Section A-Research Paper
ISSN 2063-5346

A Heuristic Approach to Protect Privacy of Patient's Sensitive Data with Prediction of Disease.

Arpita Maheriya
Research Scholar
Gujarat Technological University, Ahmedabad Gujarat, India
arpita_maheriya@gtu.edu.in.
Dr. shailesh Panchal
GTU-Graduate School of Engineering and Technology
Ahmedabad Gujarat India.
sdpanchal@gtu.edu.in

Abstract.
There is ongoing compulsion on health organization to share data for analyze purpose. The healthcare data includes patient behavior & records, DNA, laboratory test data, activity log, sensible data, cost data, and demographic data. Privacy becomes supplementally crucial in some scenarios when the data is shared with 3rd party along with the personal information of patients, and confidential record of healthcare organizations. Nonetheless, several suitable guidelines, privacy-preserving laws, and compliance requirements are there to safeguard electroclinic healthcare data. Although, privacy and security breaches remain key issues for healthcare systems. Anonymization techniques, however is liberate from the privacy related regulations. Machine learning models can imply on anonymized data. Thus, it generates an anonymized secure ML model, which provides greater protection against membership and attribute inference attacks. The heuristic approach results comparatively higher in accuracy where it does not violate data privacy and can be handled to train and test the model. Our security model's results suggest that the proposed model makes the healthcare data system secure and unauthorized access to protected patient healthcare information almost impossible.
Keywords Privacy of E-healthcare data. Anonymization techniques. Machine learning. Data protection. Privacy of electronic healthcare system.

1. Introduction

The recent pandemic situation has realized the need to collect and share healthcare data publicly on time. Realtime E-healthcare data helps policymakers and frontline workers in decision-making rapidly. This data is extremely critical, due to Sensitive personal information such as location, name, treatment plans & schedules, allergies, health condition, Mediclaim & insurance papers, payment details, and gender identity /sexual orientations. The researchers in this field mainly focused on the invention of drugs and making early & accurate diagnosis identification using E-medicate data. Many healthcare organizations, hospitals, laboratories, and institutes are sharing data for collaborative work with pharmaceutical companies, data analysts, data scientist & researchers for further studies. Sharing electronic health records with third parties would be preferable for data analysis tasks. It is essential to share electronic medical data while maintaining the confidentiality of data. Moreover, Electronic health record is enormously beneficial such as for accessing data from anywhere & at any time, reducing the cost of treatment and enabling telehealth consultancy. Fig. 1 is demonstrating the summarized major premises of E-Health care information [1].

On the other hand, electronic data spread expeditiously, which increases the chance of data breaches & cyber-attacks. Critical data of patients could be inferred by adversaries from public data repositories, e.g., Laboratory and surgery data holds normally vast amounts of personal information regarding the patient and family health history, chronic disease, vaccine and dose taken related information, etc. However, this massive amount of patient data collection tasks with maintaining data privacy and security is necessary and strenuous. Additionally, Electronic Medical Records (EMR) bring forward moral issues for doctors, nurses, patients, and clinical groups.

In the digital world, protecting patients' sensitive data become a more tough task than ever before. Although in the pandemic time healthcare organization are prime suspects of cyberattacks. Due to this healthcare industry is become a much more important feed to provide privacy of sensitive personal information. In some cases, data holders are also feeling hesitant to share information due to privacy issues. Healthcare organizations must provide a trustable environment and transparency to the patient while collecting and sharing sensitive data. There is a necessity of privacy-preserving tools and technology while sharing data publicly. This topic is cutting the edge for research studies. Security and privacy concerns of E-health data are an ethical requirement, and to meet the strict privacy of health industry data is continuedly growing with changes in laws.

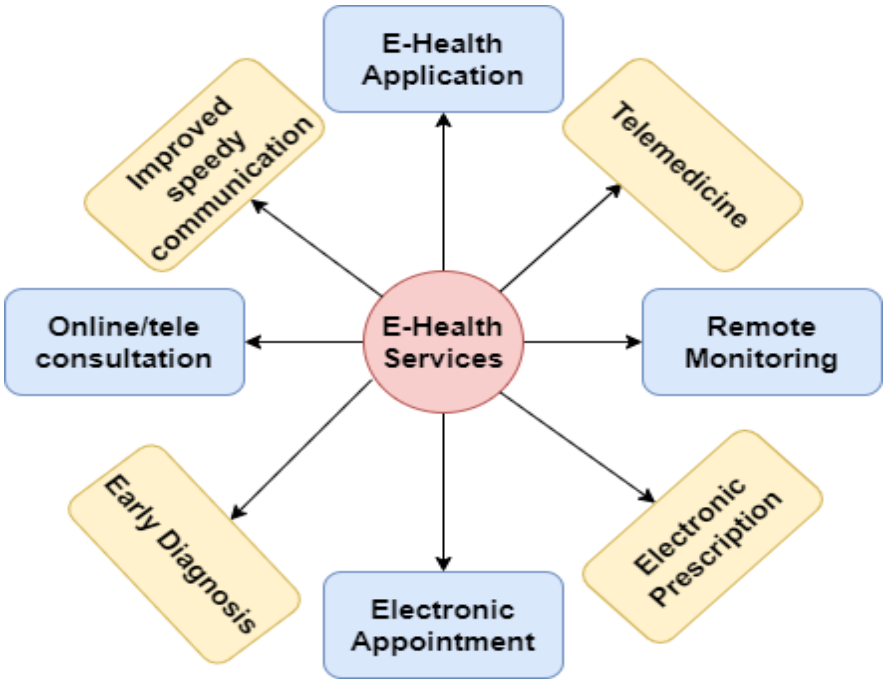

Fig.1.E-healthcare Service

Privacy is considered one of the essential human rights, therefore various degrees of nations and regions deployed privacy protection frameworks. The data privacy preserving laws, imposed by the Health Insurance Portability and Accountability Act (HIPAA). When it comes to Protect Health Information (PHI), HIPAA law comes first on the list. It is designed by the United State (US) government, enacted in 1996 and applicable to US citizens. And also, applicable each health organization in the world. Other contourites also developed HIPAA similar regulations, such as in US, the Family Education Rights and Privacy Act (FERPA), Same as to restrict the use of private information adopted in 2016, through the General Data Protection Regulation (GDPR), the European Union [2], Personal Information Protection Regulation and Election Document Act (PIPEDA) in Canada. India ranked 3rd number in the number of internet users [3].

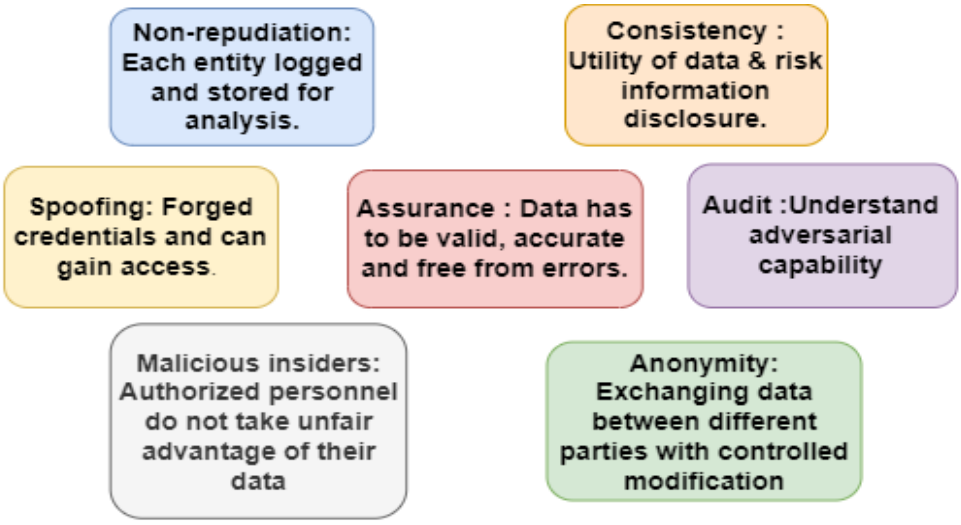

Fig.2.Requirment of Privacy in E- Healthcare

The current legal format of the India regulation act for security are Information Technology Act - 2000, Information Technology Rules - 2011 along with Digital Personal Data Protection Bill-2022(DPDP). The government of India is working to enact Digital Information Security in Healthcare Act (DISHA) is equivalent to HIPAA act [4]. To get access the information of sensitive data and transmission of data to the research community is a lengthy legal process with approvals. Nonetheless, government regulations and policies do not provide specific surety for privacy preservation [5]. However, industry, academic researchers, and scientists are preferring to use privacy-preserving methods in consideration of critical data privacy to fulfil the laws and regularities. Moreover, there must be a balance of social benefits between the advancement of scientific research and individual uses. Fig 2 is demonstrating the requirement of privacy in electronic healthcare-information.

1.1. Disclosure Risk

Violating privacy in terms of exploits result from analysis or data belonging to personal /confidential information getting re-identified is called a disclosure risk. This happens when an advisory wants to get some confidential information according to their benefit using published data. To perform discloser, adversary use background knowledge from external sources. Disclosed individual personal identity is defined as a privacy disclosure. Fig 3 shows types of privacy disclosers: 1) Identity disclosure thereat is the riskiest while sharing data publicly. It can be reidentified/revealed the actual victim identity by an adversary with high probability. Reidentification is performed by linking original data with protected data. For example, if the adversary knows X is 40 years old, then can be re-identified from the age of its record id while personal id is not available. 2) Attribute disclosure occurs when adversaries can recognize and link sensitive information directly with the victim (with high probability). However, Identity disclosure most probably leads attribute disclosure. Despite that, Attribute disclosure could possible to happen, if any identity disclosure there or not [31]. 3) Membership disclosure happens when adversaries successfully infer the existence of a selected victim in a public dataset. Table 2 demonstrated the risk disclosure parameters.

For performing any privacy attack adversary need some external knowledge and information regarding the public repository is a must. Moreover, that attacker would have other prior knowledge such as access to the public repository, having information on how the table is distributed sensitive attributes in the table 1. Attack models are used generally for performing privacy attacks. e.g., background knowledge attack linkage attack, homogeneity attack, skewness and similarity attack [5].

There are various requirements to protect e-health care data such as 1. Secure confidential identity disclosure of critical information 2. A concern of patients on personal data and spoofing 3. Anonymity and consistency 4. Confidential data membership and audit of data 5. To maximize the search space for attackers and anonymization 6. Assured, audit, authorization, authenticity, confidentiality and integrity & nonrepudiation.

According to the requirement, available privacy approaches as the same sequence of requirements: 1. Generalization 2. Suppression 3. Pseudonymization 4. Slicing 5. Randomization 6. Cryptographic approaches. Data generalization for privacy is the most preferable privacy technique. In this, values are replaced with blurry and fewer precise values. It helps in data utility and protection from some attacks [1]. As per the privacy protection scheme, (GDRP) anonymization of critical information is quite necessary. The key idea of anonymization is to secure that personal data is no longer to be traced unambiguously [2]. For this motive many fundamental privacy models are presented, such as k-anonymity [15], L-diversity [16], T-closeness [17], δ-presence [18], and differential privacy [19]. from all of these approaches to privacy, the preservation of k-anonymity is more famous [20].

Table 1. Classification of Healthcare Data

Healthcare Data is in table format and relational data. Where each tuple /row belongs to one owner of the record and a number of various attributes corresponds to each column, that could be extended into four types of classification.	**Personal (ID) /Explicit identifiers** are identified uniquely each record belonged, such as security number, name, social security IDs, Government IDs mobile number & driving license etc. these records are deleted in anonymization process, because it could identify record in first-hand
	Quasi-identifier (QID) can be used to revel personal identity (Not uniquely but with some attribute possible). such as DOB, sex, postal code, age, city etc.
	Sensitive attribute (SA) is individual related sensitive data that each wants to keep it private (must be protected). such as disease, insurance information, treatment history.
	Non-sensitive attribute (*NSA*) does not volatile individual privacy.

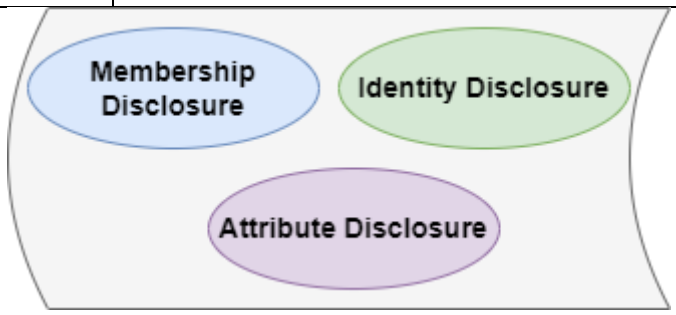

Fig.3. Privacy Disclosure Risks

Table 2. Privacy Disclosure Risk Measurement [48]

Disclosure Risk	**Privacy Risk Measurement**		**QI Datatype**
Identity Disclosure	**Outlier** Rank-based intervals, Distance based intervals, Density based intervals Numerical Standard deviation-based intervals	**Clustering** Distance-based intervals, and Density-based intervals	Numerical
	k-anonymity, k-Map, Pitmap, and Logarithmic Series		Categorical
Attribute Disclosure	**K-anonymity based** t -Closeness, Categorical σ-	**Record linkage** Probabilistic-based,	Numeric and categorical

	Disclosure privacy, Categorical, β-Likeness	Distance-based, Rank-based

2. Background Study

In the recent pandemic has observed that healthcare organizations are primary victims of attackers. The paper [6] is identified issues while pandemic in healthcare sector. And provided suitable suggesting to protect data sharing. The motivation behind paper [7] is to identify minimal cost optimization models for the security of healthcare organizations. Furthermore, demonstrated 64 district security measures linked with 70 vulnerabilities that might occur and create external or internal risk. This helps to identify vulnerabilities within their subcategories. Due to modern technologies healthcare sector also move towards digitalization. Healthcare services are made accessible for patients through distinct channels of interaction. It is created issues regarding interoperability in-between numerous interaction channel's services are present there for patients. Overview of privacy techniques with its application health organization and future challenges and suggestion mentioned in review paper [47]. This article [8] demonstrated multiple channel integration problems and provides a possible solution. In [9], the author studied the four databases (PubMed, PakMediNet, Google Scholar, and Medline) from 2011 January to 2021 June. With using the terms as keywords: ethical issue, Electronic Medical Record (EMR), digital health record, health care providers). The author followed the technique PRISMA 2020. Following ethical principles identified, in concern ethical challenges faced by healthcare organization while using EMR data in healthcare organizations: 83% privacy (n:- 20/24), and 62.5 % autonomy(n:-15/24) followed by 58% confidentiality (n:-14/24), 50% justice (n=12/24), 16% trustworthiness (n:-4/24) and 8.3% sincerity (n:-2/24). Challenges identified by the author from 24 papers. The author found a lack of commitment to ethical rules and robust monitoring in medical practice in developing counties. However, Healthcare organizations are caring out sensitive data regarding patients, and sharing this data could be led to the inadvertent reveal of privacy.

In this paper [5], the Authors represented a study on state of art - privacy enchantment techniques that guarantee of a secure environment for sharing healthcare data. In this paper [5], the Authors represented a study on state of art - privacy enchantment techniques that guarantee of a secure environment for sharing healthcare data. The authors [41] introduced two privacy notations and two schemes for discloser of data issue. The detailed overview [10] on the latest presented mechanism based on data anonymization and DP (Differential Privacy) for data privacy in the health sector. De-identified data/ anonymized dataset generates data loss, which presumably impact on data analysis and predication task [49]. Moderated error showed in result of anonymized data analysis compare with original dataset [50]. Bottom coding is playing importantly significant part in differential privacy classification. Those risky records which most probably get re-identified by adversary are get suppressed using bottom coding (hides in-between dummy data). Additionally, the pros and cons of these two mechanisms with the future scope are discussed briefly. Methodist Environment for Translation and Outcome [11] (METEOR) model is proposed based on privacy and security for healthcare security. This model prevents data breaches and unauthorized access. Hospitals and National Health Service with GP (general practitioner) surgeries collects regularly patient health-related data, while maintaining privacy. This paper [12] author purposed a privacy-preserving Generative Adversarial Network, which helps to build various ML models to provide help to members of the NHS. The proposed model calculates the privacy score and cosine similarity score of synthetic data and provides privacy and high utility. The results are based on a comparison of the ml model performance on original data and synthetic data.

This case study [13] described the achievability of using differential privacy techniques on publicly available healthcare data sharing while providing privacy and confidentiality. DP allows data to be generated more granular with strike balancing confidentiality and privacy. K-anonymization is performed on geometric perturbed data, ML classification accuracy of data is high in compared to performing k-anonymization on original data while maintaining data utility [14]. Recently research on genomics data is in trend due to that

privacy concerns needed while sharing with third parties, storing, computation E-data. In this survey [21] authors cover and summarize the privacy-preserving method with cryptographic techniques. Furthermore, training an artificial intelligence model is raising concerns about the privacy of individual personal data. This privacy risk may lead to constraints on accessing genomic biomedical data for researchers and other third-party vendors for further use or analysis. This paper [2] mentioned the overview, limitations, and strengths of methods of AI along with privacy-preserving. patient's personal data while sharing with the third-party user. That would help in encouraging and enabling scientific-research study.[1] demonstrates the new anonymization methodology for e-healthcare data privacy. This proposed scheme improves data privacy and utility while publishing data.
Healthcare 4.0 come along with privacy and security issues. [22] presented the pros and cons of existing privacy techniques. challenges and future research scopes in direction of privacy of healthcare4.0. Nowadays industry 4.0 is in trend, users want to add data remotely artificially while preserving privacy. This paper considers anonymization methods as a solution. At the same time, it would be possible to apply the AI model on anonymize result data. Implementing a machine learning model on e-data could cause data breaches/leakage of sensitive healthcare data. In this paper [23] author performs homomorphic encryption on six different datasets using a linear regression model without infecting the accuracy of the regression model and analyzing performance. Another paper [24] uses households of fully homomorphic privacy mechanisms-TFHE and HEAAN and applies regression models-linear and logistics. The [25] authors performed the most used k-anonymization algorithms with different machine learning classifiers and demonstrate performance by using distinct datasets. The result mentioned that classification accuracy is based on the degree of anonymization and dataset. This paper [26] showed the effect of the homeomorphic data space transformation anonymization algorithm on deep feedforward network performance.

3. Data Anonymization Models

In today's time, data analysis and research on medical data records will contribute in many ways to society. Many personal data related to the healthcare field need to be collected and shared in the public repository. The anonymization of healthcare records is important for personal well-being, and making available. The de-identified information is beneficial for non-hospital researchers/analysist/third parties[49]. Moreover, it will make ease in research on early disease identification. However, other historical data are also present, when both these data sets get linked with each other, it will be productive for adversaries to identify patient-related personal information such as fingerprint, bank and insurance detail, photo, contact number, etc. Other various types of data like birth place, postal code, and blood group often link uniquely help to re-identified the person. Protect the data can processes using Privacy Preserving Data Mining Methods (PPDM). The task of publishing data without any compromise of individual identity along Privacy-Preserving Data Publishing (PPDP) is major area of researchers and practitioners. Anonymization methods are purposed to provide PPDP while sharing data publicly / third party [34]. Apply anonymization, and protect the privacy of data at the same moment maintaining data utility is a challenging task. Table 2 is showing applicable transformation models base on types of attributes. And table 3 is the list of anonymization tools available.
Following are the various data anonymization techniques described [14]:

- Generalization: In these technique values of the real data are replaced with a more generic (less specific form) of that. However, it would stay semantically constant with real value. For generalization, numerical value converts into range value and categorical value converts base on hierarchy. For example, engineers' and doctors, values were replaced with profession. The generalization algorithm is further categorized into three:1) top-down VS bottom-up 2) global/single dimensional vs local/multidimensional 3) complete/optimal vs greedy/approximate.
- Suppression: This technique is preventing information disclosure by replacing value or (with "*").
- Anatomization: It diverges from generalization and suppression. It is neither altered in value of quasi-identifier /sensitive value. Here, sensitive and quasi-attributes are separated into two different tables, which are linked with the same group id in different tables [41].

- Perturbation: In this technique, the initial value is replaced with the same statistical value (synthetic value).

Table 3. Data transformation Models

Type of Attribute	Applicable Transformation Models
Categorical Attributes	Generalization Methods Multi-dimensional and Full-dimensional generalization Suppression: Cell, Attribute, and Record levels Sampling: Random and by query Aggregation: Mode and median and set
Numerical Attributes	Generalization Methods Multi-dimensional and Full-dimensional generalization Top-down coding Categorization Suppression: Cell, Attribute, and Record levels Sampling: Random and by query Aggregation Geomatic and Arithmetic mean Mode and median and set Interval

Table 4. Data anonymization Tools [40,57]

Tools	Publish to Recent Update Year	Originate	OS	licence	Open source	Language	Privacy & Utility Evaluation
μ-*Argus (*Anti Re Identification General Utility System microdata) [52]	1998-2021	Centraal Bureau voor de Statistiek (CBS)-Netherlands	All	EUPL	✓	Java, C++	✓ and
sdcMicro - Statistical Disclosure Control Methods for Anonymization [54]	2007-2021	Statistics Austria-Austria	All	GPL 2	✓ (GUI)	R	✓ and
Open Anonymizer	2008-2009	University of Vienna- Austria	All	Un-define	✓	Java	
Cornell anonymization tool-CAT	2009-2014	Cornell University-USA	All	Un-define	✓	C++	
Anonimatron [55]	2010-2018	-	-	-	✓ (GUI)		and
ARX [36]	2012-2022	Berlin -Germany	All	Apache 2	✓ (GUI)	Java	✓ and ✓
Aircloack [56]	2012-2022	Software company in Kaiserslautern, Germany	All	-	(GUI& Web)	-	✓ and ✓
Amnesia [51]	2019-	University of		BSD-	✓ (GUI	Java and	✓ and ✓

	2022	Thessaly, Greece		3cluase	&Web APP	java script
Probabilistic anonymization	2018-2018	University of Cyprus -Cyprus and Newcastle University-UK	-	Un-define	-	R

Privacy Models are aimed to protecting medical dataset record from re-identification using anonymization techniques. Here, discussed three popular anonymization technique, called k-anonymity, l-diversity, and t-closeness. Table 4 described about anonymization tools.

- K-anonymity: In rich electronically available data-world, sharing sensitive data related to the health and financial field in such a way that, personal identity does not get disclosed. one of the widely used methods is k-anonymity. It is use to minimize linking attacks by quasi-identifier [27] k-anonymity technique is purposed. In this method minimum, k number of quasi-identifier in an equivalence class in the way that it can reduce the linkage of the subset of quasi-identifier at most 1/k. While there are many others k-anonymity techniques available with combining generalization and suppression rules. There are several advantages [28]: 1) result of k-anonymity can be interpreted easily. 2) outcome determines a person's truthful identity; it is useful for healthcare assessments and traceable pattern of identity revealing. 3) this anonymity technique guarantee deforms. 4) prevent linkage of records.5) compare with another cryptographic solution cost to generate it is lower [33]. K- anonymous outcome database has at least (k-1) quasi-identifier, which is deformed with each one. k-anonymity is affected by attacks called, background knowledge and homogeneity attacks. It doesn't provide fewer diversity values of sensitive attributes in a group. Hence, k-anonymity can be disclosed sensitive values. Furthermore, k-anonymity succeeds to prevent identity disclosure but it is vulnerable to attribute disclosure.
- L-diversity: It is a refined method of previous one(k-anonymity) that measure the diversity of sensitive attribute. The database is l-diverse, if each identical quasi-identifiers in the row have at least l-number of diverse values linked with the sensitive attribute of that database. In practice, diversity is easily understandable and can address the k-anonymity shortcoming: homogeneity and background knowledge attack. Diversity in privacy is useful while publishing data when the publisher is unknown of the knowledge level of the adversary. By setting parameters, the publisher can determine protection against background attacks. Values in the group are diverse if it has represented specific values. L-diversity provides a guaranteed framework for stronger privacy. Moreover, k-anonymity and l-diversity have quite similar structures in sense of time, parameter values, and cost. Distinct l-diversity is not satisfied probabilistic inference attack. To overcome these two stronger types were used: Entropy and Recursive l-diversity notations. L-diversity is beyond the k-anonymity but it also has some shortcomings: 1. It might be tough and irrelevant to achieve. 2. Inadequate avert attribute disclosure. 3. Some attacks suck as, similarity and skewness attacks. Furthermore, whenever the dataset is skewed, fulfilment of diversity cannot able to prevent the disclosure of attributes [28].
- T-closeness: Recently, t-closeness is defined as an anonymization technique. It is an expansion of the k-anonymity and l-diversity based anonymization group, which decreases the granularity of the dataset. The dataset has T-closeness for each group of data sharing when, the distance in-between distribution of sensitive attributes in a group and the distribution of each attribute in set is not more than enough than the threshold of T [29]. If k-anonymity and t-closeness work together, it generates confidential data. And that prevents background knowledge & homogeneity attacks [32].

4. Machine Learning and Anonymization Models

Any personal data leakage caused by Machine Learning (ML) train and test models can usher to the disclosure of identity and attributes in many cases either direct way or indirect way. Some latest research has demonstrated that privacy gets violated and arouses attacks to exploit ML model data [30].ML models are used to discover the sensitive columns risk in the dataset same as upholding the utility of data [31]. It is confirmed that stronger

anonymization yields drop in execution of model [46]. Applying predictive models on anonymized datasets is one of the interesting research areas. The value of the data utility is an essential parameter for any anonymized datasets, which means data will be used in the future. in this paper [34], diverse anonymities algorithms introduced and applied regression and classification models on k-anonymized data. The ML models accuracy for trained & tested on anonymized dataset is up to the benchmark in numerous cases with preserving privacy. Anonymized models can assist in overfitting avoidance.

An approach [44] demonstrated to preserve membership inference attack on ML models with differential privacy (DF), this approach also has some downside such as performance overhead, model implementation task, complexity increased. step by step demonstration [45] of implementation of DP with model architecture and hyperparameter tunning. Case study [31], proposed the methodology of preserving privacy while maintaining the usefulness of the data. For that ML prediction model generated, which has an accuracy average 75%, besides a smaller number of privacy preservation. A novel approach [33] present is decreasing the of correlation between RFD data while data anonymization with generalization rules and prove data utility by applying decision support ML models. In this paper author [35] proposed a novel ML-based approach iterative Named Entity Recognition use on a semi-structured document, which provided F measured 99.75% accuracy. This paper also provided other beneficial tasks that can handle structured documents such as hospital records. This study [39] offers favorable use of 5 states of ML- prediction models on heart disease in south India. The data scrubbing [42] method on anonymized data for prediction of kidney injury provide better feature diversity and minor change in data utility. Eventually, use of emerging technologies with legal certain regulation will provide secure platform to drive in medical advancement while securing sensitive data [43].

5. Implementation Strategy.

For implementation purposes, the ARX Data Anonymization Tool is used [36]. ARX anonymization tool is reaccommodate for the researchers for automated anonymization process of relational / tabular data [40]. Coding is performed using python language and its libraries. Experiment work has done on windows 11 system with 16 GB RAM.ARX Data Anonymization Tool is open-source and java based with. This software has various 1) privacy and risk models 2) techniques to transform dataset attributes values 3) analysis of the utility of data. It has been used for clinical critical data sharing for further research and other purposes. ARX tool is used instinctively for customizing anonymized datasets as purpose and visualization of dataset utility and re-identification risk of the anonymized dataset with the original dataset. Publicly available repository's datasets are used in this experiment, such as Kaggle and UCI repository. The Adult Dataset-UCI Machine Learning Repository [37] and COVID-19 patient pre-condition dataset from Kaggle [38] (Anonymized Dataset) are considered for evaluation. The adult dataset has 14 attributes, multivariate characteristics, and categorical, and numerical type data. The adult data set is holding the following attributes: marital status, age, capital-loss, work-class, fnlwgt, native-country, education, capital-gain, education-num, race, occupation,income, relationship, and sex. Among these 14 attributes, considered 2 attributes (Age and Education_Num) as quasi-identifier and the COVID-19 patient pre-condition dataset contains information specifically, patient's history and habits, according to laws such as HIPAA and GDPR sharing personal health information is not possible. In this dataset, most of the data related to patients are in an anonymized format. This dataset contains the following 23 attributes: Id, COPD, sex, asthma, patient_type, entry_date, covid_res,date_symptoms,tobacco,date_die,renal_chronic ,intubated, on pneumonia, age, pregnancy, diabetes, , insurer, hypertension, other disease, cardiovascular, obesity, ICU. Among these attributes, dates, and age are considered quasi-identifier. Table 5,6 are demonstrating attribute classification for datasets.

Table 5. Adult Dataset Classification

No of Attributes	14
Attributes	marital status, age, capital-loss, work-class, income, fnlwgt, native-country, education, capital-gain, education-num, occupation, relationship, race and sex
Categorical Attributes	work-class, education, marital status, occupation, relationship, sex, race, income and native-country
Sensitive Attributes	income

Quasi-identifier	age, education-num
No of Record	48843

Table 6. COVID-19 patient pre-condition Dataset Classification

No of Attributes	23
Attributes	id, sex, patient_type, entry_date, date_symptoms, date_died, intubed, pneumonia,age, pregnancy, diabetes,diabetes,asthma,COPD, inmsupr, hypertension, other_disease, cardiovascular, obesity,ICU,tobacco, covid_res, contact_other_covid, and renal_chronic
Categorical Attributes	sex, patient_type,intubed, pneumonia,age, pregnancy, diabetes,diabetes,asthma,COPD, inmsupr, hypertension, other_disease, cardiovascular, obesity,ICU,tobacco, covid_res, contact_other_covid, and renal_chronic
No of Records	563201
Quasi-identifier	age and dates

Figure 4 represents the pre-procession /anonymization technique stages. For both datasets, first perform analysis on each attribute of the datasets, the number of attributes & it's type, analyses the personal identifiers, quasi-identifier, and sensitive identifier. Next find the vulnerabilities of attributes and separate that attribute with higher vulnerability into a table. Apply anonymization techniques on sensitive and quasi-identifier and analyse impact on dataset privacy. Perform suitable anonymization along with level of generalization & suppression rule on dataset. Figure 5 is representing ML model's stepwise flow on anonymized datasets. For ML model input, use the original dataset as well as anonymized plus original data and compare the result for both data. The Recursive Feature Elimination (RFE) method used for future selection which provide effective more relevant features to predict target result. The primary motive here to maintain trade-off of privacy utility with ML models.

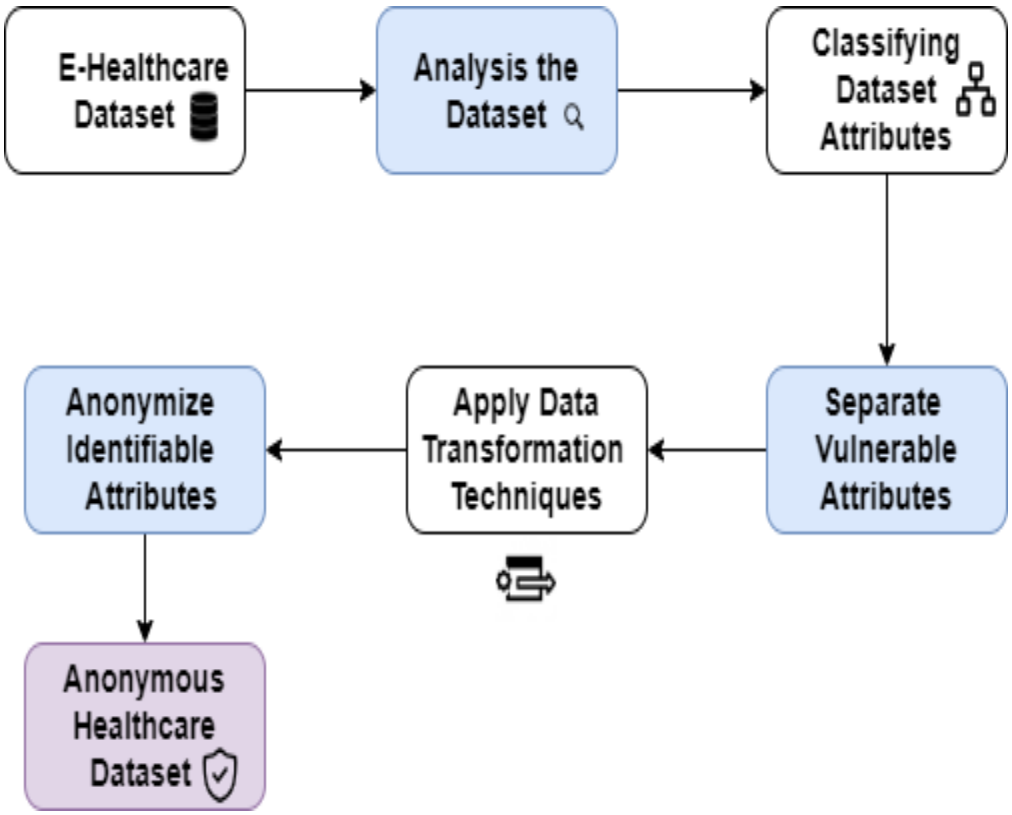

Fig.4.Pre-Processing & Anonymization Steps

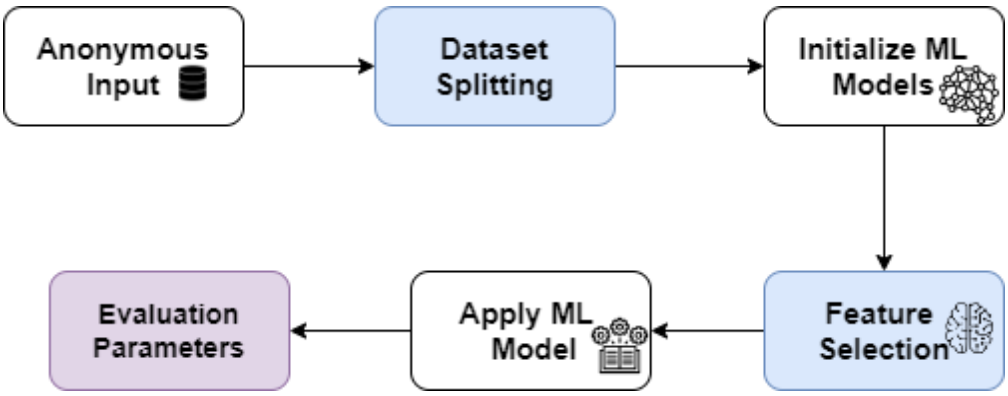

Fig.5.Machine Learning Model Steps

6. Performance Measurements

Performance Measurement is based on following parameters and functions:

- Recursive Future Selection: It is shortly known as RFE. It is well known future selection method. It is effective at selecting columns/ features or the least effective features eliminate from training set, that are most relevant while predicting the target variable.
- Root Mean Square Error: It used to calculate the average difference between the predicated values by the model with the actual values. It shortly known as RMSE. The main purpose of this standard derivation is to define how accurately model predict /model fit to the data /analysis and verify the predication result.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(Predicted\ values\ i - Actual\ values\ i)}{N}}$$

- Accuracy: It is percentage of correct prediction / regression result of model.
  T= True positive + True negative
  N= True positive + False positive+ True negative+ False negative
  Accuracy = T/N
- Precision (P): It measure all the positive values measured correctly or incorrectly as it is.
  True positive / (True positive +False positive)
- Recall (R): It measure ratio of positive values correctly classified.
  True positive/ (True positive +False negative)
- F1 Score: it calculates model's accuracy by combining precession and recall.

$$\text{F1 Score} = 2 \times \frac{P \times R}{P + R}$$

7. Results Analysis

ARX Tool is for data anonymization. It will help us to anonymize the personal sensitive information. That can configure and transform the dataset according to requirements. It provides various privacy models classification algorithms covers utility based on risk analysis utility. The result shows us risk on attacker models of Prosecutor, Journalist, and Marketer. For this experimental analysis on datasets, applied most popular model k-anonymity of anonymization with K=2 for both datasets with loss utility metric. Also applied generalization rule and suppression along with K-anonymization. As presented in Figure 6 and 7(of both dataset), after performing k-anonymity anonymization model, recorded risk for various type of risk models' percentage reduced drastically. For the value k=2 highest risk (100%) has got decreased half of the percentage and other attacks models have reduced nearby (0 to 1) % for both datasets. Furthermore, it is noticed during simulation that as value of parameter k increases, it directly impacts on highest risk percentage for various attacker risk models, which also reduced re-identification risk. Risk of percentage decreased are following: For Adult dataset, Average prosecutor risk and marketer success rate from 2.06% to 1.91% & record risk from 1.08% to 0.95 %. For COVID-19 patient pre-condition dataset, Average prosecutor and estimated market risk from 21.74% to 0.005, estimated prosecutor and journalist risk from 100% to 0.011%, sample and population uniqueness are 0%.

Figures 8-10 are showing, comparison of linkage between the quasi-identifier after partition of values using various generalization hierarchies and suppression rules along with the following approaches of anonymity in sequences: 8) K-anonymity model (K=2), 9) L-diversity model (L=2) and, 10) T-closeness (T=2) using python language and libraries on the adult data set. The diminished linkage of the quasi-identifier's impression parallelly

reducing. identification risk for sensitive data base on relation between quasi-identifiers. The linkage between quasi- identifiers gets reduced after applying the anonymization model with generalization and suppression approaches. L-diversity shows much better result than k-anonymity and T-closeness show more promising result compare to k-anonymity & l-diversity.

Fig.6. Risk Analysis Using ARX Tool (Utility=Loss)

Fig.7. Risk Analysis Using ARX Tool (Utility=Loss)

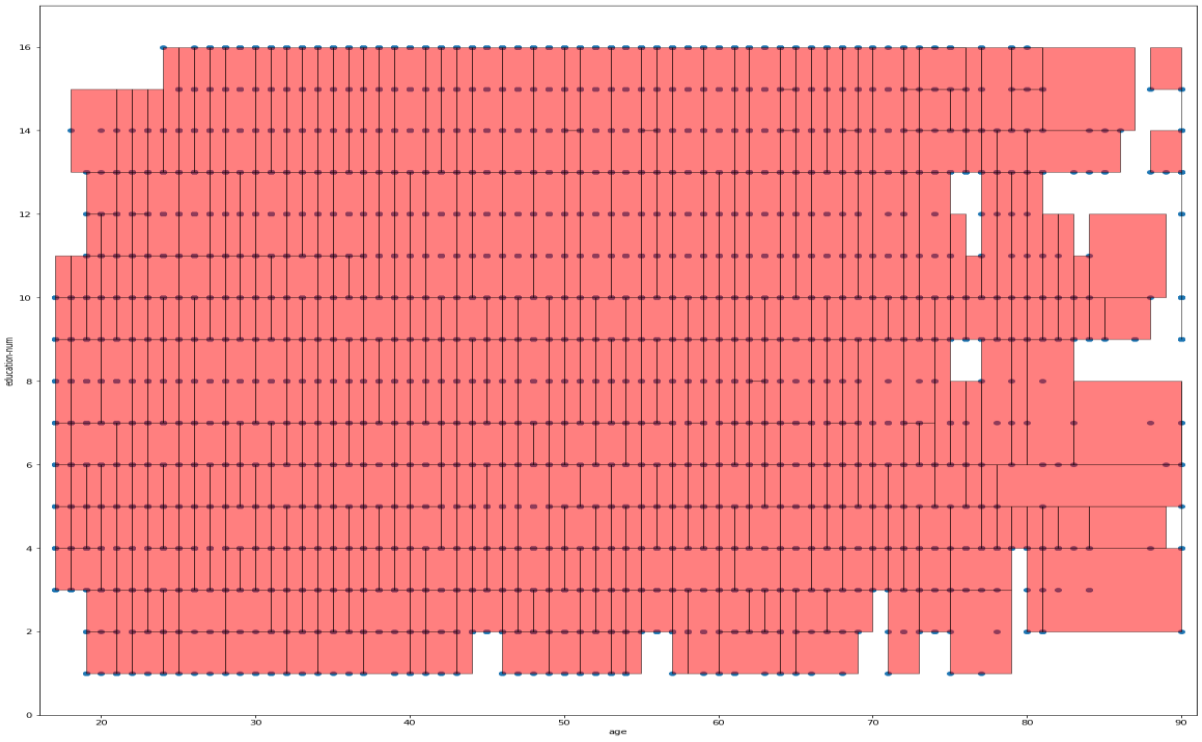

Fig.8. Linkage of Quasi-identifier using k-anonymity

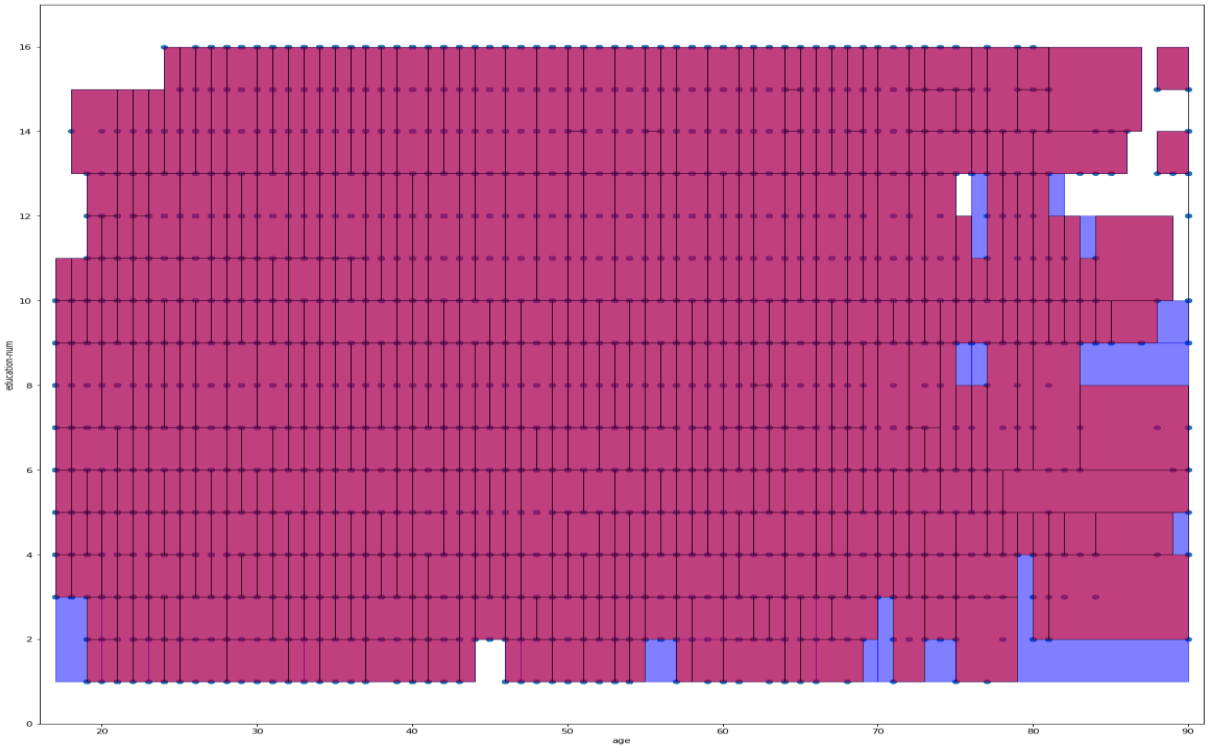

Fig.9. Linkage of Quasi-identifier using l-diversity

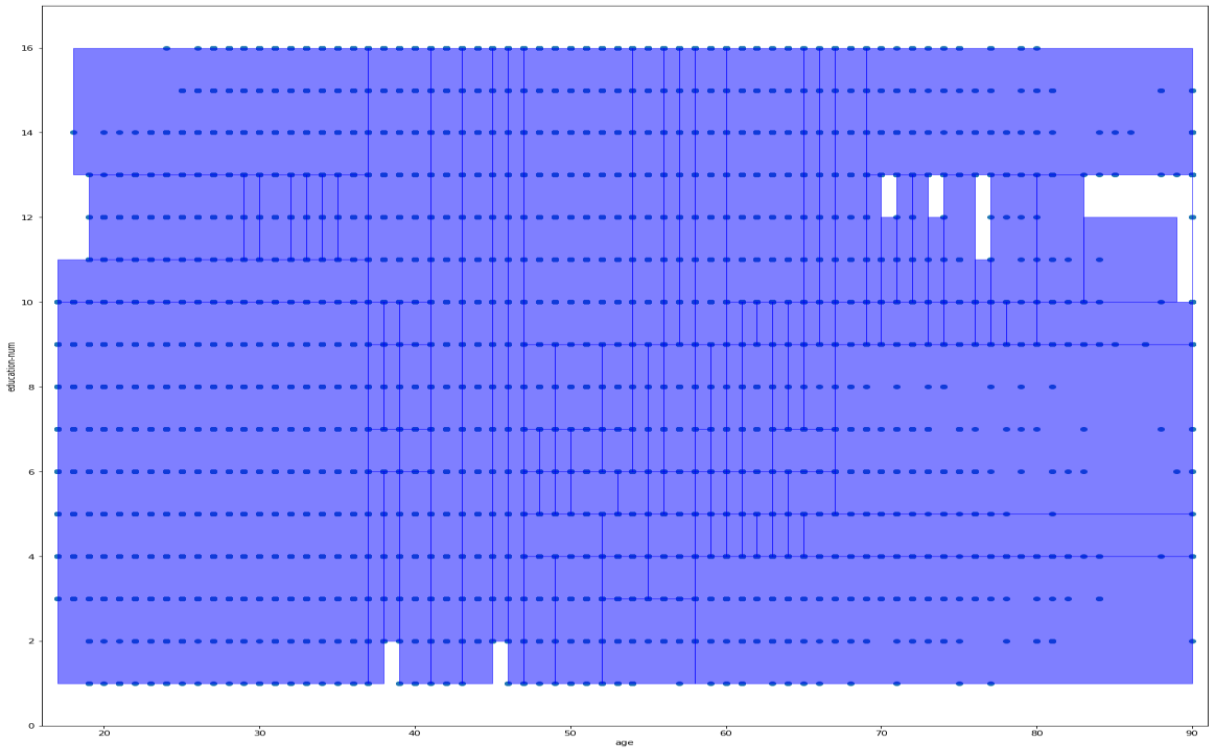

Fig.10. Linkage of Quasi-identifier using t-closeness

Table 7 is showing the classification of adult dataset columns while performing k-anonymization and k-anonymization & differential privacy. There is total 18 attributes and 15649 rows, among them 9 attributes are considered as explicit ,6 as sensitive and 6 as quasi-identifier. While performing k-anonymization model and k-anonymization plus differential privacy model is resulted 42.93% and 34.67% data loss respectively for quasi-identifiers value 3 and 5 respectively. The figure 11. is the graphical representation data loss after applied anonymization models on adult dataset. That prove that anonymization techniques cause data loss of original dataset.

Table 7. Attribute Classification

Classification	Numbers
Attributes	18
Explicit Identifier	9
Sensitive Attribute	6
Quasi Identifier	3
K-quails	3
E-quails	3

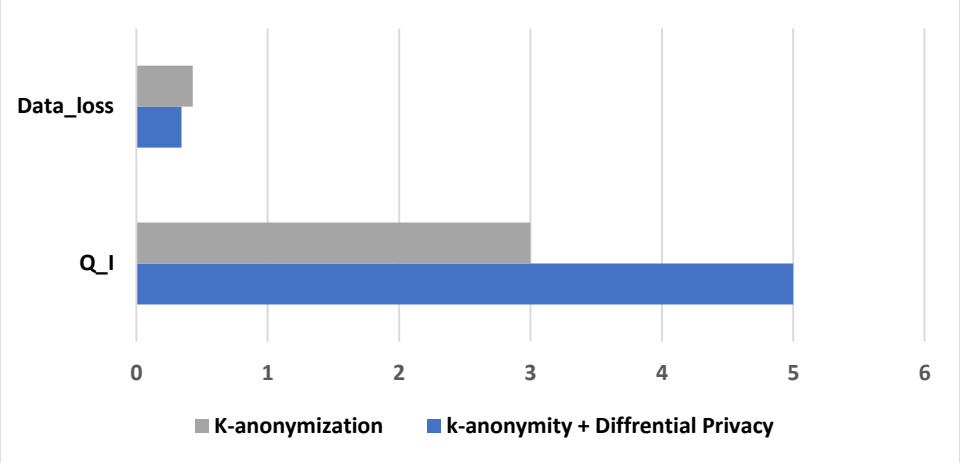

Fig.11 Data loss of Anonymized data

As discussed earlier, Machine learning models support in recognizing the utility and loss of the dataset. So, here in experiment applied several ML classification models implemented on the Pre-anonymized Covid-19 Patient Pre-condition dataset. Table 8. Demonstrating the evaluation ML models with train & test accuracy, precession, recall and f1-score.

Table 8. ML Classification Models

Machine Learning Models	Training Accuracy	Test Accuracy	Precession	Recall	F1-score
Logistic Regression	72.22%	72.30%	83%	56%	67%
K-Neighbors Classifier	92.59%	83.83%	84%	84%	84%
XGB Classifier	88.42%	88.14%	94%	81%	87%
Gradient Boosting Classifier	88.13%	88.12%	94%	82%	87%
Random Forest Classifier	93.31%	87.67%	88%	87%	88%

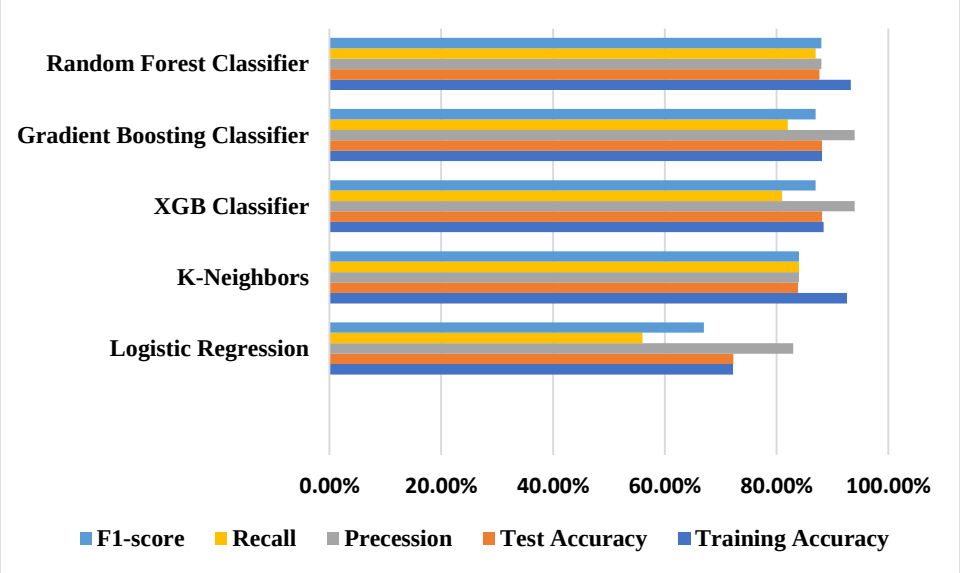

Fig .12 ML Classification Models Comparison on Anonymized dataset

8. Discussion and Conclusion

There has been continuously growing pressure and serious concern on healthcare organizations to disclose research-related records. Especially this data contains sensitive private information. Anonymization techniques can protect the disclosure of sensitive information while sharing with 3rd party organizations/researchers. There are various ways to anonymize the data precisely, k-anonymity is one of the prime privacy-preserving methods. In this paper, used 2 publicly available datasets: Adult & COVID-19 Patient Pre-Condition Datasets. On both datasets, applied k-anonymity rules with ARX Tool (utility matrix='loss'). k-anonymity anonymization methods help in a reducing risk percentage for several attacks possible on healthcare publicly available datasets. which proved with simulation performed on datasets. The diverse anonymization techniques implied, and T-closeness provided a better solution to protect against the linkage attack. Furthermore k-anonymity anonymized method resulted lesser data loss with DP method. In compare to other implemented ML model, XGB classifier produced the prominent result of train & test accuracy (88.42% & 88.14%) on anonymized data classification model.

As part of further research, implementation of proposed flow on various healthcare sensitive dataset. That provide analysis while preserving privacy. Proposed model gives us better accuracy to ML with moderate data loss on de-identified dataset. Hence, it indicates the new area of research, prediction of anonymized dataset with emerging technological advancement in healthcare organization. which also provide protection to the ML model with membership and attribute inference attacks.

List of Abbreviation

EMR: Electronic medical records
GDRP: General data protection regulation
FERPA: Family education rights and privacy act
HIPAA: Health insurance portability and accountability act
PHI: Protect health information
PIPEDA: Protection regulation and election document act
DISHA: Digital information security in healthcare act
QID: Quasi-identifier
SA: Sensitive attribute
NSA: Non-sensitive attribute
PPDM: Privacy preserving data mining methods
ML: Machine learning
RFE: Recursive Future Selection.

Declarations

Funding Not applicable
Conflicts of Interest The authors declare no relevant conflict of interest
Competing Interests The authors declare no relevant competing interest.
Author contribution All authors have equal contributions.
Ethics approval Not applicable
Consent to participate Consent was obtained from all individual participants included in the study.
Consent to publish Consent to publish has been received from all participants should appear in themanuscript.
Availability of Data and Material Datasets used in this experiment is publicly available [37][38].
Code Availability When certain code sharing requirements are met, the code is available upon request. Such requests should be sentto the first author of this paper.

Reference

[1] 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 31.4 426-435, (2019).
[2] Torkzadehmahani, Reihaneh, Reza Nasirigerdeh, David B. Blumenthal, Tim Kacprowski, Markus List, Julian Matschinske, Julian Spaeth, Nina Kerstin Wenke, and Jan Baumbach.:Privacy-preserving artificial intelligence techniques in biomedicine. Methods of Information in Medicine 61, no. S 01 e12-e27 (2022).
[3] Kesharwani S, Sarkar MP, Oberoi S.: Cyber security in India: threats and challenges. Cybernomics,Oct (2019).
[4] DISHA to give direction to digital information security in healthcare : Supratim Chakraborty, ET HealthWorld (indiatimes.com) Acessed 19 April 2023.
[5] Chong, Kah Meng.: Privacy-preserving healthcare informatics: A review. In ITM Web of Conferences, vol. 36, p. 04005. EDP Sciences, (2021).
[6] Arpita, Maheriya, and Shailesh Panchal.: Smart Health and Cybersecurity in the Era of Artificial Intelligence. In Information and Communication Technology for Competitive Strategies (ICTCS 2021) Intelligent Strategies for ICT, pp. 41-48. Singapore: Springer Nature Singapore, (2022).
[7] Sönmez FÖ, Hankin C, Malacaria P.: Decision support for healthcare cyber security, Computers & Security,122:102865,Oct 1(2), 32-4, (2022).
[8] Moreira, Ailton, Tiago Guimarães, Ricardo Duarte, Maria Manuel Salazar, and Manuel Santos.: Interoperability and Security Issues on Multichannel Interaction In Healthcare Services. Procedia Computer Science 201, 714-719, (2022).
[9] Afzal S, Arshad A.: Ethical issues among healthcare workers using electronic medical records: A systematic review. Computer Methods and Programs in Biomedicine,1.100030, (2022).
[10] Kikuchi, Hiroaki.: Differentially private profiling of anonymized customer purchase records. In Data Privacy Management, Cryptocurrencies and Blockchain Technology: ESORICS 2020 International Workshops, DPM 2020 and CBT 2020, Guildford, UK, September 17–18, 2020, Revised Selected Papers 15, pp. 19-34. Springer International Publishing, (2020).
[11] Vaishnav, Rounak, Ms Divya Deepthi Panditi, Viney Dhiman, C. Catherene Julie Aarthy, Y. Saritha Kumari, and Md Khaja Mohiddin.: Data security in healthcare management analysis and future prospects." Materials Today: Proceedings 51,pp.2202-2206,(2020).
[12] Venugopal, Rohit, Noman Shafqat, Ishwar Venugopal, Benjamin Mark John Tillbury, Harry Demetrios Stafford, and Aikaterini Bourazeri.: Privacy preserving Generative Adversarial Networks to model Electronic Health Records." Neural Networks 153,pp. 339-348,(2022).
[13] Dyda, Amalie, Michael Purcell, Stephanie Curtis, Emma Field, Priyanka Pillai, Kieran Ricardo, Haotian Weng et al.: Differential privacy for public health data: An innovative tool to optimize information sharing while protecting data confidentiality.Patterns 2, no. 12,pp.100366,(2021).

[14] Kundeti, Naga Prasanthi, and Chandra Sekhara Rao MVP.: Accuracy and utility balanced privacy preserving classification mining by improving K-anonymization." International Journal of Simulation–Systems, Science & Technology 19(6), (2018).

[15] Samarati, Pierangela.: Protecting respondents identities in microdata release. IEEE transactions on Knowledge and Data Engineering 13(6) ,1010-1027,(2001).

[16] Crainiceanu, Adina, Prakash Linga, Ashwin Machanavajjhala, Johannes Gehrke, and Jayavel Shanmugasundaram.: P-ring: an efficient and robust p2p range index structure. In Proceedings of the 2007 ACM SIGMOD international conference on Management of data, 223-234. (2007).

[17] Li, D -F.: A fuzzy closeness approach to fuzzy multi-attribute decision making. Fuzzy Optimization and Decision Making 6, pp.237-254, (2007).

[18] Nergiz, Mehmet Ercan, Maurizio Atzori, and Chris Clifton.: Hiding the presence of individuals from shared databases. In Proceedings of the 2007 ACM SIGMOD international conference on Management of data, 665-676.(2007).

[19] Dwork, Cynthia, Frank McSherry, Kobbi Nissim, and Adam Smith.: Calibrating noise to sensitivity in private data analysis. In Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3, pp. 265-284. Springer Berlin Heidelberg, (2006).

[20] Gkoulalas-Divanis, Aris, Grigorios Loukides, and Jimeng Sun.: Publishing data from electronic health records while preserving privacy: A survey of algorithms." Journal of biomedical informatics 50 ,4-19,(2014).

[21] Abinaya, B., and S. Santhi.: A survey on genomic data by privacy-preserving techniques perspective. Computational Biology and Chemistry 93 ,107538,(2021).

[22] Hathaliya, Jigna J., and Sudeep Tanwar.: An exhaustive survey on security and privacy issues in Healthcare 4.0." Computer Communications 153 .311-335,(2020).

[23] Chen, Bijiao, and Xianghan Zheng.: Implementing Linear Regression with Homomorphic Encryption. Procedia Computer Science 202.324-329,(2022).

[24] Kim, Miran, Yongsoo Song, Baiyu Li, and Daniele Micciancio.: Semi-parallel logistic regression for GWAS on encrypted data. BMC Medical Genomics 13, pp.1-13,(2020).

[25] Slijepčević, Djordje, Maximilian Henzl, Lukas Daniel Klausner, Tobias Dam, Peter Kieseberg, and Matthias Zeppelzauer.: k-Anonymity in practice: How generalisation and suppression affect machine learning classifiers." Computers & Security 111, 102488,(2021).

[26] Girka, Anastasiia, Vagan Terziyan, Mariia Gavriushenko, and Andrii Gontarenko. Anonymization as homeomorphic data space transformation for privacy-preserving deep learning." Procedia Computer Science 180.867-876, (2021).

[27] Sweeney, Latanya.: Achieving k-anonymity privacy protection using generalization and suppression. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10(05). 571-588, (2002).

[28] Machanavajjhala, Ashwin, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam.: l-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data (TKDD), 1(1).3-es (2007).

[29] Li, Ninghui, Tiancheng Li, and Suresh Venkatasubramanian. : t-closeness: Privacy beyond k-anonymity and l-diversity." In 2007 IEEE 23rd international conference on data engineering, pp. 106-115. IEEE, (2006).

[30] Senavirathne, Navoda.: Towards Privacy Preserving Micro-data Analysis: A machine learning based perspective under prevailing privacy regulations." PhD diss., University of Skövde, (2021).

[31] Viloria, Amelec, Nelson Alberto, and John Rhenals Turriago.: Machine Learning Techniques as Mechanisms for Data Protection and Privacy." In Proceedings of International Conference on Intelligent Computing, Information and Control Systems: ICICCS 2020, pp. 367-374. Springer Singapore(2021).

[32] Rajendran, Keerthana, Manoj Jayabalan, and Muhammad Ehsan Rana.: A study on k-anonymity, l-diversity, and t-closeness techniques.IJCSNS 17(12).172,(2017).
[33] Caruccio, Loredana, Domenico Desiato, Giuseppe Polese, Genoveffa Tortora, and Nicola Zannone.: A decision-support framework for data anonymization with application to machine learning processes. Information Sciences 613.1-32,(2022).
[34] LeFevre, Kristen, David J. DeWitt, and Raghu Ramakrishnan.: Workload-aware anonymization techniques for large-scale datasets." ACM Transactions on Database Systems (TODS) 33(3).1-47,(2008).
[35] Szarvas, György, Richárd Farkas, and Róbert Busa-Fekete.: State-of-the-art anonymization of medical records using an iterative machine learning framework. Journal of the American Medical Informatics Association 14(5).574-580,(2007).
[36] ARX - Data Anonymization Tool | A comprehensive software for privacy-preserving microdata publishing (deidentifier.org) (consulted in April 2023).
[37] UCI Machine Learning Repository: Adult Data Set (consulted in April 2023).
[38] COVID-19 patient pre-condition dataset | Kaggle (consulted in April 2023).
[39] Maini, Ekta, Bondu Venkateswarlu, Baljeet Maini, and Dheeraj Marwaha.: Machine learning–based heart disease prediction system for Indian population: An exploratory study done in South India.medical journal armed forces india 77(3), pp.302-311,Jul 2021.
[40] Haber, Anna C., Ulrich Sax, Fabian Prasser, and NFDI4Health Consortium.: Open tools for quantitative anonymization of tabular phenotype data: literature review. Briefings in Bioinformatics 23(6), pp.bbac440,(2022).
[41] Chong, Kah Meng, and Amizah Malip.: Bridging unlinkability and data utility: Privacy preserving data publication schemes for healthcare informatics.Computer Communications 191: 194-207, (2022).
[42] Song, Xing, et al.: The impact of medical big data anonymization on early acute kidney injury risk prediction." AMIA Summits on Translational Science Proceedings 2020: 617(2020).
[43] Vokinger KN, Stekhoven DJ, Krauthammer M.: Lost in anonymization—A data anonymization reference classification merging legal and technical considerations." Journal of Law, Medicine & Ethics 48.1: 228-231. (2020).
[44] Goldsteen, A., Ezov, G., Shmelkin, R., Moffie, M. and Farkash, A.: 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).
[45] Ponomareva, Natalia, et al.: How to dp-fy ml: A practical guide to machine learning with differential privacy. arXiv preprint arXiv:2303.00654 (2023).
[46] Larbi, Iyadh Ben Cheikh, Aljoscha Burchardt, and Roland Roller.: Which anonymization technique is best for which NLP task? --It depends. A Systematic Study on Clinical Text Processing. arXiv preprint arXiv:2209.00262 (2022).
[47] Kim W, Seok J.: Privacy-preserving collaborative machine learning in biomedical applications. 2022 International Conference on Artificial Intelligence in Information and Communication (ICAIIC). IEEE, (2022).
[48] Lordick T, Hoch A, Fransen B.: Anonymization of Electronic Health Care Records: The EHR Anonymizer." Computational Life Sciences: Data Engineering and Data Mining for Life Sciences. Cham: Springer International Publishing, 485-499 (2023).
[49] Carvalho T, Moniz N, Faria P, Antunes L.: Survey on Privacy-Preserving Techniques for Data Publishing. arXiv preprint arXiv:2201.08120 (2022).
[50] Lee H, Kim S, Kim JW, Chung YD.: Utility-preserving anonymization for health data publishing.: BMC medical informatics and decision making 17.1 (2017): 1-12.
[51] Amnesia Anonymization Tool - Data anonymization made easy (openaire.eu) consulted in April 2023.

[52] μ-Argus (Anti Re Identification General Utility System microdata),GitHub - sdcTools/muargus: Java interface of mu-argus (consulted in April 2023).
[53] GitHub - sdcTools/tauargus: Java interface to tauargus (consulted in April 2023).
[54] sdcMicro: Statistical Disclosure Control Methods for Anonymization of Data and Risk Estimation, sdcMicro,CRAN - Package sdcMicro (r-project.org) (consulted in April 2023).
[55] Anonimatron,GDPR compliant testing. | Anonimatron (realrolfje.github.io) (consulted in April 2023).
[56] Aircloack, Peace of Mind – Immediate Insights | Aircloak(consulted in April 2023).
[57] Mariana Cunha, Ricardo Mendes, João P. Vilela.: A survey of privacy-preserving mechanisms for heterogeneous data types. Computer Science Review,Volume 41,2021,100403,ISSN 1574-0137.