Journal of Advances and Scholarly Researches in Allied Education
Vol. 21, Issue No. 5, July-2024, ISSN 2230-7540

DOI: https://doi.org/10.29070/vts5p257

Check for updates

Performance Analysis of the Platfora Method for Privacy Preservation in Large Healthcare Datasets: An Empirical Study

Farendrakumar Shrawan Ghodichor [1] * , Dr. Suraj Vishwanath Pote [2]

1. Research Scholar, Department of Computer Science & Engineering, University of Technology, Jaipur, Rajasthan, India
farendrakumarghodichor@gmail.com ,
2. Professor, Department of Computer Science & Engineering, University of Technology, Jaipur, Rajasthan, India

Abstract: Multimedia files and private medical information are among the many kinds of data stored in these systems. Data mining offers a promising method for many businesses and organizations to sift through massive data stores in search of useful insights. There are legitimate worries about privacy invasion and data abuse associated with extracting sensitive information from these databases. Personal information such as names, ages, residences, and phone numbers is included in healthcare data, along with sensitive details such the names and characteristics of diseases. Improper handling of this information might lead to its misuse for personal advantage. Hence, it is crucial to conceal data from enemies before sharing it with outside parties. One new approach that promises to solve these privacy problems is privacy-preserving data mining, or PPDM. Protecting private data from unauthorized third-party suppliers is the primary objective of PPDM. Various methods for protecting personal information have been proposed in this study. To conduct the experiment, various patient pilot datasets were considered. We apply a JAYA-based genetic algorithm to conceal information in our contribution.

Keywords: databases, Preserving Data, datasets, digitalization, Health care

X

INTRODUCTION

A great deal of personally identifiable information (PII) and sensitive patient data is handled by the healthcare business. Due to growing data quantities and digitization, the integrity and security of patient data are of the utmost importance. To protect sensitive medical information and forestall data breaches, synthetic data has recently surfaced as a potential alternative. But privacy considerations and national rules prohibit direct access to individual health data.

Analyzing massive amounts of data gathered from a variety of sources, such as research institutes, government organizations, insurance companies, pharmacies, and health care providers, is a common component of health care research projects. Information security in healthcare is very concerned with patient privacy because of the delicate nature of medical records and the social and legal ramifications of their exposure.

Research involving clinical or health services makes the need to protect the privacy of patients' personally identifiable health information all the more paramount. Clinical researchers who are not considered covered entities are expressly forbidden from receiving personally identifiable health information according to the

HIPAA privacy rule. Health plans, healthcare clearinghouses, and healthcare providers that electronically communicate health information in conjunction with specific designated HIPAA transactions, such claims or eligibility queries, are considered covered entities according to this privacy rule.

Researchers often get de-identified or anonymised health data in order to prevent the dissemination of personally identifying information. Multiple sources with varying policies on the release of protected health information may provide this data. Without identifying information, it is difficult to correlate and integrate health records on an individual patient basis in a way that protects their privacy. Take, for instance, these questions that pertain to an investigation of antidepressants:

Equally important is ensuring that all patient information remains private and uncompromised. It has come to light that synthetic data might be the key to unlocking this mystery. Synthetic data allows businesses to simulate genuine patient data without revealing any personal information by creating realistic but fictional datasets. Thus, businesses may safeguard patient privacy and confidentiality while developing algorithms and software by not exposing patient data.

LITERATURE REVIEW

Sascha Welten et al. (2022). Prior knowledge As the amount of health care data continues to rise at an exponential rate, data-driven medicine is becoming more important for diagnosis, treatment, and research. Unfortunately, due to concerns about privacy issues, such as the unintentional exposure of data to other parties, data protection rules forbid data centralization for analytical reasons. So, it's important to pay close attention to alternative data use practices that are compliant with current privacy rules. Objective Our goal is to implement a paradigm shift in data analytics by creating a system that can handle sensitive patient information while still meeting the requirements of local data protection laws. This system will be dubbed the Personal Health Train (PHT). According to PHT's central premise, data instances should stay put while the analytical work is moved to the data provider. We describe our implementation of the PHT paradigm in this paper. It works with a limited number of communications channels and respects the sovereignty and autonomy of the data sources. In addition, we run a DA use case using data from three separate, geographically dispersed data suppliers. Final Product We demonstrate that data model training with dispersed data sources is possible with our technology. In order to reduce regulatory hurdles to patient data exchange, our study showcases the possibilities of DA infrastructures in the healthcare industry. We further show that it can power medical research by facilitating access to dispersed data sets for researchers and medical professionals.

Rajan. V S, Dr. (2015). With Platfora, business users and data analysts can quickly and easily see massive amounts of data, allowing them to deal with machine data, user interactions, and even the most basic types of transactions. At its core, big data retrieval is a cooperative-based database caching system designed to handle massive datasets and intercept requests sent to the database. To access data from a cache, indices are nodes that have previously requested a cache in a query. To speed up the retrieval of requested data from the cloud's distributed file system, caching mechanisms and external databases are put into place. Hadoop, the current technology, has a number of constraints, including a relatively low-level implementation for requirements analysis (such as map reduction) and a high amount of developer expertise required to run various PLATFORA. Anyone may reduce and manage PLATFORA-related datasets stored

in Hadoop by taking use of an abstraction layer that is automatically created when users submit queries. Users are able to take use of a caching decision mechanism for content and retrieve it from massive databases by analyzing the Hamlet framework. Using sky tree as an example, this article examines a data analytics platform and language for machine learning that is specifically designed to deal with Big Data.

Hamza, Rafik et.al. (2022). People may see the world in a whole new light with the help of augmented reality (AR) technology. In its most basic form, augmented reality (AR) makes use of computer-generated images to superimpose enhancements onto the user's view of reality via the use of machine learning and big data. In this article, we provide a thorough overview of the current state of privacy-preserving big data analytics for 5G-era machine learning's new frontiers. For applications based on 5G-based wearable devices, we emphasize the essential qualities of privacy and security. After that, we provide the results of a research on ML methods that protect users' privacy. At last, we go over the specifics of how a safe prediction service based on a Convolutional Neural Network (CNN) model and the CKKS homomorphic encryption method was put into production. In the age of 5G and beyond, we want to foresee how security technologies may be used to ensure the secure processing of data acquired from wearable devices. This is especially true when it comes to improving the efficiency and effectiveness of privacy-preserving machine learning analytics.

Safaei Yaraziz, Mahdi et.al. (2022). The IoT, or Internet of Things, is a smart system that automatically sets itself up by connecting autonomous objects to the web and allowing them to interact with one another. Concerns about privacy may arise due to the fact that "things" are autonomous. This research provides an overview of Internet of Things (IoT) systems, privacy and security protocols, and related topics, such as (a) methods for protecting personal information in IoT-based systems, (b) current privacy solutions, and (c) suggested privacy models for various IoT application layers. Our study's findings show that contemporary approaches to data security and privacy may benefit substantially from innovations like Blockchain, Machine Learning, Data Minimization, and Data Encryption. In addition, reducing the amount of data collected, stored, and shared by smart devices makes perfect sense from a user privacy perspective. Consequently, our research suggests a data minimization approach based on machine learning that, when applied to these networks, may greatly enhance privacy protection.

Zhang, Denghui et.al. (2023). This study delves further than just conventional remote sensing methods to explain the various data collecting platforms that have emerged as a result of the IoT. In the business sector, deep integration of computer vision and remote sensing has arrived, allowing for the application of AI to issues like autonomous data extraction and picture interpretation. There is no uniform security protection mechanism in place, and the complicated architecture of the Internet of Things makes distant sensing devices susceptible to privacy breaches while exchanging data. Due to the fact that conventional encryption techniques rely on computational complexity, it is imperative to develop a security system that is appropriate for computation-limited devices in the Internet of Things environment. When encrypted pictures are superimposed over one another, the human visual system is able to immediately decipher them, a phenomenon known as visual cryptography (VC). Ideal for privacy-preserving detection in large-scale remote sensing photos in the IoT, VC has the stacking-to-see feature and uses a simple Boolean decryption process. This work successfully accomplishes the safe and effective transfer of high-resolution remote sensing pictures using meaningful VC. The loss of quality in recovery photos is reduced by spreading the

mistake between the encryption block and the original block across neighboring blocks. Small encrypted datasets for remote sensing pictures may have their identification performance improved by fine-tuning the pre-trained model using large-scale datasets. The suggested lightweight privacy-preserving recognition framework enhances security while maintaining good recognition performance, according to the experimental findings.

RESEARCH METHODOLOGY

Method

The presentation will center on the approach that was used to get the outcomes of the thesis. The article will explain how an implementation was built, including the choices made for the methods used to produce synthetic data. This section will also detail the methods used to collect information and materials. As a first step in the deployment, the described technology will be used to make synthetic data generation easier. The "input real data" from the provided dataset will form the basis of the implementation, as shown in the first phase of the figure.1. The second step is to train the generator using a plugin. In this case, we'll be using GAN as our generator and two algorithms from that framework. Step three is to generate the synthetic data. Step four is to evaluate the synthetic data using the metrics chosen in step 3.5. The measures indicated in 3.5 will be used to analyze the synthetic data that is created. Protecting confidentiality and authenticity during data transfers, synthetic data acts as a safe framework for exchanging information. As previously stated, it enables sharing without compromising individual privacy or releasing personally identifiable data by creating randomized data thatmaintains statistical features and linkages while hiding sensitive information.

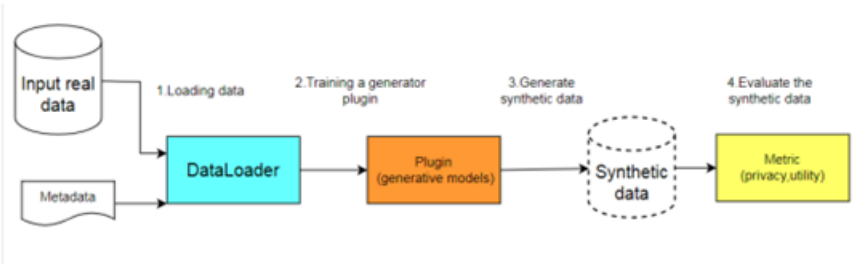

Figure. 1: overview on how the implementation functions

Tool

You may create synthetic data using one of three distinct tools: SynthCity, Data Synthesizer, or SynthPop. There are advantages and disadvantages to each of these instruments. When it comes to creating and evaluating synthetic data, SynthCity offers a more comprehensive option than Data Synthesizer and Synthpop. In addition to processing many tabular data formats, it provides a consistent interface for loading a wide variety of input data. For example, you may automate and streamline operations by using the library's utility functions to compare and contrast different data producers.

DATA ANALYSIS

This article showcases the outcomes of this thesis study, which primarily centers on the assessment of two

generative adversarial network GAN designs: DPGAN and CTGAN. Detailed below are the steps used to create the DPGAN and CTGAN pipelines, as well as the assessment approaches used to see how well they worked.

If the datasets have different distributions, then the stats.inv_kl_divergence. marginal value will be zero, and if they have the same distribution, then the value will be one. This statistic is the mean inverse of the Kullback-Leibler Divergence. Table 1 shows that for different values of epsilon, there is a significant amount of inverse KL divergence. The Kolmogorov-Smirnov test, which is part of stats.ks_test.marginal, takes on a value of zero when the distributions are totally different and a value of one when they are identical. With the numbers provided in table1, it is clear that the distributions of the synthetic data and the genuine data are quite close.

On average, the findings show that there are about 176 records in the ground truth dataset with distinct epsilon values that are indistinguishable from each other.

With a minimum value of 51 when epsilon is set to 10, the produced synthetic data falls short of attaining the same degree of indistinguishability as the ground truth. At the most stringent privacy level (epsilon 0.1), there is a significant amount of ambiguity in the findings, as seen in table1

Table 1: evaluation results using differential privacy with various epsilon values

Epsilon	0.1	5	10
privacy.k-anonymization.gt	176.0 +/- 0.0	176.0 +/- 0.0	176.0 +/- 0.0
privacy.k-anonymization.syn	78.5 +/- 29.5	71.0 +/- 3.0	51.5 +/- 11.5
privacy.distinct l-diversity.gt	0.0 +/- 0.0	0.0 +/- 0.0	0.0 +/- 0.0
privacy.distinct l-diversity.syn	0.001 +/- 0.0	0.001 +/- 0.0	0.001 +/- 0.0
stats.inv_kl_divergence.marginal	0.914 +/- 0.017	0.959 +/- 0.01	0.866 +/- 0.003
stats.ks_test.marginal	0.879 +/- 0.037	0.873 +/- 0.034	0.801+/-0.001

Here, we compared four distinct algorithms: Jaya-GA, Jaya-SSO, JayaGWO, and Jaya-PSO. See how several evolutionary algorithms do when comparing privacy preservation ratio and iteration count in Fig. 2. The suggested Jaya-GA model conceals more sensitive data than existing models, according to the model's behavior.

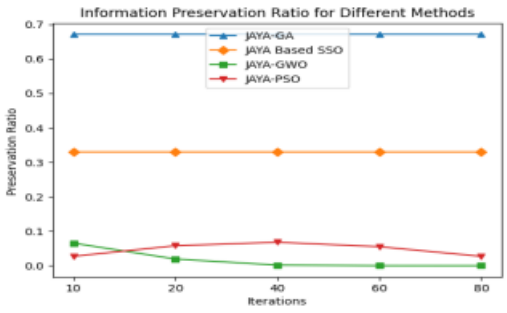

Figure 2: Simulation Output

Performance Analysis of Privacy Preservation for different values of ε

The performance analysis of privacy preservation for various values of ε has been carried out in this section. Reducing the value of ε results in excellent privacy, according to the simulation result. The performance metrics of accuracy, precision, recall, and F1-score for various values of ε before and after privacy protection are shown in Fig. 3.

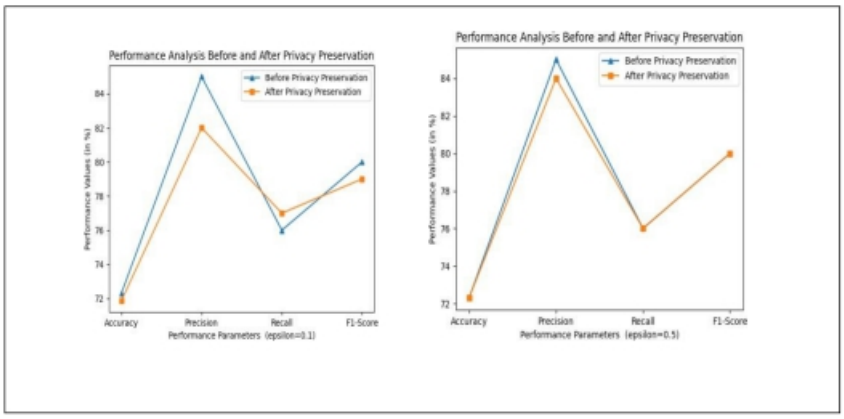

Figure 3: The change in performance for different values of ε.

CONCLUSION

Synthetic data has arisen as a means to safeguard sensitive patient information and forestall data breaches in the healthcare sector. However, direct access to health data is limited due to privacy concerns and restrictions. Using a known dataset, the author of this bachelor's thesis developed a secure data sharing infrastructure and then used a machine learning application to prove its efficiency.These results add to what is already known about how to share data while protecting individuals' privacy, which is especially important in the healthcare industry. The results of the study shed light on the potential pitfalls and opportunities of synthetic data production as it relates to the protection of personal information. The assessment of GAN algorithms and the establishment of a safe data sharing infrastructure provide the groundwork for more studies and better approaches to protecting users' privacy while exchanging information online. In sum, the findings of this study provide light on the efficacy of various solutions and deepen our comprehension of the privacy needs associated with data sharing, opening the door to more developments in this area.

References

1. Welten, Sascha & Mou, Yongli & Neumann, Laurenz & Jaberansary, Mehrshad & Ucer, Yeliz & Kirsten, Toralf & Decker, Stefan & Beyan, Oya. (2022). A Privacy-Preserving Distributed Analytics Platform for Health Care Data. Methods of Information in Medicine. 61. 10.1055/s-0041-1740564.

2. Rajan. V S, Dr. (2015). Platfora Method for High Data Delivery in Large Datasets. Indian Journal of Science and Technology. 8. 10.17485/ijst/2015/v8i33/7651.

3. Hamza, Rafik & Dao, Minh. (2022). Research on privacy-preserving techniques in the era of the 5G applications. Virtual Reality & Intelligent Hardware. 4. 210-222. 10.1016/j.vrih.2022.01.007.

4. Safaei Yaraziz, Mahdi & Jalili, Ahmad & Gheisari, Mehdi & Liu, Yang. (2022). Recent trends towards privacy-preservation in Internet of Things, its challenges and future directions. IET Circuits, Devices & Systems. 17. n/a-n/a. 10.1049/cds2.12138.

5. Zhang, Denghui & Shafiq, Muhammad & Wang, Liguo & Srivastava, Gautam & Yin, Shoulin. (2023). Privacy-preserving remote sensing images recognition based on limited visual cryptography. CAAI Transactions on Intelligence Technology. 8. n/a-n/a. 10.1049/cit2.12164.

6. Ambulkar, B., & Borkar, V. (2012, April). Data mining in cloud computing. In MPGI National Multi Conference (Vol. 2012). Academic Press.

7. Anjum, A., Ahmed, T., Khan, A., Ahmad, N., Ahmad, M., Asif, M., Reddy, A. G., Saba, T., & Farooq, N. (2018). Privacy preserving data by conceptualizing smart cities using MIDR-Angelization. Sustainable Cities and Society, 40, 326–334. doi:10.1016/j.scs.2018.04.014

8. Arthur, D., & Vassilvitskii, S. (2006c, October). Worst-case and smoothed analysis of the ICP algorithm, with an application to the k-means method. In 2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06) (pp. 153-164). IEEE. doi:10.1109/FOCS.2006.79

9. Chamikara, M. A. P., Bertók, P., Liu, D., Camtepe, S., & Khalil, I. (2018). Efficient data perturbation for privacy preserving and accurate data stream mining. Pervasive and Mobile Computing, 48, 1–19. doi:10.1016/j. pmcj.2018.05.003

10. Kikuchi, H., Hamanaga, C., Yasunaga, H., Matsui, H., Hashimoto, H., & Fan, C. I. (2018). Privacy-preserving multiple linear regression of vertically partitioned real medical datasets. Journal of Information Processing, 26(0), 638–647. doi:10.2197/ipsjjip.26.638

11. Komishani, E. G., Abadi, M., & Deldar, F. (2016). PPTD: Preserving personalized privacy in trajectory data publishing by sensitive attribute generalization and trajectory local suppression. Knowledge-Based Systems, 94, 43–59. doi: 10.1016/j.knosys.2015.11.007

12. Li, L., Lu, R., Choo, K. K. R., Datta, A., & Shao, J. (2016). Privacy-preserving-outsourced association rule mining on vertically partitioned databases. IEEE Transactions on Information Forensics and Security, 11(8), 1847–1861. doi:10.1109/TIFS.2016.2561241

13. Li, Y., Jiang, Z. L., Yao, L., Wang, X., Yiu, S. M., & Huang, Z. (2019). Outsourced privacy-preserving C4. 5 decision tree algorithms over horizontally and vertically partitioned dataset among multiple parties. Cluster Computing, 22(1), 1581–1593. doi:10.1007/s10586-017-1019-9

14. Lin, J. C. W., Wu, T. Y., Fournier-Viger, P., Lin, G., Zhan, J., & Voznak, M. (2016). Fast algorithms for hiding sensitive high-utility itemsets in privacy-preserving utility mining. Engineering Applications of Artificial Intelligence, 55, 269–284. doi: 10.1016/j.engappai.2016.07.003

15. Upadhyay, S., Sharma, C., Sharma, P., Bharadwaj, P., & Seeja, K. R. (2018). Privacy preserving data

mining with 3-D rotation transformation. Journal of King Saud University-Computer and Information Sciences, 30(4), 524–530. doi: 10.1016/j.jksuci.2016.11.009