Anonymization of Data in Healthcare Sector

Akshay Bajpai 193079002
Soham Naha 193079003
R V Satwik 193079009
Mohit Agarwala 19307R004

Abstract

Health data analytics could lead to improved medical decisions, higher treatment quality, illness prevention, and cheaper costs, as well as the development of new healthcare solutions. Health information, on the other hand, is exceedingly private and is protected by legislation like the General Data Protection Regulation, which aims to preserve patients' privacy. Anonymization or the removal of patient identifiable information is the first and most important step in conforming to the criteria and embracing privacy concerns, despite being the most frequent method. In this study, we look at how to use K-anonymity, L-diversity, and T-closeness anonymization techniques on the MIMIC-III dataset. We also provide an overview of potential data-anonymization threats. We show how anonymization, while important, has an impact on the accuracy of machine learning models such as SVC, XGBoost, and Logistic Regression. We show that as the amount of anonymization increases, the accuracy of these models decreases and hence there is a need to strike a balance between protecting the privacy and maintaining the accuracy of ML models.

I. Introduction

Hippocrates wrote a famous oath promising to keep information about his patients a secret. The motivations behind protecting patient privacy have remained the same since Hippocrates' day. It promotes honest communication between the patient and physician, which is essential for quality care. It protects patients from embarrassment, discrimination or economic harm that might result from disclosure of their health status.

AI has already been introduced in Healthcare for various detection and logistics applications. For example, mining health data can lead to illness prevention, faster diagnosis, better treatment quality, and lower the expenses of a medical procedure. In order to do so, the patient needs to be confident that his or her data will not be used for unethical practices. This is the reason for a solid infrastructure to be laid using policies.

In the paper, we first summarize various ethical themes found in the literature and explain gaps in the literature. The following section covers aspects of the dataset. Anonymization models used in the experiment and a brief discussion on possible attacks on anonymization are discussed in the data collection and research section. The following sections explain the anonymization technique and the tradeoff due to the same. The conclusions and the related inferences are summarized in the conclusion section.

II. Literature survey

Four ethical themes emerged from the literature's discussion on health applications using AI (Claire, 2022). These themes were chosen based on how clearly they were described in the literature. These themes are in many ways linked (West, 2015). There is a lack of emphasis on policies for low and middle-income countries in the literature. Most of the documentation in the literature assumes a high-income country with established companies in the AI sector which is not the proper framework for AI healthcare policy in India.

A. Privacy and security

The first theme observed is privacy and security. The data can be hacked and sold for profit, leading to a loss of privacy for a large set of citizens. A case in point is the leak of HIV status and other important medical information of 35000 patients in a Diagnostic laboratory database hack in Mumbai in 2016. A subtopic of interest here is the fact that most of the software is proprietary and is usually less likely subject to scrutiny (Crawford et al., 2016).

B. Trust in AI applications

Google Deepmind received identifiable data of 1.6 million patients for improving the management of acute kidney injuries (Denton et al., 2018). Two policy issues here: how much quantity of data was needed? Further, the data is not deleted and continues to be on the company's servers. It reduces the trust of patients in the medical system, and hence the patient might not go for AI-based solutions, thereby making the process harder for further research and hindering the upliftment of the underprivileged using lower-cost medical care facilities.

C. Accountability and responsibility for use of AI technology

Many documents covered the methods to tackle the problems of misdiagnosis from learning algorithms and

how to deal with accidents in case of malfunctioning AI based machines (Decker, 2008). The literature tries to address how citizens can be protected and helped when AI-based Technologies malfunction.

D. Adverse consequences of bias

Literature deals with bias in algorithms due to the data set considered or lack of representation (Bowser et al., 2017). It is clear that inevitably every person comes with their own implicit biases, usually reflected in the models (Monteith and Glenn, 2016).

E. Gaps in the literature

A lot of literature is focused on applications of AI in Healthcare but not so much on the ethics needed to support the infrastructure. Further, it must be understood that the themes mentioned above are not independent but reinforce each other. So to summarise, there are two asymmetries observed in the literature (Paul et al., 2018), the first being the lack of literature aiming to propose policies for AI technology in healthcare. A second asymmetry is that most of the documents that propose policies for healthcare do not consider the challenges faced in low and medium-income countries and focus more on high-income countries. The unchecked policies can cause more drift in the gap between the haves, and the have-nots in the low and middle-income countries, given already existing problems at the grassroots.

When it comes to India's policies for AI in healthcare, it is surprising that there is a single mention regarding anonymization and the problem of privacy. In the white paper Vision 2035 Public Health Surveillance in India (Blanchard et al., 2021) published by Niti Aayog, only a single mention of data anonymization was seen. Further, no additional information is provided regarding the standard and thresholds to be maintained for anonymization. Shockingly, the document did not consider the de-identification of the data as a threat in the public healthcare sector. Further, there is no guideline on the amount of time the data can be stored for. An example of a poorly anonymized dataset is the Andhra Pradesh Health Insurance dataset (kag, 2007) available publicly on Kaggle to Download. Essentially the age, gender, cast, the kind of medical procedure they have undertaken, the hospital, the date of surgery, the date of discharge, and amount paid, the Village, the Mandal, and the district are revealed in the dataset. This information can be easily used to identify the actual name from another database.

III. Data Collection and Research

A. Dataset

MIMIC-III (Johnson et al., 2020) is a large, freely-available database comprising de-identified health-related data associated with over 50,000 patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. The Medical Information Mart for Intensive Care (MIMIC-III) dataset contains 53,467 unique patients' hospital admission data, lab measurements, procedure event recordings, prescriptions, hospital duration of stay, diagnostic codes, and microbiological data. We utilised the same subset of the dataset to to predict length of stay (LOS) group. As demographic data, we used age, ethnicity, admission type, marital status, insurance, religion, gender, and language.

B. Anonymization models

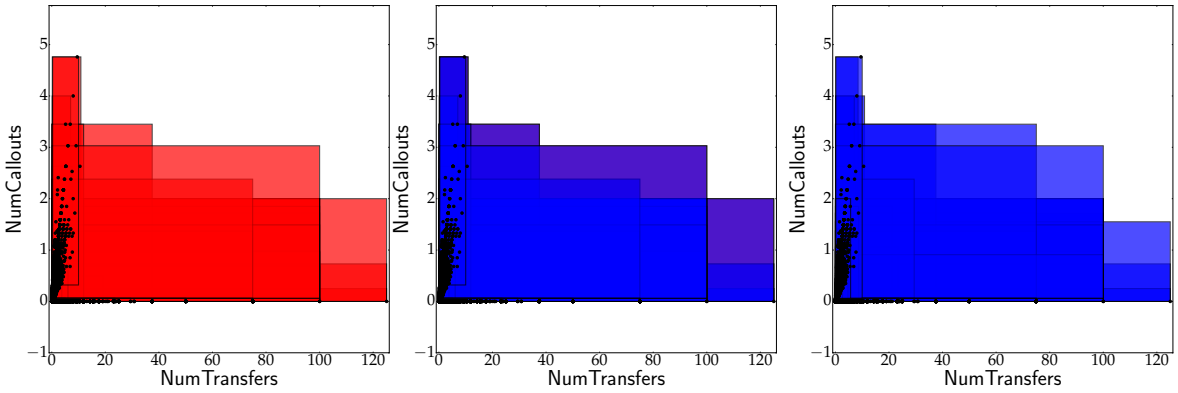
We here define three basic anonymization models proposed in the literature, namely, k-anonymity, l-diversity, and t-closeness (Olatunji et al., 2022).

- ***k*-anonymity** : k-anonymity requires that at least k individuals share the same attributes. Since QID contains fields that are likely to appear in other known datasets, k-anonymity ensures that each individual remains anonymous within their respective group (equivalence class). If $k = 10$, for example, each equivalence class should have at least 10 records that are comparable. This ensures that the attacker will not be able to determine the identity of a single document. Utility, on the other hand, depreciates when k is too high. Furthermore, the k-anonymity model's privacy is limited by the lack of significant variation in sensitive variables.
- ***l*-diversity** : l-diversity overcomes the limitations of k-anonymity by considering diversity among SAs. l-diversity ensures that there are at least l-distinct values of SA in each equivalence class.
  Example. Assume illness is the SA and the table is 3-diverse (meaning 3 distinct SA values in each equivalence class). An adversary can infer that a person has stomach-related disorders based on the attribute values of gastric ulcer, gastritis, and stomach cancer in an equivalence class where all three diseases in the equivalence class are stomach-related.
- ***t*-closeness** : t-closeness ensures that the distance between the distribution of sensitive values in each equivalence class and the original class is no more than a threshold t. Hence, a smaller value of t represents stronger privacy.
  When there are skewed attribute distributions with one sensitive value dominating, t-closeness overcomes the limitations of l-diversity.For the preceding example, ensuring t-closeness means that any equivalence class for a table that meets 3-diversity will not only contain stomach-related disorders, but also other sorts of sickness like pneumonia. Kullback-Leibler (KL) distance and Earth Mover's Distance are two popular distance functions for determining proximity (EMD).

C. Attacks on Anonymization (Olatunji et al., 2022)

In this section, we discuss about different types of attack on relational health data categorized as background knowledge attacks, linkage attacks, attribute disclosure attacks, and membership disclosure attacks.

- **Background knowledge (BK) attack**: When an adversary knows some information or QIDs



(a) Without Anonymization (b) After K-Anonymity and L-Diversity (c) After T-Closeness

Fig. 1: Mondrian Plots for different Anonymization Algorithms

about the target individual, she can reconstruct the identifiable information of the individual.

- **Linkage (LI) attack**: One where an adversary can re-identify or link a record in an anonymized dataset by combining QIDs from different sources to an individual.
- **Attribute disclosure (AD) attack**: The attacker aims to gain new information on SA. The attacker can also exploit the properties of the QIDs to estimate the SA.
- **Membership disclosure (MD) attack**: Involves an adversary aiming to infer the presence or absence of an individual in a dataset.

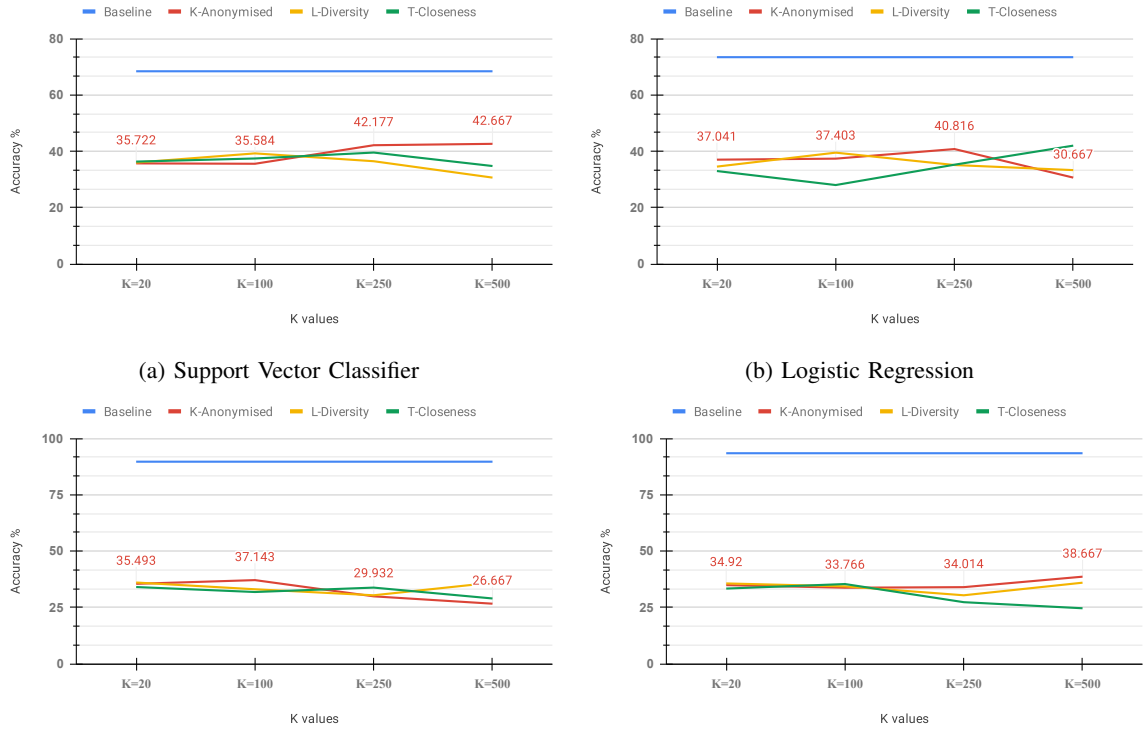
IV. Experiment Analysis

We have used MIMIC-III dataset for the purpose of analyzing the effects of anonymization on accuracy of machine learning (ML) models. We created a dummy task of predicting `LOSgroupNum` in the dataset using the remaining columns as the features. To demonstrate the true effect of anonymization of data, certain data cleaning steps were performed to ensure visible changes in the accuracy as the amount of anonymization was increased. Features like `AdmitDiagnosis`, `AdmitProcedure` were dropped as they contained inputs from medical professional in pure text format and would require NLP based procedures to make it useful for the prediction task. Also `LOSdays` was dropped as it had high correlation with `LOSgroupNum`. Now, the dataset was anonymized using 3 different techniques - K-Anonymity, L-Diversity and T-Closeness, to use for training of 4 different ML models - Support Vector Classifier(SVC), Logistic Regression(LR), Random Forest Classfier(RF) and XGBoost Classifier. Figure 1 shows the Mondrian Plots for the different Anonymization algorithms as explained previously, which we used on MIMIC-III Data. From the different subsets of combinations of features of the dataset, we choose to visualize for `NumCallouts` and `NumTransfers` features. The first Mondrian plot is for the data without any anonymization, the second one is the one anonymized using K-Anonymity combined with L-Diversity algorithm and the last one is for the data anonymized using all three anonymization algorithms: K-Anonymity, L-Diversity and T-Closeness. According to the plots, we see that as the anonymization algorithm strengthens, the overlap between the groups decreases, which suggests that de-anonymizing data from the dataset becomes harder. Table I shows the accuracy of the above models on as we increase the anonymity using the above mentioned techniques. As evident from the plots shown in Figure 2, as we introduce anonymization in the dataset, the accuracy of the ML models used here (SVC and XGBoost) decreases. As the value of K increases, the accuracy tends to increase but still very less compared to that achieved using non-anonymized dataset. Also, as we go from K-anonymization to L-diversity to T-closeness, the amount of anonymization increases and hence the accuracy achieved decreases in the same order. Hence, it is required to strike a trade-off between the accuracy we want to achieve and the amount of anonymization to be done. On one hand we will need to compromise with the effectiveness of the ML models, while on the other the privacy of the users is at risk.

V. Conclusion

In this project, we provided a comprehensive review of anonymization models and techniques applicable to health care data. We implemented deanonymization of data and demonstrated the effect of anonymization on ML model training and showed that accuracy decreases with stricter anonymization. Given the above tradeoff, It is important to propose proper guidelines on the metrics or scores used for measuring deanonymization (such as commonality between features). Further, any data made public must adhere to the said threshold or recommendation in order to avoid loss of privacy of the citizens whose data has been collected/stored.

The Government Policies must provide a guideline on the size of the dataset needed for an application based on the complexity of the model. This ensures that only representative data required is shared instead of placing the entire data at risk. The policies must also suggest the time span to which the data can be



(a) Support Vector Classifier

(b) Logistic Regression

(c) Random Forest Classifier

(d) XGBoost Classifier

Fig. 2: Accuracy vs level of Anonymization for the different machine learing models

	K=20				K=100				K=250				K=500			
	SVC	**LR**	**RF**	**XGB**	**SVC**	**LR**	**RF**	**XGB**	**SVC**	**LR**	**RF**	**XGB**	**SVC**	**LR**	**RF**	**XGB**
K-Anonymized	35.722	37.041	35.493	34.92	35.584	37.403	37.143	33.766	42.177	40.816	29.932	34.014	42.667	30.667	26.667	38.667
L-Diversity	36.014	34.578	36.071	35.669	39.276	39.535	33.075	34.367	36.486	35.135	30.405	30.405	30.667	33.333	36	36
T-Closeness	36.357	33.003	34.07	33.384	37.463	28.024	31.858	35.398	39.568	35.252	33.813	27.338	34.783	42.029	28.986	24.638

TABLE I: Variation in accuracy in % of ML models with different anonymization techniques and K values.

stored and the accountability in case the data breach occurs.

References

2007. Andhra pradesh health insurance data.

James Blanchard, Reynold Washington, Marissa Becker, N Vasanthakumar, K Madan Gopal, and Rakesh Sarwal. 2021. Vision 2035 public health surveillance in india.

Anne Bowser, Michael Sloan, Pietro Michelucci, and Eleonore Pauwels. 2017. Artificial intelligence: A policy-oriented introduction. *no. November*.

Urvashi Claire. 2022. Artificial intelligence for healthcare: Insights from india.

Kate Crawford, Meredith Whittaker, Madeleine Clare Elish, Solon Barocas, Aaron Plasek, and Kadija Ferryman. 2016. The ai now report. *The Social and Economic Implications of Artificial Intelligence Technologies in the Near-Term*, 2.

Michael Decker. 2008. Caregiving robots and ethical reflection: the perspective of interdisciplinary technology assessment. *AI & society*, 22(3):315–330.

S Denton, E Pauwels, et al. 2018. There's nowhere to hide: artificial intelligence and privacy in the fourth industrial revolution. *Wilson Center Policy Report*, pages 10–11.

Alistair Johnson, Tom Pollard, and Roger Mark. 2020. MIMIC-III clinical database.

Scott Monteith and Tasha Glenn. 2016. Automated decision-making and big data: concerns for people with mental illness. *Current Psychiatry Reports*, 18(12):1–12.

Iyiola E Olatunji, Jens Rauch, Matthias Katzensteiner, and Megha Khosla. 2022. A review of anonymization for healthcare data. *Big Data*.

Yesha Paul, Elonnai Hickok, Amber Sinha, Udbhav Tiwari, S Mohandas, S Ray, and PM Bidare. 2018. Artificial intelligence in the healthcare industry in india. *The Centre for Internet and Society, India*.

Darrell M West. 2015. What happens if robots take the jobs? the impact of emerging technologies on employment and public policy. *Centre for Technology Innovation at Brookings, Washington DC*.