Effectiveness and Limitations of Existing Techniques for Privacy Preservation Data Mining (PPDM) for Medical Data

Amol R Kulkarni
Research Scholar,
Vignan University Vadlamudi,
Guntur, India

Dr. Kamepalli Sujatha
Associate Professor,
Vignan University Vadlamudi,
Guntur, India

Abstract:

Background & Objective: This survey paper examines the application of Federated Learning (FL) and secure Multiparty Computing (MPC) considering medical data privacy, further provides an overview of FL and MPC techniques and discusses their strengths and weaknesses. This also covers other techniques like Homomorphic Encryption, data masking, differential privacy for its efficiency, and limitations.

Methods: Eligibility Criteria: For this survey PRISMA[1] framework was employed and popular electronic databases like IEEE, Computers and Security, Bioinformatics, and Google Scholar were scanned along with government websites and research web pages. Papers published between 2018 and 2023 and written in English are considered for the survey. This survey further provides direction towards the future research and potential challenges in the deployment of these techniques at a scale.

Results: The search was restricted for the period between 2018 and 2023. Initially, ~100 papers were shortlisted, and after a though review of each paper, finally selected ~35 for this work. However, papers from earlier years are also included as they are found to be relevant for this study. Fig4 below describes the article selection process. Apart from the research papers, government websites are also referred for information on various laws, regulations, and compliance for the privacy of the patient's information. Acts like HIPPA[7], GDPR[8], PDP[9] are reviewed thoroughly to ensure all the aspects are covered to present this survey.

Conclusion: The information used in the data mining process comes from data, and such data include personal data about people. Since knowledge is derived from data, the main goal of privacy-preserving data mining is the development of algorithms to conceal or protect certain sensitive information so that it cannot be disclosed to unintended users, hackers, or intruders. Various techniques like homomorphic encryption, differential privacy, data masking, data aggregation, multiparty computation is studied, one single technique may not be sufficient to provide an end-to-end solution. In the future work, we would like to explore on multiple techniques to build a secured ML model.

Key Words: Privacy Preservation Data Mining, homomorphic encryption, federated learning, secured multiparty computation.

I. INTRODUCTION

The usage of Electronic Health Record (EHR) systems has been increasing at a rapid rate in recent years. EHR[2] systems have come a long way since they were first introduced in the 1960s. These systems have evolved from being a very basic patient management system to systems which offer advanced capabilities like clinical documentation (patient medical history, diagnosis results etc.), decision support, revenue cycle management and patient engagement. The use of EHRs is becoming a standard practice in healthcare organizations as they have proven to provide better and more benefits over paper-based systems, which includes improved patient safety, improved care coordination, and increased efficiency and overall, providing Value Based Care (VBC).

Today there are various EHR systems available in the market, to name a few Epic, Cerner, Allscripts etc. Healthcare providers (hospitals) have deployed EHR systems to suit their needs, today these systems are catering to patients in a specified geography and for specified departments. These systems store patient data, which include sensitive information such as patients' demographic data, medical history, medications, treatment plans, diagnosis code (ICD/CPT codes), laboratory and test results.

Patient data provides an immense opportunity for building various decision support systems and enable data mining to provide a better population health and value-based care. With the rise in the number of medical problems across the globe, centralizing the patient data to provide a personalized treatment and medicine will be a key requirement for the future. Medical and other researchers are working on patient clinical & medical data for value discovery. As this data is restricted to the individual healthcare provider network, generalizing on the overall outcome of the population health decision and data mining is difficult. To overcome this challenge, the data needs to be centralized i.e., data can be shared over internet to a central storage location and processed to build a robust system for clinical decision making. To enable centralization of the data, there are two major challenges to be addressed one, security for the healthcare data and two, privacy for the healthcare data. Preetha[3] in her article, articulates the need to AI in the healthcare sector, how the AI would revolutionize the personal treatment and medication, effectiveness of drug and

drug research. This can further extend to address another key challenge of mental wellness and behavioral health.

Healthcare Data Security:

Even while healthcare organizations produce, store, and transfer vast volumes of sensitive data in order to deliver effective and appropriate care, they are nonetheless at risk of data breach and loss due to a lack of technical assistance and weak security. Information security has grown in importance for our society as a result of the Internet. The healthcare sector is especially susceptible to breaches of publicly revealed data due to the intricate structure of big data. In actuality, hackers could employ technology and data mining procedures to find and reveal private information to the public, leading to data breaches and data infringements. Although implementing security measures is a difficult procedure, security controls are getting smarter because of new security technologies. Therefore, it is crucial for businesses to implement health data security solutions that safeguard crucial resources and highly sensitive data while adhering to regulatory requirements for the provision of healthcare. Multiple technologies, including firewalls and encryption are in place to safeguard against data breaches caused by weak points in the company's technological and database systems.

According to the HIPPA Journal[4], between the years 2009 and 2022 there have been more than 5K healthcare data breaches reported to HHS Office for Civil Rights, which involved more than 500 records per incident. These breaches have resulted in the unveiling or improper disclosure of approximately 382M healthcare records, which is almost 1.2 times of US population. In 2018 alone, data breaches in healthcare involving more than 500 records were reported on average daily, and this rate has more than doubled in just five years. Furthermore, in 2022, the data breaches have almost doubled daily.

Rahman et al.[5] identifies and analyzes the deep learning systems built for COVID-19 in the healthcare IoT devices. The authors discuss the vulnerabilities in the deep learning models, including the “data poisoning attacks”, model stealing, and backdoor attacks. The paper also highlights the risks posed by medical IoT devices and the possible consequences of an attack on such systems. The authors propose several countermeasures to mitigate the identified threats, including the use of secure deep learning algorithms, secure training data, and secure communication protocols. They also recommend the use of intrusion detection systems and anomaly detection mechanisms to monitor the system's behavior and detect any suspicious activities. Overall, the paper provides insight into the potential security risks in healthcare IoT devices and proposes solutions to mitigate those risks. It highlights the requirement for strong security measures to safeguard the privacy, availability, and integrity of medical data and systems.

Newly developed blockchain technology may be able to overcome the biggest security problems in healthcare. Decentralized storage, cryptography, and smart contracts, among other features, give businesses a foundation for improving data security and also preserving data accuracy and avoiding unauthorized access to or updating of patient information. Jie Xu et.al.[6] proposes a privacy-preserving blockchain-based scheme called "HealthChain" that addresses the challenges of privacy, security, and interoperability in handling large-scale health data. HealthChain uses a private blockchain to secure the data and a distributed key management system to provide access control, with fine-grained access permissions and dynamic revocation. The data schema is compatible with existing healthcare data standards, and a consensus mechanism ensures data integrity. HealthChain aims to be a solution for the challenges of handling health data by ensuring privacy, security, interoperability, and data integrity.

A. Privacy Preserving:

Other than data security and theft, another challenge that invades the medical data safety is preserving the privacy of the (patient) data. Big data accessibility has created previously unheard-of potential to enhance the effectiveness and caliber of healthcare services, notably in terms of enhancing patient outcomes and cutting costs. EHR data is published so that it can be used by government agencies and the healthcare sector. Large-scale statistical analysis (such as the research of illness correlation), clinical decision-making, therapy optimization, clustering (building patient cohorts), and census surveys are a few notable examples. Sharing patient data with government agencies and other healthcare organizations poses data privacy challenges. Over a period, multiple policies and guidelines are developed to ensure patient data privacy and for the data to be published for research. US “Health Insurance Portability and Accountability Act (HIPAA)”[7] , “General Data Protection Regulation (GDPR)”[8] of EU and the “Personal Data Protection Act (PDPA)”[9], provides regulations and guidelines to protect the privacy of the patients’ data and making the data available for research for government and other healthcare organizations.

Medical data analysis often involves sensitive patient information, which requires privacy preservation techniques to ensure the confidentiality of the data. The main intent of privacy preserving data mining is to enable valuable insights to be extracted from data while still ensuring that individuals' personal information remains protected.

Common steps involved in data mining are collecting the data, pre-processing, EDA (exploratory data analysis), feature engineering, building & validating data models and finally model deployment for consumption. These steps are straight forward when working with data within a single organization network, but when we aim to generalize the outcome over multiple entities maintaining data privacy and data publishing becomes critical part, this paper focuses mainly on surveying various Privacy Preserving Data Mining (PPDM) techniques specifically for medical data. PPDM is utmost important when it comes to medical data as this type of data can contain highly sensitive information about an individuals' health and other medical conditions. Below are some of PPDM techniques that can be applied to medical data:

1. Homomorphic Encryption (HE)
2. Differential Privacy (DP)
3. Data Masking or Perturbation

4. Data Anonymization
5. Secure Multiparty Computation (MPC)

Homomorphic Encryption (HE): By using HE techniques, computations can be executed on encrypted data without requiring the data to be decrypted. This can be useful in medical data mining as it allows researchers to analyze encrypted medical data without compromising the privacy of patients. The computation's results are still encrypted. This technique is becoming more and more popular since it maintains the confidentiality and privacy of sensitive data. Depending on the type of the operation to be performed on the data, there are various HE techniques, including Partly Homomorphic Encryption (PHE), Slightly Homomorphic Encryption (SHE), and Completely Homomorphic Encryption (FHE). The four steps of the homomorphic encryption process are key generation, encryption, decryption, and evaluation. Although, HE is one of the simplest and effective way for privacy preservation of data, there are certain limitations that needs to be considered while application, these limitations include computational complexity, limited functionality, key management, encryption & decryption overhead, and sensitivity to noise.

Differential Privacy (DP): The concept of Differential Privacy (DP) is employed in privacy preservation of data analysis to safeguard the data about individuals' privacy while facilitating useful statistical analysis and data mining on the data. The basic notion behind DP is to introduce noise into the data in a manner that makes it challenging to determine if an individual's data is present or absent within the dataset. While DP is a promising technique for protecting the privacy of medical data, it has certain drawbacks that need to be considered. The addition of noise to the data can lower the precision of the analysis and data models, elevate the computational burden, and restrict the scalability.

Data masking or perturbation: This technique is commonly used in medical data to protect sensitive information about patients. For example, instead of releasing the exact age of patients, data can be transformed by adding random noise to age values or by rounding age values to the nearest integer or moving the data in past or future by a constant (undisclosed) number. While data masking or perturbation is a useful technique for preserving the privacy of medical data, it has some limitations that need attention. These limitations include a reduction in data quality, difficulty in finding the right balance between privacy and data quality, vulnerability to re-identification, difficulty in maintaining consistency, and limited scalability.

Data anonymization: This a technique is used to protect the privacy of individuals by removing personally identifiable information (PII) from the data. Although data anonymization is commonly used in medical data to protect sensitive information about patients, it has some limitations that must be considered. These limitations include the risk of re-identification, loss of data, difficulty in maintaining data quality, difficulty in combining datasets, and limited scalability.

Data aggregation: This technique involves combining data to produce statistical summaries. For example, instead of releasing the number of patients with a specific disease, a hospital may release the total number of patients treated for all diseases. This is a common and relatively easy technique used to build data warehouses from an OLTP system. This will ensure that the granularity of the data is preserved and only the aggregated information is shared with the end users for any data analytics or statistical modeling tasks. Though data aggregation is one of the easy ways to preserve privacy of the data, it may not be the best technique as the granularity of the data is lost and can result in information loss resulting in poor ML models.

Secure Multi-Party Computation (SMPC): Secure MPC is a methodology that allows 2 or more parties to compute a required output using their private inputs without revealing their inputs to each other. In the context of medical data analysis, secured MPC can be utilized to perform computations on sensitive patient data without having risk of exposing the data to unauthorized parties.

The so-called millionaire's [10] [11] problem is the classical example of MPC. Here, a group of millionaires decide to split the bill among themselves at a lunch meeting at a pricey restaurant. They do not wish to share their true wealth with one another, though. This serves as an illustration of a safe multi-party computation. The function to be computed is as follows. The inputs are the values x_i, which represent the wealth of each party.

$$f(x_i \cdots x_n) = i \; where \; x_i > x_j \; for \; all \; i \neq j$$

Another example of the MPC is the voting process. This entails computing the results of each party's vote without revealing whose party cast it. The sender, the receiver, and the adversary are the three people involved in a multi-party computation called encryption. Only the recipient has an output, and only the sender has an input (the message to be encrypted) (which is the message when decrypted).

MPC has been shown to be effective in preserving privacy in medical data analysis. For example, researchers have used MPC to perform genome-wide association studies on sensitive patient data without compromising patient privacy. Additionally, MPC can be used to perform data analysis across multiple healthcare providers, enabling data sharing without sacrificing patient privacy.

One limitation of MPC is its computational complexity. MPC requires significant computational resources, making it challenging to perform computations on large datasets. Additionally, MPC assumes that all parties are honest-but-curious[10], meaning that they will follow the set protocol but may attempt to learn information about other parties' inputs. If a party is malicious, they can compromise the security of the MPC protocol.

Fig1 below depicts a generic Multiparty Computation (MPC) diagram which involves *k*-parties. In this, each of the *k*-party shared its data with a common computation function or block which eventually processes and provides a desired

output. In this scenario the data shared by any party is not known to the other party, hence securing the data privacy.

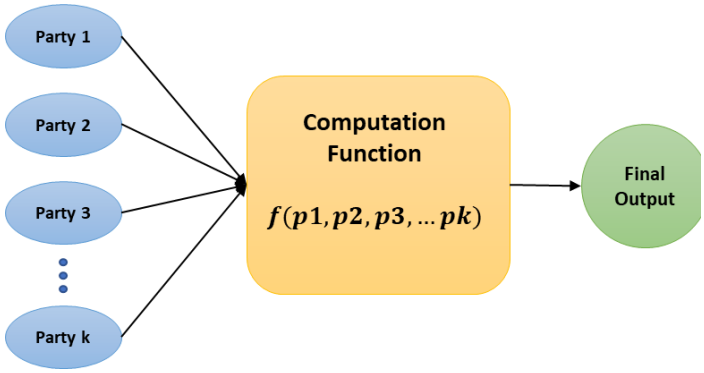

Fig. 1. Multiparty Computation with *k* parties

Federated Learning (FL): FL was first introduced in 2017 by Google[12]. FL is a technique that allows multiple parties to train an ML model without having to share their data. In FL, each party trains a local model on their data, and these locally trained models are then aggregated to form a common global model. In the context of medical data analysis, FL can be used to train machine learning models on sensitive patient data without exposing the data to unauthorized parties.

FL has been shown to be effective in preserving privacy in medical data analysis. For example, researchers have been using FL to train machine learning models on electronic health records without compromising patient privacy. Additionally, federated learning can be used to perform data analysis across multiple healthcare providers, enabling data sharing without sacrificing patient privacy.

One limitation of federated learning is that it assumes that all parties are trustworthy, i.e., they will follow the procedure and not attempt to learn information about other parties' data. If a party is malicious, they can compromise the security of the federated learning protocol. Additionally, federated learning requires significant communication between parties, making it challenging to implement in environments with slow or unreliable communication networks.

In the Fig2 depicted below by Oh et al.[13] provides a clear distinction in various ML model building techniques. In this, (A) Local ML: Entities build their own local ML models with their own data and computation power (B) Centralized ML: Here the data from participating entities is pooled and global ML models are built. Mostly these entities belong to the same network and hence data privacy may not be a concern. (C) Distributed ML: The participating entities receive the data and computing tasks from the central aggregator to construct a global ML model. (D) FL can be used to train machine learning models on sensitive patient data without exposing the data to unauthorized parties.

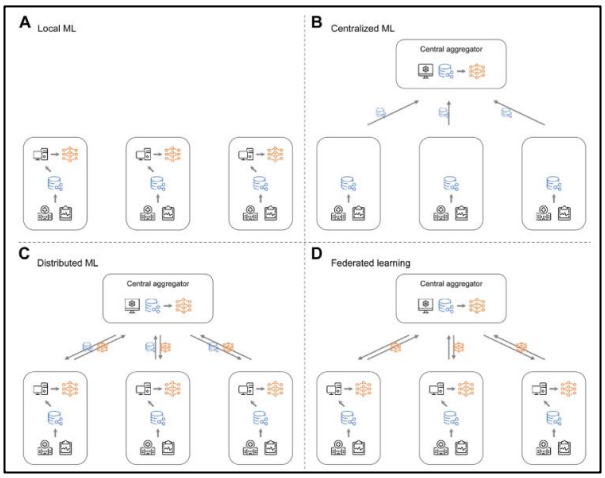

Fig. 2. Various ML model building techniques. [13]

Federated Learning Architecture for Healthcare[14]

Fig3 below depicts a federated learning architecture for a healthcare system. In this setup, each of the hospitals (Hospital 1 to 3) train their own ML model on their respective data and the model is eventually shared with the central aggregation server for model aggregation, which then creates a consensus data model that can be finally used for predictions or classifications. In this, the data is privately secured with each hospital, and only the final model is shared which ensures data privacy.

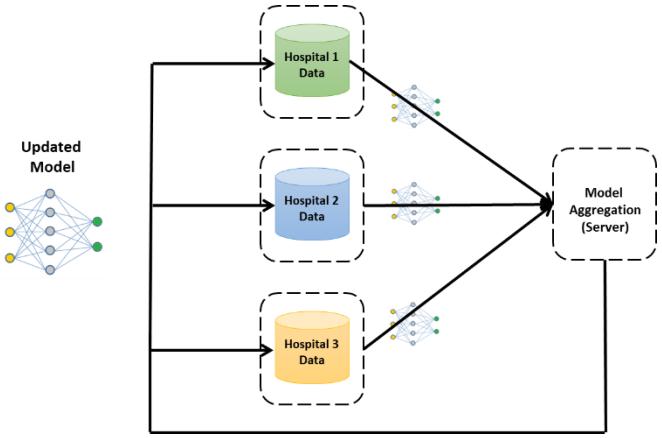

Fig. 3. Federated Learning Architecture for Healthcare[14]

TABLE I. BELOW SUMMARIZES SOME OF THE KEY FINDINGS OF THE LITERATURE SURVEY.

Paper Title	Model/Method	Data Set	Evaluation metrics	Comments
"Application of Homomorphic Encryption in Machine Learning"[15]	Liner Regression Homomorphic Encryption using Pailler cryptosystem.	General	Comparing the linear regression model build using encrypted and plain text	This technique works better when the data relationship is linear. However, the homomorphic capabilities of the Paillier cryptosystem are limited, which can make it difficult to perform complex operations on medical data. Other challenges can be security of the (encrypted) data and adherence to local government regulations.
"Privacy-Preserving Linear Regression on Distributed Data by Homomorphic Encryption and Data Masking"[16]	Privacy Preserving Linear Regression (PPLR) "Paillier" homomorphic encryption	6 UCI Dataset: Auto MPG Wine Quality Bike Sharing Forest Fires Communities and Crime YearPredictionMSD	Compared with other protocols.	This implementation has asserted that the combination of Homomorphic encryption along with data masking to perform a linear regression across multiple servers is a viable and more accurate solution. This can further be evaluated on other real data in the healthcare domain.
"Privacy Preserving Deep Learning Using Secure Multiparty Computation"[17]	Deep Learning	MNIST	Compared with other methods.	This paper implements a privacy preserving technique using multiparty computation. This technique needs further evaluation on more complex data sets.
"Privacy-preserving SVM on outsourced genomic data via secure multi-party computation"[18]	Support Vector Machine (SVM)	Genomic Data (HIV Sequence data) Synthetic data sets	Cross validation Comparison between data sets	This implementation includes both real data and synthesised data. SVM in multiparty computation has shown promising results and can be extended to medical data.
"Privacy-preserving self-serviced medical diagnosis scheme based on secure multi-party computation"[19]	Euclidean distance and similarity function	Experimental data	Model comparison	This paper describes an implementation of a patient-provider (hospital) communication on a secured line. The diagnosis is based on the Euclidean distance and similarity of pre-built disease vectors. This paper does not discuss about any specific ML algorithms to make predictions.
"Privacy-Preserving Feature Selection with Secure Multiparty Computation"[20]	Data Pre-processing (Feature selection) Using Gini impurity	Real world data set	Model comparison	This paper describes approaches for data pre-processing steps with MPC by making use of the Gini impurity. This is a useful implementation before any ML Model is built in a federated way. Feature selection with MPC can further be explored for healthcare data for data privacy
"Secure Multi-Party Computation based Privacy Preserving Data Analysis in Healthcare IoT Systems"[21]	Bayesian Deep Learning (Federated Learning)	Medical IoT Device data from multiple hospitals	Comparison with base model (without federated learning)	This paper presents a Bayesian NN model built in a Federated way on the IoT devices. This new model is said to be performing equally good as the base model. This provides an opportunity to extend the same to the other healthcare data and securing the data privacy.

"Privacy-preserving multi-party computing for K-means"[22]	k-means clustering	Seeds (grains) – 210 samples with 7 attributes Wine Data set	Comparison with base model	The privacy-preserving MPC technique for K-means clustering (PPMCK) seems promising and can be applied to the medical data to build patient cohorts in a privacy mode.
"Mainzelliste SecureEpiLinker (MainSEL): Privacy-preserving record linkage using secure multi-party computation"[23]	Privacy Preserving Record Linkage (PPRL)	Medical data	Benchmarking with other methods	This paper introduces a fault-tolerant Record Linkage method, which can be implemented to build a patient lineage across multiple hospitals without compromising on the privacy (PII) data of the patients. This can further be extended to build ML models for better prediction without any loss of data privacy.
"Secure Multiparty Computation for Synthetic Data Generation from Distributed Data"[24]	Novel approach to generate synthetic data from distributed DBs	Synthetic Data	NA	Introduced a novel way to generate synthetic data using MPC. This method is useful for experimentation of various methods in ML and data analytics.
"EasySMPC: a simple but powerful no-code tool for practical secure multiparty computation"[25]	Arithmetic Secret key (Secure multiparty computation)	Medical data. Patients' data for Phenylketonuria (PKU)	Comparison between multiple data sharing scenarios.	No-code, desktop-based application to perform joint calculations. This creates a dependency on the third-party software installation which might not be feasible in certain scenarios. Also, this tool has a limitation of addition and subtraction which would be a major limitation.
"Sequre: a high-performance framework for secure multiparty computation enables biomedical data sharing"[26]	Multiparty Computation algebraic operations	Genomics Data	Comparison between bioinformatics tasks.	Python based tool, which has the capability of converting the simple syntax of python code an equivalent MPC code, which can then be distributed to the group of computing parties. This tool can be implemented for the healthcare data to build a generic common code across all the parties without having to worry about the MPC implementation.
"Proportionally Fair Hospital Collaborations in Federated Learning of Histopathology Images"[27]	Federated Learning for fairness of the data.	Medica data: Histopathology Image dataset Non-medical data: MNIST, FMNIST	Comparison between multiple data sets for speed and accuracy. (FedAVG, FedSGD)	Proposed a novel framework to improve the model fairness using federated learning. (Prop-FFL). FedSGD (Federated Stochastic Gradient Descent) method can be applied on the medical data for evaluate the model weights which can further be passed on to a deep learning model to train the data based on the weights.

II. LITERATURE SELECTION PROCESS AND SURVEY:

A. Literature Selection Process:

The objective of the survey is to identify work and techniques that are available for privacy preservation of individuals' data, in general and for medical data, that includes techniques like secured multi-party computation and federated learning. The article selection process for this study is influenced by the "Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA)"[1]. The rest of this section describes a methodical approach in selecting the research articles and the search criteria used to search renowned databases, like IEEE, Computers and Security, Bioinformatics, and Google Scholar. The search was restricted for the period between 2018 and 2023. Initially, ~100 papers were shortlisted, and after a though review of each paper, finally selected ~30 for this work. However, papers from earlier years are also included as they are found to be relevant

for this study. Fig4 below describes the article selection process. Apart from the research papers, government websites are also referred for information on various laws, regulations, and compliance for the privacy of the patient's information. Acts like HIPPA[7], GDPR[8], PDP[9] are reviewed thoroughly to ensure all the aspects are covered to present this survey.

Keywords used for the search are as follows:

"Privacy Preserving Data Pre-processing in healthcare", "Privacy Preserving Data Mining in healthcare", "Muti-party Computation in healthcare", "Federated Learning", "Multi-party Computation and Federated Learning", "Homomorphic Encryption", "Healthcare data security".

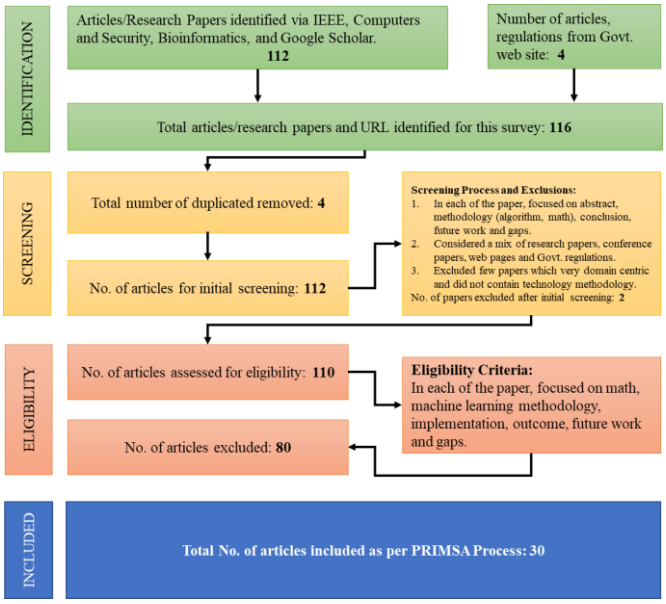

Fig. 4. Research paper selection process flow[1]

B. Literature Survey:

Behera et al.[15] have implemented the Paillier cryptosystem which follows additive HE property. They started with a plain text which is encrypted using HE method and this encrypted data is then fed as input to a linear regression model. This model takes in encrypted data as input and outputs encrypted form. At the output side the same homomorphic decryption method is used to decrypt the data. It is discovered that if addition and multiplication are applied to encrypted text, the results obtained after decryption are identical to those obtained after applying the same operation to plain text. There are several limitations of this application. Apart from computational overhead, public key size, the Paillier cryptosystem can only perform limited homomorphic operations, such as addition and multiplication of encrypted numbers. It cannot perform more complex operations like division or exponentiation, which limits its use in some applications.

Li et al.[16] in this study, describe the homomorphism operations in ciphertext space and present a novel HE framework over non-abelian rings. Based on the Conjugacy Search Problem, the approach can establish one-way security. Afterwards it was recommended to use homomorphic encryption over a matrix ring. It enables rapid homomorphic comparison of ciphertexts without requiring the intermediate outcomes of any ciphertext operations to be decrypted and supports real number encryption based on the homomorphism of a 2-order displacement matrix coding function. A data ciphertexts environment was used by the author to achieve training and classification of machine learning models which used privacy preservation of the sensitive data. The analysis demonstrates the effectiveness of the suggested techniques for homomorphic operations and encryption/decryption. In the future work the author propose to improve the efficiency of the proposed encryption scheme, targeting primarily on reducing the ciphertext expansion rate. Further, the future work will consider the privacy preservation for machine learning in more complex environment.

Bonawitz et al.[17] presented secure aggregation for federated learning in their study. According to the author, their protocol is tolerant to client dropouts. To prevent access to local models, they use blinding with random values, Shamir's Secret Sharing (SSS)[18], and symmetric encryption. The drawback is that each iteration of their aggregation needs at least 4 rounds of communication between each of the client and the aggregator. For clients who are frequently connected through WAN and have constrained resources, this results in a large overhead.

Reichert et.al[19] proposes a decentralized contact tracing system that aims to preserve users' privacy while still enabling effective tracing of COVID-19 cases. The system uses GPRS signals to detect potential exposures between devices, and it stores data on a user's phone rather than a central server, reducing the risk of data breaches. The system also uses cryptographic techniques to protect the privacy of users and their contacts while still allowing health authorities to track potential infections. The authors argue that this approach is more privacy-preserving and effective than centralized contact tracing systems that rely on GPS tracking or server-based storage of sensitive data. However, this technique is not effective when dealing with large data especially because most client computers are mobile phones with restrictions on connectivity, power, and computation.

Suhel et.al [20], in their paper demonstrated a privacy preservation technique which is based on MPC. In this technique a secret shared key is created which solves the privacy issue for the data used to train the model. The author has built a neural network by utilizing the MNIST data set to demonstrate the implementation. This technique needs further evaluation for more complex data sets and other security issues.

By employing garbled circuits, Dong et.al[21] demonstrated that secure MPC is a practical solution for dealing with medical use cases that call for cross-institutional data sharing and collaboration. The author proposed two secure MPC techniques that were evaluated using sizable and realistically simulated datasets and are based on Yao's garbled circuits. Authors also demonstrated that the Cuckoo hashing made their protocol run extremely fast. Lastly, these protocols do away with the requirement for a reliable third-party "honest

broker" to facilitate the linking and exchanging of information. Authors exhibit confidence that this developed protocol using the garbled circuits is real world ready and can be implemented to build and deploy clinical informatics.

Jayabalan et al.[22] in their article discusses three privacy-preserving techniques for medical data, l-diversity, k-anonymity, and t-closeness. k-anonymity aims to ensure that each record in a dataset cannot be linked to a unique individual by grouping similar records together. By assuring that each group of related records contains a diverse set of sensitive qualities, L-diversity enhances k-anonymity. By ensuring that the distribution of sensitive qualities in each group matches that in the entire dataset, t-closeness enhances l-diversity. The article provides a detailed comparison of these three techniques, along with their strengths and weaknesses in protecting medical data privacy.

Chen et al.[23] proposed a privacy preserving support vector machine (SVM) model on HIV sequence data to predict the effectiveness of a certain antiviral drug. This protocol is based on the secure MPC, that utilized the semi-honest adversary model and oblivious transfer. This was run on combination of data from multiple sources and without compromising the data privacy. In this study two different solutions based on Beaver's multiplication triples over arithmetic secret sharing and the Goldreich-Micali-Wigderson (GMW) protocol over boolean secret sharing are proposed. Furthermore, boolean circuits were discovered to be the best in terms of execution speed. Both of these solutions have proved to be quite successful.

Li et al.[24] in their paper proposes a self-service diagnosis method based on HE and secure MPC. In this method, the first step involves data encryption by the registered patient and then the patient sends this encrypted data to the hospital medical server, which then, based on the similarity vectors, performs a diseases diagnosis, and sends the treatment method back to the patient. This works on the principle of privacy preservation of patients' health data and diagnosis provided by the hospital.

Qiu et al.[25] in their study proposed a novel method for performing linear regression using privacy preservation on horizontally partitioned data. The technique is based on the Paillier HE and a new data masking technique. C++ and Java were the choice of programming language to implement the protocol that uses 6 real data sets from UCI. The approach discusses how multiple servers (data providers) can collaborate to perform a linear regression without having to disclose their own data i.e., maintaining data privacy. The authors assert that the combination of HE along with masking of data can realize greater security and accuracy of outcome.

Li et al.[26] in their study propose a first MPC based protocol for feature selection based on filter methods. Gini Impurity is used as a feature scoring method. The author asserts that the proposed feature selection technique based on MPC and Gini Impurity, improves the accuracy of the classifier without leaking the feature value. The proposed method is MS-GINI (mean-split Gini score), that avoids the need for sorting by selecting the mean of the features values as the split point. Further, the authors confirm that this protocol can be used in combination with any other feature selection methods.

The primary goal for Şahinbaş et al.[27] in their study is to build DL models for various hospitals using the pertinent patient records. For IoT medical equipment, federated learning-based privacy protection along with multi-party computation has been proposed. This study shows that as the number of clients varies, the accuracy of the DL models' prediction performance is very equal, and the accuracy of the Bayesian neural network (NN) model is generally identical to the accuracy of the base model's prediction performance. In their upcoming research, the author proposes to implement partial and a little HE based techniques to safeguard private and sensitive information in IoT.

The inability of HE to handle division operations and ciphertext value size comparison is one of its drawbacks. To address this, Fan et al.[28] in their study proposed a MPC technique for K-means clustering (PPMCK), which can preserve the data privacy at the local sider as well in the cloud. This technique uses HE to protect data privacy, which overcomes the above-mentioned drawback and makes it work smoothly. The data privacy-preserving protocol in this work is implemented on both the local and cloud sides, and it uses the PPWAP protocol to protect data during local contact. The authors assert that this technique improves computing efficiency and ensures data preserving privacy.

Record linkage from various sources is an effective way to collate data for statistical analysis and data mining. Especially with medical data, linking the patient data from various hospitals has great value and potential. But considering the privacy laws around PII (personally identifiable information) data, this linkage based on name, data of birth, SSN or zip codes is a challenge. In their study Stammler et al.[29] propose a novel method for privacy-preserving record linkage (PPRL) technique which prevents any of the data leakage issues and still allowing for record linkage. The software MainSEL (Mainzelliste) is used in the study and employed secured multi-party computing.

Pereira et al.[30] proposes a novel approach for generating differentially private synthetic data from distributed databases based on MPC. The three main contributions are as follows (1) Present a method for creating synthetic data from networked databases using Secure Multiparty Computation (MPC) protocols, which simulate a trusted curator and are executed by two or more computer parties. (2) To generate synthetic data with DP guarantees based on real data coming from numerous data holders and without relying on a single point of failure, modified the Multiplicative Weights with Exponential Mechanism (MWEM). (3) Using the exponential process, suggested an MPC procedure for safe sampling from distributed data.

Wirth et al.[31] have developed a simple "no-code solution" called as EasySMPC. This tool is very easy to adapt by scientists who do not possess IT skills and does not have any specific infrastructure requirements for doing secure collaborative calculations on biological data (Phenylketonuria (PKU)) using SMPC protocols. This tool was developed in Java Programming language and is available as desktop

software and does not need any additional software. This tool utilizes the Arithmetic Secret Sharing approach of SMPC (Secure Multi-Party Computation) to securely add up predetermined sets of variables across multiple parties in two communication rounds. Additionally, it incorporates this technique into a user-friendly graphical interface. Though the author asserts that this tool is scalable and reliable, addition and subtraction is one of the big limitations. Apart from this, the tool uses email as basic communication medium, when dealing with large number of messages, the email server may tag this as spam.

Smajlović et al.[32] introduces Sequre framework in this paper. This is a powerful and user-friendly framework for creating effective MPC applications. The performance of MPC applications is greatly enhanced by this tool's set of automatic compile time optimisations. The Python programming language, which permits quick application creation, is what makes this possible. For demonstration of efficacy, this framework was implemented on multiple domains, some of them are medical genetics, pharmacogenomics, metagenomics etc. The results showed that effectiveness and scalability of the framework. This can further be extended to other healthcare datasets without worrying much about the implementation of the MPC.

Xiao et al.[33] in their study proposes Multi-party Computation for Drug-drug interaction (MPCDDI). In this, they built a neural network model for DDI predictions by utilizing drug-related feature data from multiple universities. MPCDDI makes use of secret sharing technologies. With MPCDDI, all deep learning and data transmission activities are incorporated DDiinto secure MPC frameworks, allowing pharmaceutical institutions to collaborate effectively without disclosing confidential drug-related information. The author asserts that MPCDDI outperforms the other five baselines and performs on par with matching plaintext partnerships. More intriguingly, MPCDDI performs noticeably better than approaches that rely solely on institution-specific private data. In conclusion, the MPCDDI is a useful platform for fostering cooperative and private drug development.

Maryam Hosseini et al.[34] in their study proposes a model called as Proportionally Fair Federated Learning (Prop-FFL), which is expected to improve the "fairness" among the participating hospitals. Prop-FFL is bases on novel optimization method to decrease the performance variations among the participating hospitals. This model was validated using two histopathology datasets and two general (non-medical) data sets (MNIST, FMNIST), authors assert promising outcome from these experiments in terms of learning speed, accuracy, and fairness. Prop-FFL is built on top of FedSGD (Federated stochastic gradient descent), proposed by McMahan et al.[35]. FedSGD defines a benchmark for federated learning, in this the model weights are updated only for one batch of data by the participants and the central server updates the model parameters by taking the averaging the training results. The aggregation rule at the central server majorly differentiates these two approaches.

Fereidooni et al.[36] presents SAFELearn in this work, a general private federated learning approach that effectively foils strong inference attacks that require access to clients' individual model updates. SAFELearn is more effective than earlier works in terms of communication and computing because it tolerates dropouts and does not necessitate costly cryptographic operations. Furthermore, it doesn't depend on a reliable outsider. Analysis demonstrates that, on common hardware, aggregating models with more than 300K parameters takes less than 0.5 seconds.

III. FUTURE WORK:

For the future work, we would like to extend our study by implementing some of the research reviewed for this paper. Some of the techniques to be included as part of the future papers are Privacy preserving SVM model on HIV sequence data as proposed by Chen et al.[23] PPMCK as proposed by Fan et al.[28], "no-code solution" EasyMPC proposed by Wirth et al.[31], [32] proposed by Smajlović et al., SAFELearn Framework proposed by Fereidooni et al.[36], MPCDDI proposed by Xiao et al.[33] and Prop-FL[34], FedSGD[35] proposed by Maryam Hosseini et al. and McMahan et al. respectively. We would implement these techniques on the MIMIC[37] Patient note data.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCE

[1] D. Moher et al., "Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement," PLoS Medicine, vol. 6, no. 7. Jul. 2009. doi: 10.1371/journal.pmed.1000097.

[2] S. Doyle-Lindrud, "The evolution of the electronic health record," Clin J Oncol Nurs, vol. 19, no. 2, pp. 153–154, 2015, doi: 10.1188/15.CJON.153-154.

[3] "How AI, Data Analytics are revolutionizing the healthcare ecosystem." https://indiaai.gov.in/article/how-ai-data-analytics-are-revolutionizing-the-healthcare-ecosystem?utm_source=LinkedIn&utm_medium=Generic&utm_campaign=WBS-2023-how-ai-data-analytics-are-revolutionizing-the-healthcare-ecosystem-29th-March-2023 (accessed Mar. 30, 2023).

[4] "Healthcare Data Breach Statistics." https://www.hipaajournal.com/healthcare-data-breach-statistics/ (accessed Mar. 15, 2023).

[5] A. Rahman, M. S. Hossain, N. A. Alrajeh, and F. Alsolami, "Adversarial Examples - Security Threats to COVID-19 Deep Learning Systems in Medical IoT Devices," IEEE Internet Things J, vol. 8, no. 12, pp. 9603–9610, Jun. 2021, doi: 10.1109/JIOT.2020.3013710.

[6] [J. Xu et al., "Healthchain: A Blockchain-Based Privacy Preserving Scheme for Large-Scale Health Data," IEEE Internet Things J, vol. 6, no. 5, pp. 8770–8781, Oct. 2019, doi: 10.1109/JIOT.2019.2923525.

[7] "Health Insurance Portability and Accountability Act of 1996 (HIPAA) | CDC." https://www.cdc.gov/phlp/publications/topic/hipaa.html (accessed Mar. 15, 2023).

[8] "What is GDPR, the EU's new data protection law? - GDPR.eu." https://gdpr.eu/what-is-gdpr/ (accessed Mar. 15, 2023).

[9] "Personal Data Protection Act 2012 - Singapore Statutes Online." https://sso.agc.gov.sg/Act/PDPA2012 (accessed Mar. 15, 2023).

[10] N. P. Smart, "Cryptography Made Simple," 2016. [Online]. Available: http://www.springer.com/series/4752

[11] I. Ioannidis and A. Grama, "An Efficient Protocol for Yao's Millionaires' Problem," 2002.

[12] "Federated Learning: Collaborative Machine Learning without Centralized Training Data – Google AI Blog." https://ai.googleblog.com/2017/04/federated-learning-collaborative.html (accessed Apr. 11, 2023).

[13] W. Oh and G. N. Nadkarni, "Federated Learning in Health care Using Structured Medical Data," Advances in Kidney Disease and Health, vol. 30, no. 1, pp. 4–16, Jan. 2023, doi: 10.1053/j.akdh.2022.11.007.

[14] "Federated Learning in Medicine: Facilitating Multi-Institutional Collaboration Without Sharing Patient Data | CBICA | Perelman School of Medicine at the University of Pennsylvania." https://www.med.upenn.edu/cbica/federated-learning-in-medicine-facilitating-multi-institutional-collaboration-without-sharing-patient-data.html (accessed Apr. 11, 2023).

[15] S. Behera and J. R. Prathuri, "Application of Homomorphic Encryption in Machine Learning," in 2020 2nd PhD Colloquium on Ethically Driven Innovation and Technology for Society, PhD EDITS 2020, Institute of Electrical and Electronics Engineers Inc., Nov. 2020. doi: 10.1109/PhDEDITS51180.2020.9315305.

[16] J. Li, X. Kuang, S. Lin, X. Ma, and Y. Tang, "Privacy preservation for machine learning training and classification based on homomorphic encryption schemes," Inf Sci (N Y), vol. 526, pp. 166–179, Jul. 2020, doi: 10.1016/j.ins.2020.03.041.

[17] K. Bonawitz et al., "Practical secure aggregation for privacy-preserving machine learning," in Proceedings of the ACM Conference on Computer and Communications Security, Association for Computing Machinery, Oct. 2017, pp. 1175–1191. doi: 10.1145/3133956.3133982.

[18] A. Shamir, "How to Share a Secret," 1979.

[19] S. B. B. S. Leonie Reichert, "Privacy-Preserving Contact Tracing: current solutions and open questions," Apr. 2020, [Online]. Available: http://arxiv.org/abs/2004.06818

[20] Proceedings of the Second International Conference on Inventive Research in Computing Applications (ICIRCA-2020) IEEE Xplore Part Number: CFP20N67-ART; ISBN: 978-1-7281-5374-2, Privacy Preserving Deep Learning using Secure Multiparty Computation. 2020.

[21] X. Dong, D. A. Randolph, C. Weng, A. N. Kho, J. M. Rogers, and X. Wang, "Developing High Performance Secure Multi-Party Computation Protocols in Healthcare: A Case Study of Patient Risk Stratification," 2019.

[22] M. Jayabalan, M. Ehsan Rana, and K. Rajendran Manoj Jayabalan Muhammad Ehsan Rana, "A Study on k-anonymity, l-diversity, and t-closeness Techniques focusing Medical Data Use of Cloud Computing in Software Engineering Education View project Continuous and Transparent Access Control Framework for Electronic Health Records View project Keerthana Rajendran A Study on k-anonymity, l-diversity, and t-closeness Techniques focusing Medical Data," 2017. [Online]. Available: https://www.researchgate.net/publication/322330948

[23] H. Chen, A. B. Ünal, M. Akgün, and N. Pfeifer, "Privacy-preserving SVM on outsourced genomic data via secure multi-party computation," in IWSPA 2020 - Proceedings of the 6th International Workshop on Security and Privacy Analytics, Association for Computing Machinery, Inc, Mar. 2020, pp. 61–69. doi: 10.1145/3375708.3380316.

[24] D. Li, X. Liao, T. Xiang, J. Wu, and J. Le, "Privacy-preserving self-serviced medical diagnosis scheme based on secure multi-party computation," Comput Secur, vol. 90, Mar. 2020, doi: 10.1016/j.cose.2019.101701.

[25] G. Qiu, X. Gui, and Y. Zhao, "Privacy-Preserving Linear Regression on Distributed Data by Homomorphic Encryption and Data Masking," IEEE Access, vol. 8, pp. 107601–107613, 2020, doi: 10.1109/ACCESS.2020.3000764.

[26] X. Li, R. Dowsley, and M. De Cock, "Privacy-Preserving Feature Selection with Secure Multiparty Computation," 2021.

[27] K. Şahinbaş and F. O. Catak, "Secure Multi-Party Computation based Privacy Preserving Data Analysis in Healthcare IoT Systems," Sep. 2021, [Online]. Available: http://arxiv.org/abs/2109.14334

[28] Y. Fan et al., "PPMCK: Privacy-preserving multi-party computing for K-means clustering," J Parallel Distrib Comput, vol. 154, pp. 54–63, Aug. 2021, doi: 10.1016/j.jpdc.2021.03.009.

[29] S. Stammler et al., "Mainzelliste SecureEpiLinker (MainSEL): Privacy-preserving record linkage using secure multi-party computation," Bioinformatics, vol. 38, no. 6, pp. 1657–1668, Mar. 2022, doi: 10.1093/bioinformatics/btaa764.

[30] M. Pereira, S. Pentyala, A. Nascimento, R. T. de Sousa, and M. De Cock, "Secure Multiparty Computation for Synthetic Data Generation from Distributed Data," Oct. 2022, [Online]. Available: http://arxiv.org/abs/2210.07332

[31] F. N. Wirth, T. Kussel, A. Müller, K. Hamacher, and F. Prasser, "EasySMPC: a simple but powerful no-code tool for practical secure multiparty computation," BMC Bioinformatics, vol. 23, no. 1, Dec. 2022, doi: 10.1186/s12859-022-05044-8.

[32] [32] H. Smajlović, A. Shajii, B. Berger, H. Cho, and I. Numanagić, "Sequre: a high-performance framework for secure multiparty computation enables biomedical data sharing," Genome Biol, vol. 24, no. 1, Dec. 2023, doi: 10.1186/s13059-022-02841-5.

[33] X. Xiao, X. Wang, S. Liu, and S. Peng, "MPCDDI: A Secure Multiparty Computation-Based Deep Learning Framework for Drug-Drug Interaction Predictions," 2023, pp. 263–274. doi: 10.1007/978-3-031-23198-8_24.

[34] S. Maryam Hosseini, M. Sikaroudi, M. Babaie, and H. R. Tizhoosh, "Proportionally Fair Hospital Collaborations in Federated Learning of Histopathology Images," IEEE Trans Med Imaging, pp. 1–1, Jan. 2023, doi: 10.1109/tmi.2023.3234450.

[35] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," Feb. 2016, [Online]. Available: http://arxiv.org/abs/1602.05629

[36] H. Fereidooni et al., "SAFELearn: Secure Aggregation for private FEderated Learning (Full Version)," 2021. [Online]. Available: https://github.com/TRUST-TUDa/SAFELearn.

[37] A. E. W. Johnson et al., "MIMIC-IV, a freely accessible electronic health record dataset," Sci Data, vol. 10, no. 1, Dec. 2023, doi: 10.1038/s41597-022-01899-x.