The Significance of Big Data Analytics in reducing Data Breaches and Scams in the Medical Industry

ESWAR SIDDHARTH MANIVANNAN[1], DEEPANJALI CHANDRASEKARAN[2], AKSHAYA SHREE[3]

[1] UG Scholar, SRM Institute of Science and Technology, Ramapuram, Chennai, India

[2] PG Scholar, The University of Texas at Arlington, Texas, United States

[3] UG Scholar, SRM Institute of Science and Technology, Ramapuram, Chennai, India

***Abstract*— The advancements in data analytics have made it possible for analysts and healthcare workers to extract useful information and simulate clinical insights in such a network with respect to the availability of contemporary computer simulations in the healthcare industry. As per Fortified Health Security's mid-year report, there were 337 breaches in the medical sector within the initial half of 2022. Over 19 million records were involved in medical data breaches during the first half of the year. The health sector has become a prime target for those seeking to make quick cash using unethical means. Since life expectancy increases, it is projected that healthcare scams will become more prevalent. Fraud can be frequently resolved by big data,but perhaps employees make unintentional data entry errors. Despite the fact that such advancements undoubtedly increased the medical sector's efficiency, they have also given rise to various sets of issues involvingpatient data breaches. This paper describes the opportunities of big data analytics in the field of fraud detection in healthcare services and use of Machine Learning Algorithms to reduce data breach.**

***Indexed Terms*— Big Data Analytics, Healthcare 4.0, Data Breach, Fraud Management.**

I. INTRODUCTION

As the digital transformation age reached the medical industry, the number of scams and frauds began to rise. To continuously reinvent healthcare, clinical technology, data management, electronic health records, and portable and mobile devices are used. Healthcare 4.0 is more about acquiring enormous data sets and putting them to use in implementations, which will further guide business decisions for the medical industry and lead to significant performance and cost reduction benefits. The efficacy of any healthcare domain or renowned hospital depends on records and data. The interaction between the hospital, the doctor, and the patient relies greatly on the patient's medical database. As the need for healthcare is now increasing, especially in light of the pandemic, maintaining enormous amounts of data is already a major problem. However, big data analytics comes to the rescue by keeping everything automated despite the great amount of data that is stored and is never lost. Big data analytics are essential for a structured and appropriate use of the data in the healthcare industry. Additionally, there is a significant chance that a data breach will occur, endangering some crucial information and confidential data that might lead to much more problems. The database enables patients or even doctors to utilize the information for anyone at any moment, but also improved technology enables efficient, accurate, and creative variability detection and quick responses. In order to explore detection and mapping, associations as well as other observations, big data analytics evaluates huge volumes of data. It's indeed feasible to examine the data and also get responses instantly with current technology, an attempt with more common analytical tools is weaker but also less effective. Fraud detection is really a series of practices conducted to avoid the acquisition of valuables under illegal activities. Fraud detection can be used in many fields, including financial services, medical coverage, employment, hospital administration, and so on.

II. RELATED WORKS

Doctors who are part of medical schemes and who supply services to other healthcare providers are the fraudsters who commit the crime the most frequently. The most frequent instance of fraud is an insurance claim that is duplicated or not submitted at all. By gathering data and comparing the types and laws that apply to fraudsters in different nations, researchers may be able to identify the flaws in the fraud detection method, which could then be taken into account when deciding whether to impose legal sanctions on fraud, regardless of the status of the employee or profession.[1] A new and stronger approach to data integrity is required in the healthcare sector. Use of inclusion and exclusion criteria to sort the research that is most pertinent. At the initial stage, 110 experiments were listed, out of which 89 accounts were verified through database examination, and an additional 21 accounts were verified through additional offline sources like conference proceedings, symposium reports, books, etc.[2]

Investigators can more effectively use such approaches to spot trends and patterns, address problem areas, and even predict future cyber crimes given the association between analysis proficiency and the features of various cybercrimes. The detection procedure needs to be adaptable so that the system can handle the way that crimes are committed, which is always evolving. Comparability tests are a crucial tool for identifying unsolved crimes in crime patterns.[3] Over an increasing number of years, CMS has made numerous Big Data Medicare claims datasets available for public use. A novel method (combining multiple Medicare datasets and leveraging cutting-edge Big Data processing and machine learning approaches) for evaluating the fraud detection abilities of three Medicare datasets, separately and together, using three learners, against actual fraudulent doctors and other healthcare providers taken from the LEIE dataset are presented. Methods for processing the individual CMS datasets, the combined dataset, and the provider fraud label mapping are presented [4].

Almost every area of our real-world lives can benefit from changes and improvements brought about bydata science modeling, provided the necessary data is available for analysis. The key to data science modeling in any application domain is to gather the appropriate data and extract relevant information or actionable insights from the data for making wise decisions.[5] Through allocating some computational requirements, our medical reporting framework leverages the strength of Big Data Analytics applications. A Big Data Analytics-based platform that has the potential to peruse a vast percentage of statistical information in really a short span of time, such that proper cybercrime could be used for financial activities. Scam identification and protection are specifically illustrated as follows: fraudulent audit guidelines and the scorecard for scam analysis.[6] Today, leading departments are now using Industry for obtaining higher relative growth performance. Medical records comprise private communications regarding patients including such names, addresses, and specifics of the disease. The designed methodology implements its encryption mechanism to secure private information.[7]

Each phase in managing big data is characterized by different complexities that must be solved using powerful computational strategies for Bigdata analytics. Few violations of encryption, malware, cyber attempts, and incidents of vulnerabilities make network security concern for medical institutions. [8] There is some privacy inpersonal data from healthcare information. Some privacy protection methods which are used worldwide are protection laws and preservation of personal information which includes De-identification. [9] Big data analytics in clinical practice helps millions of individuals to evaluate massive datasets, recognize groups but also database associations, and create statistical models utilizing machine learning methods.[10]

Many drawbacks of conventional techniques were solved by Big Data and therefore will introduce developments in healthcare. Efficient processing of massive data sources is Hadoop-enhanced computation, and Hadoop-based Data Analytics have performance, and accuracy, including extensibility benefits.[11] The scientific world was already drawn to the rapid growth of the importance of health data to obtain and analyze its latest knowledge through big data analytics. Throughout the medical industry, several broad data alternatives are required, like documentation, biometrics, digital medical records, scans, patient reviews, and online activities.[12]

Forgery in medical care is constantly seen being one of the most significant societal issues. Medical abuse is obviously a concern for the country as well as the need for even more efficient identification techniques. It takes quite a bit of time and can identify hospital abuse through better diagnostic expertise.[13] Among the most prevalent incidences of corruption was its medical insurance. Scams may occur in several processes within the medical field, such as 'Up-coding: payments for much more costly procedures and treatments that have been currently rendered or undertaken' or 'paying for just a procedure never offered' or 'doing unnecessary diagnostic services solely and obtaining medical expenses.[14]

The identification of data breaches is mostly discovered via manual processes by accountants or researchers looking at different documents to identify irregular or deceptive activities. Especially in comparison to much more advanced statistical and machine learning techniques for detecting abuse, the human method of filtering fake information and predicting data breaches from a large volume of data could be repetitive and also very unreliable.[15] Since the patient and healthcare information for fraud prevention primarily resides in insurance providers which also include individuals and organizations' medical firms, it's really a tough problem to analyze information. The unprocessed report is high via doctor injury claims records, the doctor prescribing information, and payments and financial transactions policy customers and providers.[16]

As crime grows increasingly challenging as well as the number of information increases, it becomes hard to comprehend crime from an immense amount of information. They might not even eradicatecorruption, and we would definitely be able to minimize it. In addition to providing information, machine learning reveals trends concealed in data.[17] Frauds are blowing a hole in the insurance market. Medical insurance is a cash-strapped market with a large claim ratio. As a result, if the health insurance market is to be clear of fraud, it must depend on the removal or minimization of false claims earned from medical insurance. [18]

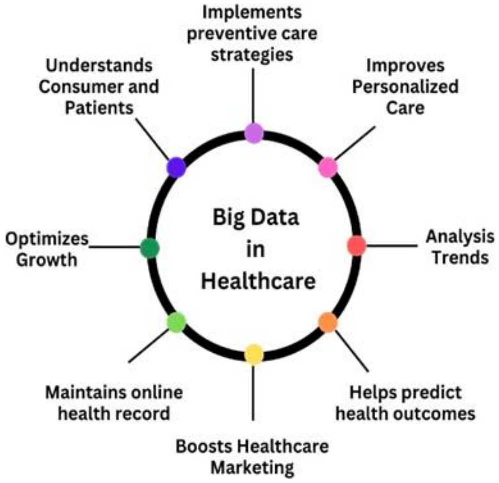
Implements preventive care strategies
Understands Consumer and Patients
Improves Personalized Care
Big Data in Healthcare
Optimizes Growth
Analysis Trends
Maintains online health record
Helps predict health outcomes
Boosts Healthcare Marketing

Fig.I Big Data Analytics in Healthcare

III. BIG DATA ANALYTICS IN HEALTHCARE

Big data analytics is still a framework that explains massive data sets produced through the use of technology that captures information from users and monitors the output of hospitals which is too wide and complicated with conventional methods. Its aim is to provide quality care, will save costs, and create productivity throughout all divisions. Fig 1 shows how Big Data Analytics is leading the healthcare industry in various ways. Healthcare providers must lay money on data analytics. Recently, big data is assisting doctors and hospitals to provide a treatment that is more personalized and has improved accurate outcomes. There are varieties of medical scams whichincludes the following fraudulent activities:

- Clinical Abuse of Personality: The abuse ofclinical identification includes the misuse of a patient's identities to procure personal care products, facilities, or funding fraudulently
- Excessive billing amount: Charging amounts for unreasonable products or services.
- Upcoding: Upcoding is a deceptive clinical payment where a payment submitted for a healthcare service is much more costly than the treatment which was offered could be focused on.
- Unbundling: When different process protocols were charged for a community of operations protected by either a comprehensive measurement system,unbundling arises.

Kickbacks: Kickbacks might also be described as providing, requesting, charging, or obtaining compensation to encourage and in exchange for all the recommendations of entities to supply or arrange some products or services over which fee might have been made through Federal insurance programs.

IV. EXISTING SYSTEM

The emergence of modern technology has various types of big data techniques which help in crime detection. Many Firms can monitor electronic payments for anomalies and identify unauthorized activity through advanced machine-learning techniques. These instruments include classification trees, data science, the study of clusters, and laws of correlation, which can produce statistical fraudulent methodologies. Some of the existing algorithms like K-Nearest Neighbors, Naive Bayes, Support Vector Machine, Decision Tree, and so on. There are a few inconveniences in the usage of the KNN Algorithm, asit is always known as lazy algorithmic learning and it relies on the performance of the dataset. Healthcare datasets are always massive, and the use of the KNN Model might be slow in the detection of fraudulent activities. Same to the KNN model, SVM, and Naive Bayes techniques would underperform when it comesto the analysis of large datasets.

V. PROPOSED SYSTEM

Scam is indeed a problem that several businesses are facing, particularly in the financial sector. Using some original information, such sectors must continuously forecast so that the fraudulent allegations could be recognized and then focused on.

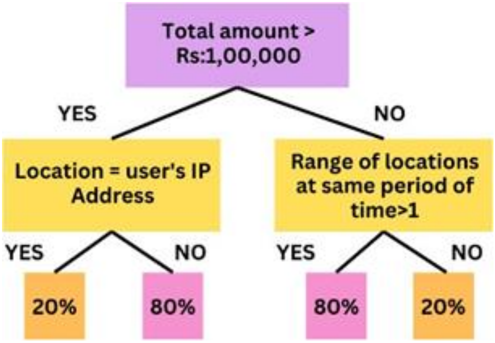
Total amount > Rs:1,00,000
YES
NO
Location = user's IP Address
Range of locations at same period of time>1
YES
NO
YES
NO
20%
80%
80%
20%

Fig.II Fraud Detection using Decision Tree

Researchers will detect reports which come as actual data to assess if they are difficult to disprove, costing the health insurer a lot of money. Since the Support vector machine is much more complicated than those the Decision tree, it is even harder to grasp or analyze. Fig 2 is an example of the Decision Tree method in Fraud Detection in the medicinal field. The proposed model shows that the decision tree algorithm and Logistic Regression Model are more suitable for fraud detection in the medical sector.

VI. METHODOLOGY

Datasets are the most important feature of fraud detection in healthcare. Most predictive models depend on the performance of information. The datasets may have different features, labels, specialties, and distinct providers. We must focus mainly on the features of fraud labels. Data Transformation is the next phase after the selection ofdatasets. At this phase, all the missing values and unwanted data could be modified or removed. Analysis of the best predictive model from different machine learning algorithms.

VII. TECHNOLOGIES TO OVERCOME CLINICAL FRAUDULENT ACTIVITIES

Data from the healthcare sector is thought to be very valuable. This has grown to be a significant allure forthe theft and misuse of medical data. In order to address this anomaly, the current study uses time series analysis to investigate the pattern of healthcare data breaches and their cost. It does this by using the simple moving average approach and the simple exponential soothing method. The simple moving average method produced more accurate forecasting results than the other two methods.

The process of data analysis is done by collecting data from various sources and tabulated. Different kinds of patterns are derived from this data using the total, percentage, and average methods. These patterns will assist us in better comprehending the causes and effects of healthcare data breaches, therise and fall of data breaches, the behavior of various forms of attacks, and other crucial issues covered in the analysis section of this paper. For the purpose of forecasting healthcare data breaches, a time series analysis is done.

A. Challenges: Health insurance companies arebeing burdened by complex auditing practices and asset assessments, creating stress on the sector. Healthcare's accounting and financial elements are now low in creativity, relying too heavily on paperwork and inefficient procedures. The challenges ofthe financial health care system make it vulnerable to faults that are often hard to detect by the naked eye. There is a greater vulnerability to fraud with the advent of online purchases and multi-device connectivity. In conventional accountingsystems, with strict structures, clients are deeply frustrated with services. Medical conditions are comparatively strong due to the nature of the business, and mistakes can also have significant consequences for the company.
B. Solutions: Fraudsters frequently exhibit observable behaviors that suggest their intention to deceive. Monitoring and responding to staff might benefit the detection of possible forgery. The cyber-riskpolicies, including varieties of deception andthe penalties connected to it, must be known by almost everyone in the medical industry aswell as to the patients. People attempting to commit fraud would be conscious when the administration is monitoring and, presumably, this would prevent them.

Internal management measures must be reviewedand updated on a regular basis to verify that they will be always beneficial and keep up with technology and many other advancements like Fraud detection using ML techniques, Minimization of breaches of data in the medical industry, Reduction of unwanted and fraudulent medical expenses and so on. When choosing auditors, fraudulent investigators, and other experts who have access to the relevant business information including such banking information, it is indeed vital for being assured such firms or individuals get a track record of providing outstanding service and also being trustworthy. Workplace theft might result in substantial losses, attorney's expenses, and tarnished identities, many of which can contribute to an organization's demise. Providing the correct measures is essential and could help organizations avoid illegal transactions or decrease the losses if one has actually happened.

Table.I - Accuracy of different ML Algorithms

Algorithms	Accuracy
Logistic Regression	0.87
Decision Tree	0.83
K-Nearest Neighbor	0.78
Support Vector Machine	0.73
Naive Bayes	0.71

VIII. BIG DATA ANALYTICAL TOOLS FOR FRAUD DETECTION

dramatically enhance the performance of treatment they deliver through computer analytics, artificial intelligence, and deep learning by expanding accessibility for critical and preventive services instead of spending time and energy finding for corrupt individuals. Table 1 shows different accuracy levels of various algorithms, from which the Logistic Regression and Decision Tree algorithm shows the highest accuracies among all the other algorithms.

Result - Decision Tree algorithms for fraud protectionare being used because there is a need for the authorized person to identify suspicious behaviours ina money transfer.

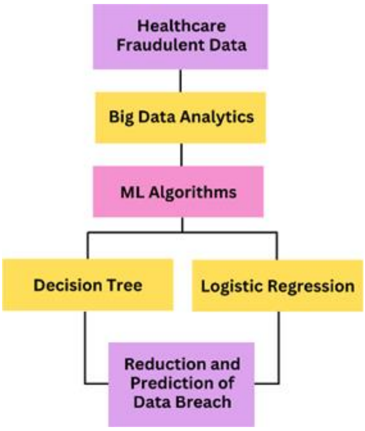
Healthcare Fraudulent Data
Big Data Analytics
ML Algorithms
Decision Tree
Logistic Regression
Reduction and Prediction of Data Breach

Fig.III Framework for Fraud Detection

Such testing methodology of specifications for categorizing fraudulent activities focused mostly on the dataset. Logistic Regression is a method in classification problems used whenever the judgment is unambiguous. Fig.3 shows the fraud detection framework using the decision tree method and logistic Making sense of data that has been mined, both structured and unstructured, is the core task of big data analytics. Effective anti-breach security solutions are being created by examining data from various virus sources, infiltration patterns, and fraudulent methods used to steal valuable information. To detect anomalies and suspicious activity from intruders, big data analytics is applied. More effective solutions for cybersecurity breach protection can be developed by researching the methods used by cybercriminals to enter networks. Doctors and clinical experts will regression model. It implies that when a scenario happens, that outcome will be 'corruption' or 'anti-corruption'. Immediate identification and prevention of corruption and theft in the medical industry can assist in the restoration of a huge amount of money.

CONCLUSION

Healthcare data breach and fraudulent activities is now the biggest problem in considerably higher medical spending and costs, specifically for senior citizens. Ensuring good health, reducing abuse, and restoring expenses seem to be of great concern. We also discussed a few forms of corruption and their ramifications. Healthcare fraud detection technology examines numerous money transfers and other analyses to predict typical behavioural patterns and spot possible fraudulent activity using ML algorithms. According to the proposed model, the algorithms like a decision tree and logistic regression are the most suitable algorithms of big data techniques for predicting crimes in the healthcare sector.

REFERENCES

[1] Andi Yaumil Bay R. Thaifur a,b,∗, M. Alimin Maidinc, Andi Indahwaty Sidinc, Amran Razak d(2021),” How to detect healthcare fraud? “A systematic review” Published by Elsevier Espana, ˜ S.L.U.

[2] Zarour, M., et al.: Ensuring data integrity of healthcare information in the era of digital health. Healthc. Technol. Lett. 8, 66–77 (2021)

[3] K. Chitra Lekha Dr. S. Prakasam, 2018 “IMPLEMENTATION OF DATA MINING TECHNIQUES FOR CYBER CRIME”, InternationalJournal of Engineering, Science and Mathematics

[4] Matthew Herland*, Taghi M. Khoshgoftaar and Richard A. Bauder, 2018, “Big Data fraud detection using multiple medicare data sources” Journal of Big data

[5] Iqbal H. Sarker1,2, 2019, “Data Science and Analytics: An Overview from Data-Driven Smart Computing, Decision-Making and Applications Perspective” Springer Nature Singapore Pte Ltd 2021

[6] Konasani, V R , M Biswas, P K Keloth, 2012, ‘Healthcare Fraud Management using Big Data Analytics’, Trendwise Software Solutions LLP, pp 1:5.

[7] Manogaran, G., Thota, C., Lopez, D. and Sundarasekar, R. (2017). Big Data Security Intelligence for Healthcare Industry 4.0. Springer Series in Advanced Manufacturing, pp.103–126.

[8] Dash, S., Shakyawar, S.K., Sharma, M. and Kaushik, S. (2019). Big data in healthcare: management, analysis and future prospects. Journal ofBig Data, 6(1).

[9] Abouelmehdi, K., Beni-Hessane, A. and Khaloufi, H. (2018), Big healthcare data: preserving security and privacy. Journal of Big Data, 5(1).

[10] Ristevski, B. and Chen, M. (2018). Big Data Analytics in Medicine and Healthcare. Journal of Integrative Bioinformatics, 15(3).

[11] Wang, L. and Alexander, C.A. (2019). Big Data Analytics in Healthcare Systems. *International Journal of Mathematical, Engineering and Management Sciences*, 4(1), pp.17–26.

[12] Nazir, S., Khan, S., Khan, H.U., Ali, S., Garcia-Magarino, I., Atan, R.B. and Nawaz, M. (2020). A Comprehensive Analysis of Healthcare Big Data Management, Analytics and Scientific Programming. *IEEE Access*, 8, pp.95714–95733.

[13] Waghade, S S, A M Karandikar, (2018) A Comprehensive Study of Healthcare Fraud Detection based on Machine Learning”, International Journal of Applied Engineering Research, 13(6), pp. 4175-4178.

[14] Georgakopoulos, S.V., Gallos, P. and Plagianakos, V.P. (2020). Using Big Data Analytics to Detect Fraud in Healthcare Provision. 2020 IEEE 5th Middle East and Africa Conference on Biomedical Engineering (MECBME).

[15] Herland, M., Khoshgoftaar, T.M. and Bauder, R.A. (2018). Big Data fraud detection using multiple medicare data sources. Journal of Big Data, 5(1).

[16] Bauder, R. and Khoshgoftaar, T. (2018). Medicare Fraud Detection Using Random Forest with Class Imbalanced Big Data. 2018 IEEE International Conference on Information Reuse and Integration (IRI).

[17] Rao, K. & Lakshmi, D. & P N, Jyothi. (2021). Performance on Fraud Detection in Medical Claims ofHealthcare Data. International Journal of Innovative Technology and Exploring Engineering (IJITEE), 8(7), pp. 1158:1165

[18] Rawte, V. and Anuradha, G. (2015). Fraud detection in health insurance using data mining techniques. 2015 International Conference on Communication, Information & Computing Technology (ICCICT).