine learning (ML) to enhance the predictive performance by integrating the predictions of various individual models without human intervention. Heterogeneous ensemble approach combines decisions of different heterogeneous base level models and makes a final decision based on the majority voting of the expert decisions. That not only enhances the predictive performance of the hybrid model but also improves the reliability of results.

Above discussed methodology has been practiced in three phases to conduct the experimental work and generate results for the current research

study. The first phase results of this research study reveal that hacking incidents, unauthorized access (internal), theft or loss, and improper disposal of unnecessary data are the main strategies that have been used to disclose sensitive healthcare records. Since 2010 to 2015 the number of hacking and unauthorized access (internal) incidents were low in number in comparison to theft/ loss. But from 2016 to 2020 the number of hacking attacks followed by unauthorized access (internal) shows an abrupt growth not only in number but also in magnitude. Whereas, the number of theft/ loss incidents depicts a decline in number. It depicts that digital transformation in the healthcare sector has made healthcare data more vulnerable.

More than 87% of the total hacking-related data breaches have been reported in the last five years (2016-20). It is not the only concern for healthcare service providers and patients but the cost of each breached health record increases rapidly year by year. The cost of each breached health record in the year 2010 was $294 which reach $444 in 2020. That is a 51% increase in just ten years which is another serious issue for both care providers and patients. In addition to this healthcare service providers face other problems due to these data breaches. These problems are reputation defamation, customer attrition, and customer legal complaints. Generally, sensitive healthcare data disclosure puts patients at risk, but when it is about celebrities it becomes a more serious and problematic issue for them. In the case of celebrities sometimes it may lead them to the verge of suicide, because of their privacy sensitiveness regarding their health issues. Further, it may lead to the loss of precious human lives if the integrity of the electronic health records of patients has been compromised.

With the help of the second phase methodology and expert consultation, we have identified a set of seven cyber-security attributes from an ML perspective to find out the most indispensable user intrusion

detection model for digital healthcare infrastructure. These attributes are Anomaly detection, Misuse detection, DoS attack detection, Spam detection, Phishing detection, and Implementation complexity. Experimental results depict that Anomaly detection followed by misuse detection are two important cyber-security attributes that have got the highest priority with respect to healthcare cyber-security using the fuzzy-ANP approach.

Finally with the help of the third phase of our proposed methodology first we collected a dataset of 2000 user log records from a UK-based hospital and labeling of the dataset has been performed to make it suitable for supervised ML. Then we conducted a baseline experiment on seven well-known supervised ML classifiers. Data set of 2000 user log records have been used to conduct the baseline experiments for these seven classifiers. We have examined from the baseline experiments that Decision Tree, Logistic Regression, and Support Vector Machine classifier are most ideal and effective for our proposed ensemble learning model as baseline classifiers. Grid SearchCV and Random SearchCV hyperparameter optimization techniques have been incorporated to improve the performance of our proposed model.

The results of this study depict that our proposed hybrid ensemble learning model not only improves the accuracy by approximately 3 to 10% with respect to the existing proposed models but also improves the F1-score by up to 5% which is a significant achievement of our proposed study. It controls the false negative rate up to the highest level and improves sensitivity (recall) by approximately 7 to 23% as compared to the existing proposed studies which reveal the significance of our proposed model. The accuracy of our proposed model reaches up to 99.9% and the F1-score is up to 0.99. Our proposed model not only improves accuracy and F1-score in comparison to existing proposed models but also provides maximum control

on false negative predictions that remain a challenging and serious issue in healthcare cyber-security like domains. As false negative predictions have a high cost impact and calamitous consequences as compared to false positive predictions in such areas.

Conclusively, our proposed study will help healthcare service providers and researchers to get deep insights and implications about data privacy breaching issues in the healthcare sector and use their potential and resources accordingly. Further, our proposed proactive security mechanism will help healthcare service providers to detect suspicious user accesses carried out to gain access to electronic health records of patients, and save their resources to a good extent in the form of time, money, and manpower.