Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Boosting Privacy of Data While Applying with Machine Learning Algorithms

Arpita Maheriya, Shailesh Panchal
Research Scholar, Gujarat Technology University,
arpita_maheriya@gtu.edu.in Orchid Id:0000-0001-5703-4410
Professor, Gujarat Technology University-Graduate School of Engineering Technological, sdpanchal@gtu.edu.in OrchidId:0000-0003-4597-4063
Gujarat Technological University – Graduate School of Engineering and Technology Academic Block 5, GTU-Chandkheda Campus, Nr. Vishwakarma Government Engineering College, Nr. Visat Three Roads, Visat – Gandhinagar Highway, Chandkheda, Ahmedabad–382424–Gujarat

ARTICLE INFO

Received: 20 Sep 2024

Accepted: 24 Dec 2024

ABSTRACT

Machine learning (ML) models are widely used in various domains, such as healthcare, finance, and social media. However, ML models may also pose privacy risks, as they can reveal sensitive information about the training data or the users who interact with them. To protect the privacy of data and users, several techniques have been proposed, such as anonymization and differential privacy. Anonymization aims to remove or modify the identifying attributes of the data, such as names, addresses, or phone numbers. Differential privacy adds random noise to the data or the model outputs, such that the presence or absence of any individual in the data does not affect the results significantly. However, both techniques have limitations and challenges, such as information loss, utility degradation, or high computational cost. In this paper, we propose a novel hybrid algorithm that combines anonymization and differential privacy to enhance the security of data and ML models. Our algorithm applies k-anonymity with n-gram to the data before sending it for training with ML model, which further processed with differential privacy. Differential privacy allows performing computations on encrypted data without decrypting it, while blowfish encryption is a fast and secure symmetric-key algorithm. Our algorithm ensures that the data and the model are protected from unauthorized access or modification by the malicious third parties. We evaluate our algorithm on several benchmark datasets and ML models, and show that it achieves high accuracy and privacy while reducing the communication and computation overhead. We also compare our algorithm with existing methods and demonstrate its advantages and limitations.

Keywords: Anonymization techniques. ML model security. Data protection. Privacy of electronic healthcare system. Differential Privacy.

1. INTRODUCTION:

The healthcare sector has become a focal point of interest for academicians, industry professionals, healthcare organizations, companies, and individuals seeking to harness personal data mining [44]. In recent times, the advancement of technology and the growth of healthcare industries have generated vast amounts of electronic medical and patient data. The emergence of digital medical systems and technologies has brought about significant improvements such as decision support systems, enhanced treatment quality, image segmentation, digital communication systems, patient monitoring, and expert consultation based on personal e-healthcare data [5]. These advancements have revolutionized the quality and innovation of medical services. However, there has been ongoing debates and concerns about data privacy and ownership.Privacy-sensitive data can possess numerous significant and valid applications, serving diverse objectives. The recent Covid-19 outbreak has further underscored the urgent need to employ state-of-the-art technologies and develop strategic models for preserving privacy while sharing data in the healthcare domain. Nonetheless,determining the permissible uses of data while upholding security measures and respecting individuals' right to privacy is an exacting task.

In this paper, we briefly elaborateexisting methods for anonymization &encryption methods, techniques to apply ML model without decrypt the sensitive patient's information and securitychallengesof application of ML/DL in healthcare

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

along with possibleattacks on MLmodel. The aim of the paper to furnishmethodology to provide confidential and safe model for healthcare system.

The comparison of this paper with other survey and research paper is described in table 1.

Framework of the paper: rest of the paper is structured as follows. In section 2,brief description of background study with literaturereview.In section 3, described ml application in healthcare with its challenges and vulnerabilities in each stage of building model moreover possible ML model security and applicability challenges in healthcare system.In section 4. demonstrated existing model of privacy and security preserving techniques in healthcare system. In section 5, demonstrate implementationstrategies, proposed security model and algorithm. in section 6, experimental result with comparison and in last conclusion of the paper.

2. PROACTIVE STUDY OF PRIVACY PRESERVING METHODOLOGY WITH ML/DL

Efforts to develop efficient, confidential, robust, and safe ML-based healthcare applications are crucial. In a recent study [8], the author provides an overview of privacy-preserving methods, ranging from traditional approaches to federated learning, and discusses their applications in different scenarios. The practical feasibility of performing anonymization on E-Healthcare data is demonstrated based on an analysis of studies [6]. The survey concludes that, anonymization of healthcare data is achievable, but striking a balance between utility and privacy necessitates further research.In study [1], a novel framework is proposed for generating anonymized networks using data synthesis, promoting the use of open data and sharing while ensuring privacy. Figure 2 presents a technical perspective of Personal Data Privacy (PDP), depicting how the proposed model mitigates the risks of data breaches and re-identification of confidential patient data, thereby offering a secure, ethical, and legal solution for data sharing in support of advancements in AI technology in the medical field.An efficient and scalable data scrubbing method for anonymized data is proposed in [2], which incorporates measures for moderate data utility and feature diversity. The proposed approach [3] demonstrates high utility of data and is tested on a dataset comprising health and demographic surveillance system events (totaling 280,381) with demographic and time-varying data information.Moreover, a novel approach [6] is introduced for anonymized ML models, ensuring secure privacy during the training process. The study also explores the potential integration of privacy methods such as federated learning and homomorphic encryption with ML models.

Table 1. ComparisonOf Research Paper with Other Model Secure and PrivateML Robust Models on Sensitive Healthcare Data (Covered-Y, Not Covered-N, Partial-P, Research Paper-R, Survey Paper-S)

No. & Year	Type	Application of ML in Healthcare	Risk on Healthcare Sensitive Data Disclosers	Privacy Preserving Method of Health care sensitive Data/proposed	Attacks on ML	ML Model on Anonymized/Encrypted Data in Healthcare	Proposed method for Secure ML Model
[1] 2020	R	N	N	Y	N	Y	Y
[2] 2020	R	Y	N	N	N	Y	Y
[4] 2020	S	Y	N	Y	N	N	N
[23]2020	R	N	N	Y	N	N	N
[35] 2020	R	N	N	N	Y	N	N
[3] 2021	R	N	Y	Y	N	N	N
[5] 2021	S	N	Y	Y	N	Y	N
[6] 2021	R	Y	N	Y	Y	Y	Y
[7] 2021	S	N	Y	Y	N	N	N
[9]2022	R	Y	Y	Y	Y	Y	N

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

[10]2022	R	N	N	Y	N	N	N
[11]2022	R	Y	Y	N	Y	Y	N
ProposedResearch 2023	R	Y	Y	Y	Y	Y	Y

Following are the four major concerns on ethics and policy for Machine Learning in healthcare:

1. Data Quality- Quality of data is utmost matters, when it comes to build healthcare application.Lack of availability of data and bias datacan lead unfair outcomes and disparities in healthcare base ML application.Preserve accuracy of ML model directly impact on individuallife. It is crucial to establish rigorous data collection protocols, address biases in the data, and regularly evaluate and validate the performance of AI systems to mitigate potential biases and improve the overall quality of AI-driven healthcare applications.
2. Privacy- It is essential to safeguard individual privacy and protect sensitive health information from unauthorized access or disclosure. Sharing healthcare data without maintain privacy can lead to data disclosure, discrimination, cyber- theft, identity theft, financial fraud, or other malicious activities,harm of individual reputation.It will be directlyinfluencingtrust of individuals on healthcare organizations.
3. Data excess & commercialization- In era of digitalbio-medical data, data collected by the technicist of government agencies canre-aggregate/ exploited in network, non-commercial and commercial companies without due concern,autonomy of patient.
4. Patient's concern-Use of AI for healthcare industrywith ethical rules is paramount important between the data collectors and source persons.it is challenging task to inform to patient about each update and potential current use of data.

In this paper, ourfocus to solve the privacychallengesand provide protection against data disclosures.

3. PRIVACY AND CONFIDENTIALITY PRESERVING TECHNIQUES FOR HEALTHCARE SYSTEM

To reconcile the conflicting demands of ensuring privacy and facilitating legitimate use, various techniques have been suggested, encompassing the following discussed extensive techniques used for privacy preserving data publishing methods (PPDP) to preserve security of clinical data.

- Cryptographic Techniques: It is branch of cyber security techniques to protect data.These approaches are to securecommunication and only allow the selected authentic person to change and interpret. In short, sensitive data not get disclose There are three types of crypto technique -Symmetric, asymmetricand one way Hash method.Here we discussed existing methods use for public sharing data of healthcare [14]. In [24] surveyed techniques of cryptographywithits unique safety features.

Rivest, Shamir, Adleman (RSA): This method is purposed for securely sharing information with authentication. The telecare is also useful in current scenario condition of patient and doctor communication.In[12], proposedRSA method for tele media healthcare information system provide security against polynomial timeadversaryattack. If any end user shares the key with unauthorized user or it gain the key access somehow then it would leak all the information [13].

Advance Encryption Standard (AES):in AES, common key generatedto encrypt/decrypt data is shared between the end parties on securemedia channel.With use of this technique trust is developed between data controller and third party. But, after the completion of agreement of sharing key and unauthorized key access there is no more security. Moreover,brute force and Man in the Middle attack can possible [15].

Unlike AES (one single key) it uses private and public key concept with maintain the integrity of data.However, it takes too much exponential time consuming for time and space[16]. AES and RSA demised while on secure large amount of information [18].Other cryptographic techniques also we can apply to preserve security of information exchange such as hashing, Diffie -Hellman[21], DES [22],3 DES, RC4, SHA-256[50], and digital signature [23]. However, it might be tough to apply mathematical operation or machine model without decryption. Furthermore, sharing key in network

Research Article

increase the risk of information exposure.In the study [17] suggested advance crypto graphic techniques are also fails to achieve privacy and convince healthcare stockholders to adopt in daily uses.In addition, cryptography techniques not guaranteed to incorporate with ethical regularities policy for health care data publishing.

- Homomorphic encryption (HE): It converts data into encrypted form that can be useful same as original form.it able to apply mathematical and ML operation with balancing privacy [31]. Surveyed various homographic encryption tools and its comparison [32]in bigdata environment. Withal, [33] briefly explained SMC via HE. HE enables privately publish and train the ML/DL models. SEAL, HElib, Palisade are the most used opensource tool-kits to execute HE of different types.It has some drawbacks such as excessive computational cost & structural change and time consuming.
- Anonymization: It also known as de-identification.Anonymization emergedtool on healthcare data to preserve the re-identified data of individual patient /person by removing personal dataor adding noise in data.A comparative review [19] of anonymized techniques on graph-based data. Table 4 discriminate types of anonymized data. By applying anonymization information loss is possible,which make analysis process worthless. That put the healthcare specialist aim nonfulfilled clearly. In same time, not reducing data up to the level to preserving confidentiality generate threat of breaches. Such as any lab / clinical worker try to identified the person out of its curiosity. Our goal is to anonymized the data at the level that it preservesprivacy,utility,confidentiality and regulation of healthcare policies.As well, analysist canapply sate of art methods on anonymized data with preferrable accuracy results [20]. Figure demonstrated the five safe parameters to keep in mind while performing anonymization model on healthcare data. Amidst those approaches, anonymization techniques have emerged as a pivotal means of safeguarding privacy. In recent times, numerous anonymization methods have been introduced, each grounded in distinct privacy definitions existed foranonymized model:k-anonymity,l-diversity, t-closeness [25] and attribute centric [27,28].
- Differential privacy (DP): DP is one of the preferable models to preserve availability of data while providing privacy to sensitive data.A randomized algorithm R is considered (ε, δ)-DP. For any two adjacent databases D and D0 that differ in the data of only one user, and for all sets O of potential outputs, the following inequality holds:

$$Pr[R(D) \in O] \leq e^{\varepsilon} * Pr[R(D0) \in O] + \delta \quad (1)$$

To collect, combine and share the data with preserve privacy and ML techniques can easily aggregate with DP. We can differentiate private data with general basic information of patient using this approach. General information is not secure with DP, means if sensitive healthcare data covers in general information may get disclosed. It provides preservation against the inference attacks on sensitive data and ML model.it not provide full guarantee the sensitive information stays sensitive. Laplace mechanism to satisfy DP by including noise in data.

- Federated Learning (FL): This approach globally trains the ML model on distributed data. It allows organizations to work collaboratively without share raw data of their own [29]. Overcome the challenge of healthcare sensitive data sharing raw information sharing, expect that it shares the train model and end user can use it to train their local data. FL vulnerable with inference attacks. In [26], used DP on federated learning ML model to mitigate inference attack. In federated learning, we need to use robust encryption technique to protect system and privacy of user [44].To address the issue of potential information leakage in federated learning (FL), In [50], propose a novel framework called Noising before model Aggregation FL (NbAFL). By incorporating the concept of differential privacy (DP), framework aims to effectively prevent the disclosure of private information. This is achieved by adding artificial noises to the parameters at the clients' side before aggregation. By doing so, it enhances the privacy protection of FL, reducing the risk of private data exposure during the model training process
- Secure multiparty computation (SMC): It is one the type of privacy preserving bigdata protocols. This protocol computes ordinary function for multiple independent parties/analysists/data collectors to get insight of data and only output of the function is shared with outside user [30]. Analyst can perform operation on encrypted dataset. Major obstacle of SMC, not guarantee to overcome the data breaches. we can apply this protocol with cryptography /encryption models.

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

4. IMPLEMENTATION STRATEGIES

Proposed Approach to ApplyMachine Learning Towards Anonymization

Anonymization methodology is applied to execute Privacy Preserving Data Publishing with maintain privacy and utility trade-off. Once datagetsshared with third party organization, it is assigned to data analysis, researcher, and developers fordistinct analytic techniques such as data mining, machine learning, statistics, visualization etc to get more insight from data. Anonymized dataset has noisy / embedding data (such as "*"). Machine learning may not be able to recognize and createaccurate model on embedded/ anonymized data set. Our proposed model is more focused overcome the challenges related to accuracy of ML personality de-identified model.As we know anonymized dataset is de-identifiedand confidential so here our approach to generate synthetic dataset with use of anonymized data. Thebenefit of synthetic dataset is independent from original set while preserving the statistics values(similar matching values). Natural Language Processing (NLP) aiming for read, analyse, understand, interpret the text. With advancement of pre-trained & transfer learning model for NLP text generation is in trend in recent research.

N-gram technique is used over anonymized data to enhance data security. N-grams are essentially contiguous sequences of n items within a given text. By applying n-grams to anonymized data, we could extract patterns and relationships without directly revealing sensitive information. One way to utilize n-grams for data security is through the process of tokenization. This involves breaking down the anonymized data into smaller units, such as words or characters, and then generating n-grams from these units. By analysing the frequency and distribution of these n-grams, we can identify potential vulnerabilities or patterns that may pose a risk to data security.

Hyperparameter and Tuning section

First in start we identified the columns /variablesfrom dataset need to anonymized with suitable and appropriatetechniques for it. Based on datatypes of identified variables define and perform the techniques such as suppression, generalization, masking with identical threshold values for each parameter.

In second step select the signific value of k with considering privacy protection, identification risk, data utility,legal consideration. In purpose to apply differential privacy, we have to consider: epsilon Ɛ (quantifies the privacyguarantee), delta, LaPlace value, sensitivity, query compositions, utility and privacy budget (total privacy loss).sensitivity refers to how much a change in the underlying dataset can influence the outcome of a query. Let'sdenote two datasets, xa and xb, as representatives of all potential datasets in X.

$$\text{Sensitivity}=\text{MAX} \,||\, q(a) - q(b) \,||_1 \quad (2)$$

Equation utilizes the L1-norm distance, denoted as $\|\cdot\|_1$, to measure the dissimilarity between datasets that vary by, at most, one element. Epsilon is a metric to gauge the degree of privacy loss when there's a differential change in data, such as adding or removing a single entry. Smaller values of ε indicate a stronger level of privacy protection.

The Laplace distribution with these parameters is used to generate random noise that is then added to the query

result to ensure differential privacy. The specific range of truncation depends on the characteristics of the query.

The Laplace distribution is characterized by two parameters: location (μ) and scale (b). To create a truncated

Laplace distribution that is bounded within a specified range, with the following equation:

$$F(x \mid \mu, b) = (1 / (2b)) * \exp(-|x - \mu| / b) \quad (3)$$

$$b = \Delta f / \varepsilon \quad (4)$$

In last for ML model implementation on anonymized data, we have to keep the ratio of train and test data splitting

other hyper parameters have to be based on model we applied on dataset.

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

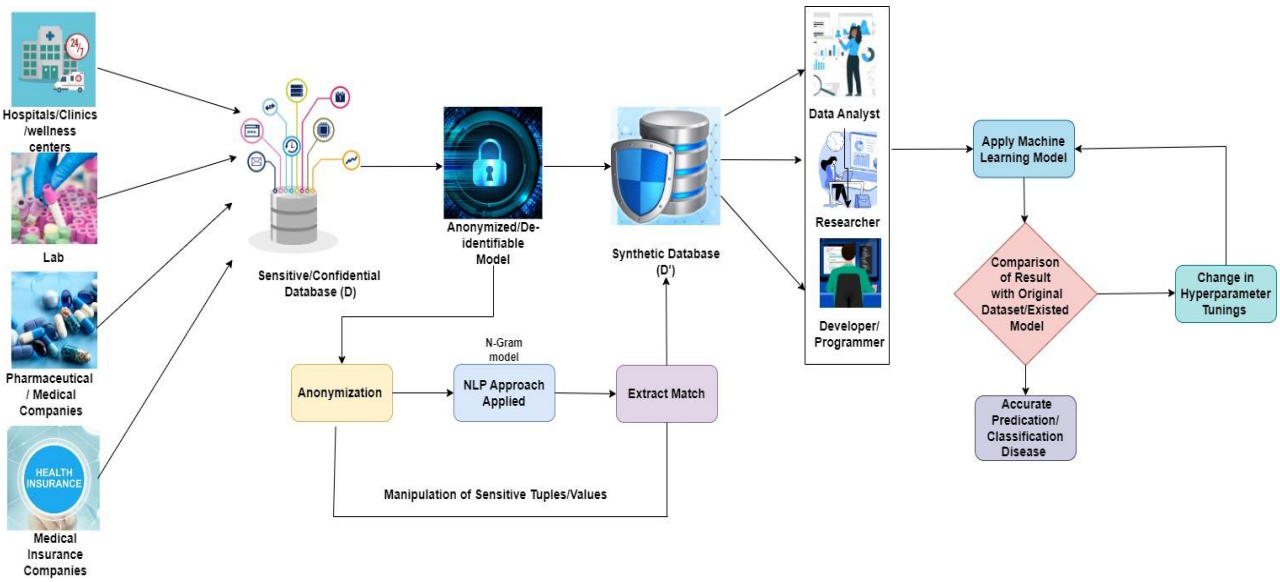

Fig 7. Proposed Model of Robust Healthcare ML Model

Using differential privacy (DP) in combination with k-anonymity can offer several benefits.:

1. Enhanced Privacy Protection- Adding differential privacy provides an extra layer of protection by introducing controlled noise or randomness into query responses, making it even more difficult to extract sensitive information.
2. Fine-Grained Privacy Control- privacy protection by tuning the privacy parameter ε. This enables organizations to strike a balance between privacy and data utility based on their specific needs.
3. Guaranteed Privacy Bounds- DP provides formal privacy guarantees. By setting ε and δ (privacy parameters) appropriately.
4. Utility-Preserving Data Sharing: K-anonymity can sometimes result in significant data distortion or utility loss, especially when k is set to a high value. By incorporating DP, it's possible to limit data distortion and maintain higher data utility while still achieving strong privacy guarantees.
5. Robustness Against Linkage Attacks- DP can protect against linkage attacks.
6. Research and Innovation: Privacy-preserving data analysis techniques, including the combination of k-anonymity and DP, enable researchers to access sensitive data for analysis without compromising privacy.
7. Improved Data Sharing: Privacy-preserving techniques like k-anonymity and DP can facilitate data sharing collaborations, as organizations can share data while minimizing the risk of data breaches and privacy violations.

Algorithm 1: Anonymization Technique

Input: Dataset D [quasi attribute-Q_i, sensitive attribute S_i, explicit identifiers E_i=no of attributes, privacy loss Ɛ, synthetic datasetSD,sensitivity S (Δ f),μ location parameter, b scale parameter]
Output: Anonymous dataset D`
Step 1: Attribute Classification (categorical/numerical)
Step 2: Identify Q_i and encode categorical data
Step 3: Delete E_i and label encoding perform
Step 4: Apply suppression /semi-suppression/ generalization on Q_i
Step 5: k-anonymization model implementation
Step 6:N-gram on anonymized data
Step 7: Laplace truncation (Ɛ=0.3, S=3)
Step 8: Apply epsilon (Ɛ) -differential privacy
Step 9: Plot risk of linkage attack, calculate data loss and error date

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Table 5. K-anonymized healthcare dataset

NO	Age	work class	education	educational-num	marital-status	occupation	relationship	gender	hours-per-week	native-country	income
0	8	Private	***h	7	Never-married	Machine-op-inspct	*	1	40	********ates	0
1	21	Private	*****ad	9	Married-civ-spouse	Farming-fishing	*	1	50	********ates	0
...	...	...	...	...	...	...	...	...	...	...	...
10990	46	Self-emp-not-inc	********lege	10	Married-civ-spouse	Sales	*	1	60	********ates	1
10991	21	Private	*****th	4	Married-civ-spouse	Craft-repair	*	1	40	********ates	0

Table 6. Differential Privacy with K-anonymity on Healthcare Dataset

No.	Age	workclass	education	educational-num	marital-status	occupation	relationship	gender	hours-per-week	native-country	income
0	8	Private	***h	7	Never-married	Machine-op-inspct	*	1	40	********ates	50K
1	21	Private	*****ad	9	Married-civ-spouse	Farming-fishing	*	1	50	********ates	18,3042
...	...	...	...	...	...	...	...	...	...	...	...
10990	46	Self-emp-not-inc	********lege	10	Married-civ-spouse	Sales	*	0	60	********ates	50K
10991	21	Private	*****th	4	Married-civ-spouse	Craft-repair	*	1	40	********ates	50K

Data utility typically involves measuring the usefulness or value of data while considering privacy or information loss. A basic equation to express this trade-off might be: Data Utility = Benefits - Loss of Privacy or Information.However, the exact equation for measuring data utility can vary widely depending on the specific application, data type, and the

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

metrics used to quantify benefits and loss of privacy or information. It often involves complex mathematical or statistical models that are tailored to the particular problem at hand using.

5. EXPERIMENT RESULTS

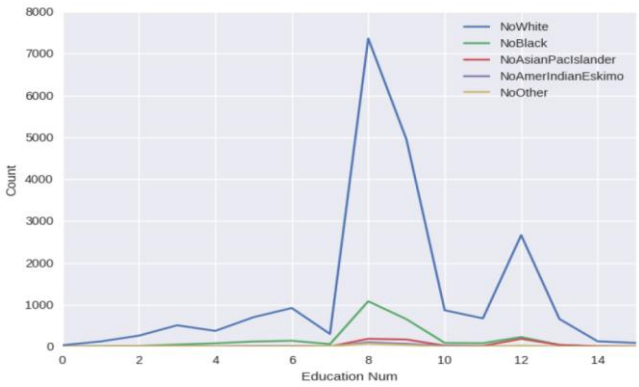

Fig 8. Impact of data after K-anonymization

The data sparsity is increased data sparsity, especially when the quasi-identifiers are highly generalized. Sparse data can make it difficult to perform certain statistical analyses or machine learning tasks. The process has a significant impact on data by reducing the granularity and detail of the information in order to protect individual privacy. While it is a valuable technique for privacy preservation, it requires thoughtful trade-offs between privacy and data utility, and it may not be suitable for all types of data analysis and applications.

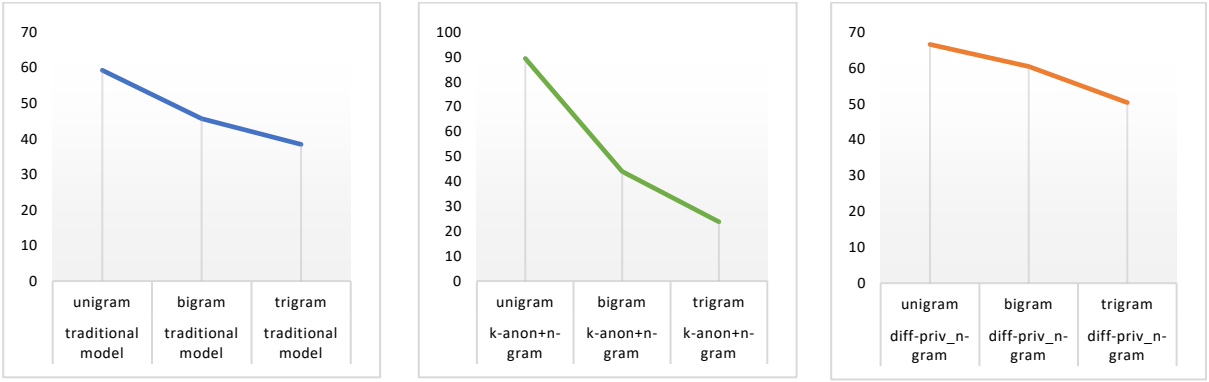

Fig.9Traditional,k-anonymity+n-gram and DF+n-gram

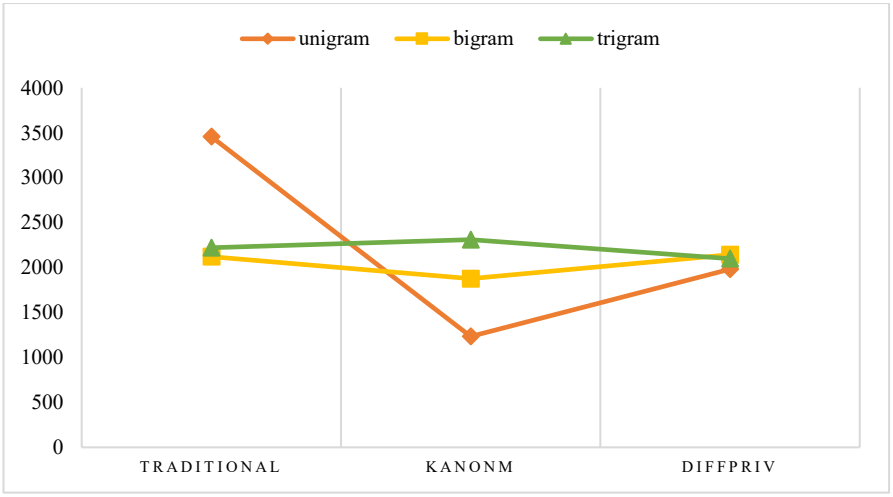

Fig 10. Impact of data after K-anonymization with DF

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Data loss is dependent on distribution of data with multiple factors. The distribution of income between the clusters is significantly different. The first cluster has a median income of $50,000, while the second cluster has a median income of $100,000. This difference is statistically significant, and it suggests that there is a relationship between income and the other variables that were used to create the clusters. One possible explanation for this difference is that the first cluster contains a higher proportion of people who are employed in low-paying jobs, while the second cluster contains a higher proportion of people who are employed in high-paying jobs

Data loss = (Original data - Anonymized data) / Original data (5)

(23456 - 18874) / 23456 = 0.198

Therefore, the data loss calculation over anonymized data over the adult dataset is 19.8%.

Data loss = (Original data - Anonymized +DP data) / Original data (6)

(23456 - 15901) / 23456 = 0.322

The data loss after the proposed approach applied is 32.2%. data loss in the anonymization process can also lead to a loss of trust in the data. If people believe that their data is not being protected, they may be less likely to share it, which could make it difficult to conduct research and develop new products and services. In conclusion, the ethical implications of data loss in the anonymization process are significant. It is important to be aware of these implications when making decisions about how to anonymize data.

MSE = (1/n) * Σ(actual – predicted)^2 (7)

Mean Squared Error for the adult dataset after processing the anonymization with k-anonymity and differential privacy is 0.0235.

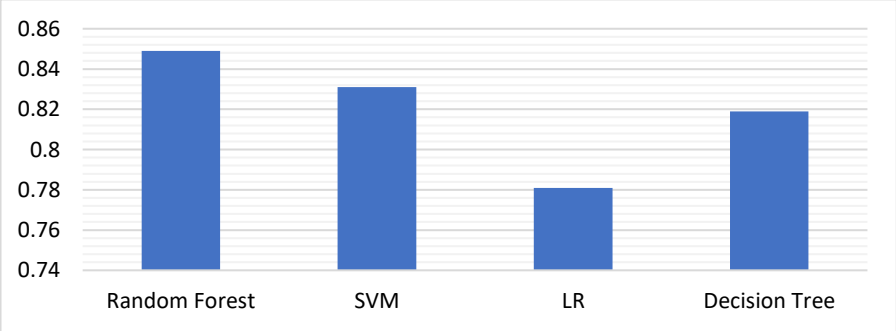

Fig 11. Accuracy comparison of ML algorithms

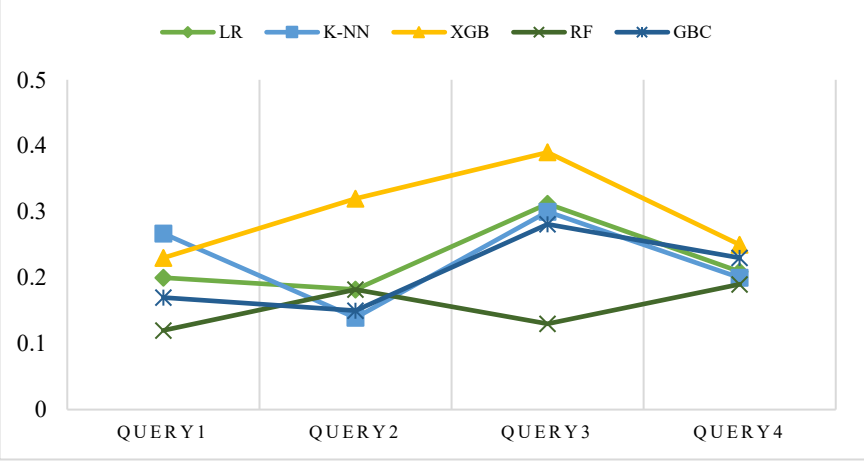

Fig 12. Failure Rate of ML Classification Models

Here, N-grams are used in conjunction with machine learning algorithms to build predictive model. By training these models on anonymized data and n-grams, we could identify unusual or suspicious patterns that may indicate a

Research Article

security breach or unauthorized access.While n-grams can enhance data security, they are used as part of a comprehensive security strategy that includes other measures such as encryption, access controls, and data anonymization techniques.

6. CONCLUSION

The security and privacy of healthcare data are of paramount importance in an era where digital health records are becoming the norm. This research paper has explored the application of K-anonymity, N-gram analysis, and differential privacy as mechanisms to safeguard sensitive healthcare data while enabling its utilization in machine learning classification algorithms. Through our investigation, several significant conclusions and contributions have been identified:

1. Privacy Preservation: K-anonymity has been demonstrated as an effective means of privacy preservation in healthcare data. By ensuring that each record in the dataset is indistinguishable from at least k - 1 other records, we mitigate the risk of re-identification and attribute disclosure.

2. Contextual Understanding: N-gram analysis has proven invaluable in enhancing the context-based understanding of healthcare data. By capturing the relationships between words or characters, it enables more accurate machine learning models that can classify medical records, detect anomalies, or predict patient outcomes.

3. Quantifiable Privacy Guarantees: Differential privacy, with its quantifiable privacy guarantees, allows for robust protection of individual data while enabling the application of machine learning algorithms. It empowers healthcare organizations to confidently leverage the advantages of data-driven insights without compromising privacy.

4. Enhanced Machine Learning: Combining K-anonymity, N-gram analysis, and differential privacy with machine learning classification algorithms offers an advanced approach to healthcare data security. The privacy-enhanced models enable more accurate predictions and classifications, thereby improving patient care, administrative efficiency, and healthcare research.

5. Ethical Considerations: Beyond the technical aspects, this research underscores the ethical obligation of healthcare organizations to protect patient data. The implementation of these privacy techniques not only aligns with regulatory compliance but also ensures the trust and confidence of patients, thus upholding the ethical principles of medical practice.

6. Future Directions: As healthcare data security continues to evolve, future research may explore more advanced privacy-preserving techniques, including homomorphic encryption, federated learning, and secure multi-party computation. Additionally, there is a need to further investigate the trade-offs between privacy preservation and data utility in healthcare analytics.

In conclusion, the research presented here emphasizes the critical role of privacy-enhancing technologies in the healthcare sector. The adoption of K-anonymity, N-gram analysis, and differential privacy, in conjunction with machine learning, offers a promising avenue to harness the full potential of healthcare data while respecting the confidentiality and privacy of patients. As the field of healthcare informatics progresses, the ethical and technical dimensions of data security will remain at the forefront of research and practice.

This conclusion encapsulates the research's key findings, contributions, and the broader implications of applying these privacy techniques to healthcare data. It leaves room for future research and underscores the importance of ethical considerations in healthcare data security.

REFERENCE

[1] Yoon, Jinsung, Lydia N. Drumright, and Mihaela Van Der Schaar. "Anonymization through data synthesis using generative adversarial networks (ads-gan)." *IEEE journal of biomedical and health informatics* 24, no. 8 (2020): 2378-2388.

[2] Song, Xing, Lemuel R. Waitman, Yong Hu, Bo Luo, Fengjun Li, and Mei Liu. "The impact of medical big data anonymization on early acute kidney injury risk prediction." *AMIA Summits on Translational Science Proceedings* 2020 (2020): 617.

Research Article

[3] Templ, Matthias, Chifundo Kanjala, and Inken Siems. "Privacy of study participants in open-access health and demographic surveillance system data: Requirements analysis for data anonymization." *JMIR Public Health and Surveillance* 8, no. 9 (2022): e34472.
[4] Vokinger, Kerstin N., Daniel J. Stekhoven, and Michael Krauthammer. "Lost in anonymization—A data anonymization reference classification merging legal and technical considerations." *Journal of Law, Medicine & Ethics* 48.1 (2020): 228-231.
[5] Zuo, Zheming, et al. "Data anonymization for pervasive health care: Systematic literature mapping study." JMIR Medical Informatics 9.10 (2021): e29871.
[6] Goldsteen, Abigail, et al. "Anonymizing machine learning models." Data Privacy Management, Cryptocurrencies and Blockchain Technology: ESORICS 2021 International Workshops, DPM 2021 and CBT 2021, Darmstadt, Germany, October 8, 2021, Revised Selected Papers. Cham: Springer International Publishing, 2022.
[7] Templ, Matthias, and Murat Sariyar. "A systematic overview on methods to protect sensitive data provided for various analyses." International Journal of Information Security 21.6 (2022): 1233-1246.
[8] Onesimu, J. Andrew, et al. "Privacy preserving attribute-focused anonymization scheme for healthcare data publishing." IEEE Access 10 (2022): 86979-86997.
[9] Qayyum, Adnan, et al. "Secure and robust machine learning for healthcare: A survey." IEEE Reviews in Biomedical Engineering 14 (2020): 156-180.
[10] Zhang, Ziqi, Chao Yan, and Bradley A. Malin. "Membership inference attacks against synthetic health data." Journal of biomedical informatics 125 (2022): 103977.
[11] Jammu, Aashmeen, and Harjinder Singh. "Improved AES for Data Security in E-Health." International Journal of Advanced Research in Computer Science 8.5 (2017).
[12] Haq, Hafiz Burhan Ul, et al. "E-Healthcare Using Block Chain Technology and Cryptographic Techniques: A Review." Pakistan Journal of Engineering and Technology 5.4 (2022): 21-28.
[13] Lewis, Nehama, et al. "Factors Influencing the Adoption of Advanced Cryptographic Techniques for Data Protection of Patient Medical Records." Healthcare Informatics Research 28.2 (2022): 132-142.
[14] Kavitha, S., P. J. A. Alphonse, and Y. Venkataramana Reddy. "An improved authentication and security on efficient generalized group key agreement using hyper elliptic curve based public key cryptography for IoT health care system." Journal of medical systems 43 (2019): 1-6.
[15] Olatunji, Iyiola E., et al. "A review of anonymization for healthcare data." Big data (2022).
[16] Arbuckle, Luk, and Felix Ritchie. "The five safes of risk-based anonymization." IEEE Security & Privacy 17.5 (2019): 84-89.
[17] Naresh, Vankamamidi S., et al. "Quantum Diffie–Hellman Extended to Dynamic Quantum Group Key Agreement for e-Healthcare Multi-Agent Systems in Smart Cities." Sensors 20.14 (2020): 3940.
[18] Narula, Gaurav, et al. "A Novel Review on Healthcare Data Encryption Techniques." International Conference on Innovative Computing and Communications: Proceedings of ICICC 2022, Volume 3. Singapore: Springer Nature Singapore, 2022.
[19] Tulu, Bengisu, et al. "Implementing digital signatures for healthcare enterprises: the case of online disability evaluation reports." International journal of healthcare technology and management 6.4-6 (2005): 470-488.
[20] Narula, Gaurav, et al. "A Novel Review on Healthcare Data Encryption Techniques." International Conference on Innovative Computing and Communications: Proceedings of ICICC 2022, Volume 3. Singapore: Springer Nature Singapore, 2022.
[21] Rajendran, Keerthana, Manoj Jayabalan, and Muhammad Ehsan Rana. "A study on k-anonymity, l-diversity, and t-closeness techniques." IJCSNS 17.12 (2017): 172.
[22] Choudhury, Olivia, et al. "Differential privacy-enabled federated learning for sensitive health data." arXiv preprint arXiv:1910.02578 (2019).
[23] Gadekallu, Thippa Reddy, et al. "Federated learning for big data: A survey on opportunities, applications, and future directions." arXiv preprint arXiv:2110.04160 (2021).
[24] Majeed, Abdul. "Attribute-centric anonymization scheme for improving user privacy and utility of publishing e-health data." Journal of King Saud University-Computer and Information Sciences 31.4 (2019): 426-435.
[25] .Dong, Xiao, et al. "Developing high performance secure multi-party computation protocols in healthcare: a case study of patient risk stratification." AMIA Summits on Translational Science Proceedings 2021 (2021): 200.

Research Article

[26] Hamza, Rafik, et al. "Towards secure big data analysis via fully homomorphic encryption algorithms." Entropy 24.4 (2022): 519.
[27] Kumar, A. Vijaya, et al. "Secure Multiparty computation enabled E-Healthcare system with Homomorphic encryption." IOP Conference Series: Materials Science and Engineering. Vol. 981. No. 2. IOP Publishing, 2020.
[28] Fang, Haokun, and Quan Qian. "Privacy preserving machine learning with homomorphic encryption and federated learning." Future Internet 13.4 (2021): 94.
[29] Arpita, Maheriya, and Shailesh Panchal. "Smart Health and Cybersecurity in the Era of Artificial Intelligence." Information and Communication Technology for Competitive Strategies (ICTCS 2021) Intelligent Strategies for ICT. Singapore: Springer Nature Singapore, 2022. 41-48.
[30] Digital Health Laws and Regulations Report 2023 India (iclg.com)
[31] Holohan, Naoise, et al. "($ k $, $\epsilon $)-Anonymity: $ k $-Anonymity with $\epsilon $-Differential Privacy." arXiv preprint arXiv:1710.01615 (2017).
[32] Maheriya, Arpita, and Shailesh Panchal. "A Neoteric Approach to Improvise Privacy Shielding of Sensitive Healthcare Information & Prediction of Disease." International Journal of Computing and Digital Systems 13.1 (2023): 1-12.
[33] Shafer J, Rixner S, Cox AL. The hadoop distributed filesystem: balancing portability and performance. In: Proceedings of 2010 IEEE international symposium on performance analysis of systems & software (ISPASS), March 2010, White Plain, NY. p. 122–33.
[34] Raval, Aanal, and Arpita Maheriya. "A Survey on Deep Learning Methods for Addressing COVID-19 Issues." Information and Communication Technology for Competitive Strategies (ICTCS 2022) Intelligent Strategies for ICT. Singapore: Springer Nature Singapore, 2023. 61-73.