A survey on health care data using Machine learning Algorithms

CURRENT STATUS: POSTED

Manjulatha B
Vignana Bharathi Institute of Technology

Correo: manjulathareddy86@gmail.com Corresponding Author
ORCiD: https://orcid.org/0000-0002-9094-1941

Suresh Pabboju
CBIT

DOI:
10.21203/rs.3.rs-16736/v1

KEYWORDS
Big Data, Electronic health records, Analytical tools, machine learning and Security algorithms
Abstract
Big data in the healthcare has been growing rapidly in recent years. A massive amount of patient’s data is generated through various sources in this digitalized environment including wearable devices, mobile devices, electronic health records and medical centres. The medical data, which are exploding exponentially includes a large volume of structured and unstructured data. Patients information is shared and managed by many sources such as insurance, in the form of prescription, healthcare providers including the attackers which causes a severe security threats their confidential data. Cybercriminals use a variety of social engineering attacks like phishing, leakage of data, embedding virus and trojan horse into the attachments, DNS snooping, sniffing, and many more. In this paper a brief overview is done on different data mining and security techniques which are implemented in healthcare data. The study and its presentation in this context would be helpful in not only analysing but also predicting chronic diseases depending upon the available stored data which has been collected from various sources in the area of medical expert system.

1. Introduction:
Big Data has been a pure revolution for the healthcare industry. Starting from understanding the consumers or patients to personalized services it has come a long way. With technologies like trend and predictive analysis, it has optimized the growth by multiple times. Though, this is just the beginning. With the development of technologies, new and better treatments and diagnosis saving more lives by curing more diseases are possible.

2. Role Of Big Data Analytics In Health Care:
Health data [1] plays a vital role in today’s environment includes clinical metrics along with environmental, socioeconomic, and behavioural information pertinent to health and wellness. Big data helps to improve care personalization and efficiency with comprehensive patient profiles, identifies geographic markets with a high potential for growth and also provides straightforward identification of patterns in health outcomes, patient satisfaction, and hospital growth. Also helps in boosting the healthcare marketing efforts with information about consumer, patient, and physician needs and preferences. Thus, analytics is becoming very crucial in tracking different types of healthcare trends.

2.1 Applications of Big Data Analytics in Healthcare:
2.1.1 Patient’s Health Tracking:
Big Data and Analytics along with the Internet of Things (IoT), is revolutionizing the way one can track various user statistics and vitals. Apart from the basic wearables like Fitbit, apple watch that can detect the patient's sleep, heart rate, exercise, distance walked, etc. there are new medical innovations that can monitor the patient’s blood pressure, pulse Oximeters, glucose monitors, and more. The continuous monitoring with these sensor devices allows the organizations to keep people out of visiting the hospitals.

**2.1.2 Reducing the cost:**
Predictive analytics will reduce the cost and also saves time for the patients by using android apps and IOT devices for tracking of their health.

**2.1.3 Preventing Human Errors:**
Sometimes most experienced persons also do mistakes by giving wrong medicines or dispatching a different medication which causes a great problem to the patient's health. Such errors, in general can be reduced since Big Data analyses such user’s data and the prescribed medication.

Big Data also helps in increasing the ability of healthcare sectors to cure chronic diseases, begin early preventive care, improves the quality of life.

**3. Health Care Data Breaches: Hidden Dangers And Causes:**
The current situation with healthcare data is extremely dangerous, as patient health information can be sold or used for crimes such as identity theft and insurance fraud, or to illegally obtain prescription drugs.

Outsider threats continue to present new challenges, but hidden insider threats are even more dangerous.

**3.1 Electronic Health Record Systems**
Most of the healthcare providers use Electronic Health Records (EHR) to store patient’s sensitive data such as Name, Bank account information, Insurance information, Family history, Contact information etc.

**3.2 Protecting the network**
Using firewalls, antivirus software and some security tools which limits the damage when hackers use a variety of methods to steal the confidential information.

**3.3 Educate Staff members**
With or without user’s intervention employees are often involved in healthcare data breaches. So, training is to be given to avoid the errors and equips healthcare employees in making smart decisions and using appropriate caution when handling patient data.

3.4 Delete unnecessary data
The more data present in the organization, the more there is a chance for criminals to steal. So, it is the responsibility of the organization to delete the patient information which is no longer needed.

3.5 Encrypt portable devices
Encryption is one of the most useful data protection methods for healthcare organizations. One thing healthcare organization should always do is to prevent security breaches is by encrypting all devices which holds patient data.

3.5 Secure Mobile Devices
Every organization should create a mobile device policy that governs what data can be stored on those gadgets.

4. Iot And Big Data Analytics In Healthcare Industry
IOT adds great value to the healthcare industry. The devices generate data about health of a person and send it to the cloud will lead to a plethora of insights about an individual’s heart rate, weight, blood pressure, lifestyle and much more. Through Big Data real-time monitoring of patients can be done. This will help in proactive care. The sensors and wearable devices will collect patient health data even from home and will help to monitor the healthcare institutions. This will also provide remote health alerts and lifesaving insights to their patients.

Smartphones have added a new dimension. The apps enable the smartphone to be used as a calorie counter to keep track of calories; pedometers to keep a check on how much you walk in a day. All these have helped people live a healthier lifestyle. Moreover, this data could be shared with a doctor, which will help with personalized care and treatment. Patients can make lifestyle choices to remain healthy.

Related Work:
Recent studies briefly describe about techniques used for handling healthcare data and providing security for the confidential patient’s data which generates vast amount of data on daily basis. This
survey gives a brief review which many researchers have done.

Fadoua Khennoua [3] investigated the existing EHRs to analyse the data which is unstructured and present the data in a meaningful model through analytical and decision-making tools such as Kibana (Analytics reports/dashboards/graphs), Couchbase /NOSQL for Database etc.

Aris Gkoulalas-Divanisa [4] reviewed more than 45 algorithms and highlighted their advantages and disadvantages. To achieve privacy in while preserving patient’s data by using cryptography and access control techniques, data partitioning, data clustering, space mapping, genetic search, horizontal data partitioning, space clustering, vertical data partitioning, top down space partitioning, greedy search, suppress control etc.

J. Archenaa [5] given an insight about how health care and government organisations generate data and security issues in handling that big data. Big data helps in identifying expiry date of medicine and surgical instruments based on RFID data. Improving the treatment methods improves the quality of healthcare data. To address the basic needs in government sectors such as telephone, water, ration card and gas connection, providing quality education, to reduce unemployment rate etc. And also discussed about security tools and methods used for storing, processing and encrypting the data using Attribute based encryption method, HDFS, Flume, Sqoop, MapReduce, HBase, Hive, Storm, Intelicius and hunk.

Tapas Ranjan Baitharu [6] extended his work in detecting Liver disease at the early stage by using various classification models in WEKA tool which gives better performance.

Hiba Asria [7] conducted comparison between different classification algorithms in WEKA tool by taking the performance indicators like accuracy, precision, sensitivity and sensitivity. Conclusion of this paper is, SVM has proven its efficiency and reaches its accuracy of 97.13% in predicting and diagnosing the Breast Cancer in women.

Mehrbakhsh Nilashi [8] used machine learning techniques such as SOM, PCA and NN for clustering, noise removal, PCA for dimensionality reduction etc to classify diabetes disease by taking Pima aboriginals diabetes dataset using Matlab7.10. He also concluded that PCA-SOM-NN has better accuracy (92.28%) compared to other classification systems. In future work, diabetes data can be
classified and analysed by using incremental machine learning approaches.

Abdullah [9] identified effectiveness of different treatment types for different age groups by using Oracle Data Miner, SVM which predicts diabetes patients and also its treatment plan.

Dr Saravana Kumar N M [10] used Hadoop environment in predicting the diabetes and its complications in rural areas to get proper treatment at low cost.

Wullianallur Raghupathi [11] given a brief overview on how Big Data helps in healthcare sectors and also discussed about frameworks used and its challenges.

S. Saranya [12] focused on storing and processing huge amount of data using HDFS, MongoDB and also proposed a triple encryption scheme which uses security algorithms using Apache Sentry.

Kevin Peterson [13] presented a Blockchain-based approach which provides publicly accessible APIs for exchanging Electronic Health Records which gives a rapid improvement in computational power.

Shalini Bhartiya [14] proposed an access control framework that applies hierarchy similarity analyser using XACML policies and these policy conflicts can be used in semantic, syntactic, temporal constraints in future scope.

Xueping Liang [15] implemented a blockchain based system on Hyperledger Fabric [16], which gives a clear scenario on how personal health care data is secured. Future enhancements of this study describe how personal health data and medical data are combined.

Xueping Liang [17] proposed a ProvChain which provides security features with low overhead for the cloud storage applications.

Jigna J Hathaliya [18] designed an AVISPA tool which identifies various security threats in accessing EHR from the database repository.

Sha Liu [19] developed a new scheme to provide a secure communication channel using Bluetooth. In future work AVISPA tool is used to identify security issues which are found while processing the data between devices.

Palanisamy [20] reviewed on the works carried out by various researchers from various sources ranging from electronic health records to medical images using big data frameworks.

Christy. A [21] has clearly shown a better accuracy between cluster based and distance-based
methods by taking esoph and diabetics datasets.

Ioannis Kavakiotis [22] focussed on diagnosis, predictions and complications present in diabetic patients using machine learning algorithms.

Kambiz Ghazinour [23] proposed and implemented a model to collect user’s personal credentials like email, address, mobile number etc and provides privacy preferences which are collected from the apps installed on their wearable devices.

Ms. Ashwini Mandale [24] addresses the challenges like heterogeneity, scale, timeliness, complexity, privacy problem by using machine learning tools.

Muni Kumar N [25] provides better outcomes in health care at lower costs for rural people which contains massive amount of data using Hadoop and MapReduce techniques using healthcare datasets.

Minerva Panda [26] implemented an application which provides a solution for patients as well as doctors to protect the patients during emergency.

Dasari Madhavi [27] provided a solution for de-identifying personal health information, Map Reduce application uses jar files which contain a combination of MR code and PIG queries. This application also uses advanced mechanism of using UDF (User Data File) which is used to protect the health care dataset.

Dhina Suresh [28] proposed a framework for protecting the confidential information of patient which is stored in electronic health records by using ECC cryptographic algorithm.

Veena H. Bhat [29] adapted SEMMA methodology to calculate highest accuracy by using CART and Genetic Algorithm to detect occurrence of diseases and also for tracking the various parameters which influences the occurrences of diseases. The proposed model evaluated with 82.33% accuracy.

Sabibullah M [30] in his proposed work, a Genetic Algorithm (GA) based model which is implemented using .Net framework [Visual studio 2008, C#.Net to predict heart attack or other chronic diseases.

Anders Andersena [31] focused on providing privacy for the clinical data and implemented a practical approach to meet the security challenges by integrating with SNOOP framework.

Haider DhiaZubaydi [32] a reviewed that most health care applications are developed on the
Ethereum framework. The ECDSA Algorithm [39] is used in blockchain to prevent from hacking and he also proposed Byzantine Fault Tolerance algorithm with IoT devices because it does not require high hashing power.

Youssef M. Essa [33] implemented an Intelligent Framework for Healthcare Data Security which provides security for the sensitive data in healthcare by using DES encryption algorithm using Hadoop cluster.

Xiong Li [34] used biometric system to overcome the drawbacks in existing systems in enhancing the security measures taken to protect the data.

In this paper, the author, Asaph Azaria [35] proposed MedRec approach which manages all crucial considerations when handling sensitive information which is stored in EMRs.

Karim Abouelmehdi [36] discussed about existing challenges towards security various attacks and suitable cryptographic algorithms that are used for protecting the security and privacy of healthcare data.

In this paper, the author proposed Guang Yang [37] an independent architecture that implements blockchain technology in EHR systems for the integrity of healthcare records and improves the interoperability of the current systems which includes the creation, verification, and appending of new blocks.

Peng Zhang [38] applies blockchain technology to share the clinical data by analysing the ONC requirements and providing an FHIRChain architecture to authenticate digital health data by using decentralized app.

In this paper, authors i.e. Lobna Yehia [39] reviewed about security techniques to be used and applied in IOT devices to provide security for the data as there is a huge amount of information related to health care is stored.

Cheryl Ann Alexander [40] predicted heart attack and also proposes medical treatment to every individual by using Big Data Analytics techniques. Also evaluated security challenges to protect the health care data.

Arash Ghazvini [41] explored and analysed the various security challenges faced by the e-health
systems and also discussed about security mechanisms to protect patient’s data.

K. Rajesh [42] worked on Diabetes dataset to achieve greater accuracy by applying classification algorithms and showed the result that among many classification algorithms C4.5 algorithm has 91% classification rate.

D. Peter Augustine [43] discussed about the benefits of Big Data technologies such as Hadoop which stores, process a massive amount of data which can be used by various sectors like healthcare, government sector etc.

Jorge Almeida [44] presents the methodology for exporting data to standard formats and performing analysis on openEHR repository to standard tabular formats, enabling further form reports and data analysis using other software.

S. Packiyam [45] presented approaches for collecting and storing Big Data for analytics using classification algorithms and also predicts HIV/AIDS, TB and silicosis by implementing a framework.

Conclusions:
By this literature survey, we can come to conclusion that as data generated from various sources is increasing day by day handling and securing that massive collection of data become a hectic task. So, by implementing big data techniques and security measures patient’s confidential data is protected from cyber criminals. Therefore, implementing Big Data Analytics techniques into healthcare plays a vital role using various preventive measures to protect the sensitive data from cybercriminals.

Future Scope:
This paper can be enhanced further by applying various machine learning algorithms for predicting liver diseases in an early stage and also implementing some security mechanisms to protect the data which has been generated from various health care sources.

Abbreviations:
RFID: Radio Frequency Identification; IOT: Internet of Things; EHR: Electronic Health Records; UDF: User Data File; EMR: Electronic Medical Records; ECC: Elliptic Curve Cryptography; HDFS: Hadoop Distributed File System.

Declarations
Acknowledgements:

The authors are grateful and pleased to all the researchers in this research study.

Funding:

The authors state that this work has not received any funding.

Author’s contributions:

B. Manjulatha as an OU Scholar under the guidance of Dr. Suresh Pabboju took a supervisory role and reviewed various journals to identify the problems in healthcare data. All authors read and approved the final manuscript.

Competing interests:

The authors declare that they have no competing interests.

Declarations:

NA

Availability of data and materials:

NA

References

[1] https://healthinformatics.uic.edu/blog/how-health-care-analytics-improves-patient-care/

[2] https://www.datapine.com/blog/big-data-examples-in-healthcare/

[3] Fadoua Khennoua, Improving the Use of Big Data Analytics within Electronic Health Records: A Case Study based OpenEHR. Procedia Computer Science 127 (2018) 60–68.

[4] Aris Gkoulalas-Divanisa, Publishing data from electronic health records while preserving privacy: A survey of algorithms. Journal of Biomedical Informatics 50 (2014) 4–19.
[5] J. Archenaa, A Survey of Big Data Analytics in Healthcare and Government. Procedia Computer Science 50 (2015) 408 - 413.

[6] Tapas Ranjan Baitharu, Analysis of Data Mining Techniques for Healthcare Decision Support System Using Liver Disorder Dataset. Procedia Computer Science 85 (2016) 862 - 870.

[7] Hiba Asria, Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis. Procedia Computer Science 83 (2016) 1064 - 1069.

[8] Mehrbakhsh Nilashi, Accuracy improvement for diabetes disease classification: A case on a public medical dataset. Fuzzy Inf. Eng. (2017) 9; 345-357.

[9] Abdullah, Application of data mining: Diabetes health care in young and old patients. Journal of King Saud University- Computer and Information Sciences (2013) 25,127-136.

[10] Dr Saravana Kumar N M, Predictive Methodology for Diabetic Data Analysis in Big Data, Procedia Computer Science 50 (2015) 203 - 208.

[11] Wullianallur Raghupathi, Big data analytics in healthcare: promise and potential, Raghupathi and Raghupathi Health Information Science and Systems 2014.

[12] S. Saranya, Dynamic Preclusion of Encroachment in Hadoop Distributed File System, Procedia Computer Science 50 (2015) 531 - 536.

[13] A Blockchain-Based Approach to Health Information Exchange Networks by Kevin Peterson, Rammohan Deeduvanu, Pradip Kanjamala, and Kelly Boles.

[14] Shalini Bhartiya, Proposing hierarchy similarity-based access control framework: A multilevel EHR data sharing approach for interoperable environment, Journal of King Saud University- Computer and Information Sciences (2017) 29,505-519.

[15] Xueping Liang, Integrating Blockchain for Data Sharing and Collaboration in Mobile Healthcare Applications,2017 IEEE.

[16] C. Cachin, Architecture of the Hyperledger blockchain fabric,“ in Workshop on Distributed Cryptocurrencies and Consensus Ledgers, 2016.

[17] Xueping Liang, Prov Chain: A Blockchain-based Data Provenance Architecture in Cloud Environment with Enhanced Privacy and Availability, 2017 17th IEEE/ACM International Symposium on
Cluster, Cloud and Grid Computing, 468-477.

[18] Jigna J Hathaliya, Securing electronics healthcare records in Healthcare 4.0: A biometric-based approach, Computers and Electrical Engineering 76 (2019) 398-410.

[19] Sha Liu, A novel asymmetric three-party based authentication scheme in wearable devices environment, Elseiver, 2015.

[20] Palanisamy, V., Thirunavukarasu, R. Implications of big data analytics in developing healthcare frameworks - A review. Journal of King Saud University - Computer and Information Sciences (2017).

[21] Christy. A, Cluster Based Outlier Detection Algorithm for Healthcare Data, Procedia Computer Science 50 (2015) 209 - 215.

[22] Ioannis Kavakiotis, Machine learning and data mining methods in Diabetes research, Computational and Structural Biotechnology Journal 15 (2017) 104-116.

[23] Kambiz Ghazinour, A Model to Protect Sharing Sensitive Information in Smart Watches, Procedia Computer Science 113 (2017) 105–112.

[24] Ms. Ashwini Mandale, Big Data Analytics: Challenges, Tools, International Journal of Innovative Research in Computer Science & Technology (IJJRCST), 2015.

[25] Muni Kumar N, Role of Big Data Analytics in Rural Health Care - A Step Towards Svasth Bharath, International Journal of Computer Science and Information Technologies, 2014.

[26] Minerva Panda, Big Data in Health Care: A Mobile Based Solution, 2017 IEEE.

[27] Dasari Madhavi, De-Identified Personal Health Care System Using Hadoop, International Journal of Electrical and Computer Engineering (IJECE), 2015.

[28] Dhina Suresh, Securing Personal Health Record System in Cloud Using User Usage Based Encryption, Springer, 2019.

[29] Veena H. Bhat, An Efficient Prediction Model for Diabetic Database Using Soft Computing Techniques, Springer, 2009.

[30] Sabibullah M, Diabetes Patient’s Risk through Soft Computing Model, International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), 2013.

[31] Anders Andersena, Privacy preserving distributed computation of community health research
data, Procedia Computer Science 113 (2017) 633–640.

[32] Haider DhiaZubaydi, A Review on the Role of Blockchain Technology in the Healthcare Domain, MDPI, 2019.

[33] Youssef M. Essa, IFHDS: Intelligent Framework for Securing Healthcare Big Data, Springer, 2019.

[34] Xiong Li, A new authentication protocol for healthcare applications using wireless medical sensor networks with user anonymity, SECURITY AND COMMUNICATION NETWORKS Security Comm. Networks 2016; 9:2643–2655.

[35] Asaph Azaria, MedRec: Using Blockchain for Medical Data Access and Permission Management, IEEE, 2016.

[36] Karim ABOUELMEHDI, Big data security and privacy in healthcare: A Review, Procedia Computer Science 113 (2017) 73–80.

[37] Guang Yang, A blockchain-based architecture for securing electronic health record systems, Wiley, 2019.

[38] Peng Zhang, FHIRChain: Applying Blockchain to Securely and Scalably Share Clinical Data, Computational and Structural Journal 16(2018) 267-278.

[39] Lobna Yehia, Hybrid Security Techniques for Internet of Things Healthcare Applications, Advances in Internet of Things, 2015, 5, 21–25.

[40] Cheryl Ann Alexander, Big Data Analytics in Heart Attack Prediction, Journal of Nursing and Care, 2017.

[41] Arash Ghazvini, Security Challenges and Success Factors of Electronic Healthcare System, Procedia Technology 11 (2013) 212 – 219.

[42] K. Rajesh, Application of Data Mining Methods and Techniques for Diabetes Diagnosis, International Journal of Engineering and Innovative Technology (IJEIT), 2012.

[43] D. Peter Augustine, Leveraging Big Data Analytics and Hadoop in Developing India’s Healthcare Services, International Journal of Computer Applications (0975 – 8887).

[44] Jorge Almeida, Exporting data from an openEHR repository to standard formats, Procedia Technology 16 (2014) 1391 – 1396.
Figures

Figure 1

Legends are not available in this version.
Diagnostics
Data mining and analysis to identify causes of illness

Preventative medicine
Predictive analytics and data analysis of genetic, lifestyle, and social circumstances to prevent disease

Precision medicine
Leveraging aggregate data to drive hyper-personalized care

Medical research
Data-driven medical and pharmacological research to cure disease and discover new treatments and medicines

Reduction of adverse medication events
Harnessing of big data to spot medication errors and flag potential adverse reactions

Cost reduction
Identification of value that drives better patient outcomes for long term savings

Population health
Monitor big data to identify disease trends and health strategies based on demographics, geography, and socio-economics

Figure 2