

# Comparative Analysis of Traditional Machine Learning and Privacy Preserving Federated Learning



MCS

by

Nasir Mehmood


A thesis submitted to the faculty of Information Security Department, Military College of Signals, National University of Sciences and Technology, Rawalpindi in partial fulfilment of the requirements for the degree of MS in Information Security


August 2023


# THESIS ACCEPTANCE CERTIFICATE

## THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS Thesis written by Mr. Nasir Mehmood, Registration No. 00000318740, of Military College of Signals has been vetted by undersigned, found complete in all respects as per NUST Statutes/Regulations/MS Policy, is free of plagiarism, errors, and mistakes and is accepted as partial fulfillment for award of MS degree. It is further certified that necessary amendments as pointed out by GEC members and local evaluators of the scholar have also been incorporated in the said thesis.

Signature:   
Name of Supervisor Dr. Fawad Khan  
Date: 7/19/23

Signature (HOD):  HoD  
Information Security  
Date: 7/19/23 Military College of Sigs

Signature (Dean/Principal)   
Date: 7/19/23  
Brig  
Dean, MCS (NUST)  
(Asif Masood, Phd)

# Declaration

I hereby declare that no portion of work presented in this thesis has been submitted in support of another award or qualification either at this institution or elsewhere.

# Dedication

“In the name of Allah, the most Beneficent, the most Merciful”

I dedicate this thesis to my mother, sister, and teachers who supported me each step of the way.

# Abstract

With the rapid development in the IT field, thousands, even millions of IoT devices were developed. IoT devices play a vital role in the field of healthcare. Nowadays smart wearable devices are used in the field of healthcare to monitor the health of patients like heartbeat, fitness, blood pressure, etc. These IoT devices generate a vast variety of data, but in healthcare, the generated data is related to patients. This data contains the private and sensitive information of the patient.

In real world, there are lot of big data generated on daily bases from different sources. These data are in hundreds of gigabytes, and it requires large storage devices. Artificial Intelligence and Machine learning is a technique that is used to predict the result on the base of given data. In machine learning, it requires the data to be present in a centralized location, which is a major security concern for the users.

FL is a machine learning technique that trains algorithms across multiple decentralized devices. FL works on decentralized heterogeneous computing devices. It varies in many ways from traditional Machine Learning like time saving, resource saving, etc. FL is one of the types of machine learning that improve the privacy and security concerns. In FL one is a server that contains the main model. The server shares the model with clients and clients train the local model collaboratively on the bases of data. This technique protects the user from transferring data, and it minimizes privacy issues. For privacy preserving, we will use the Pailliar Homomorphic Encryption.

# Acknowledgments

All praises to Allah for the strengths and His blessing in completing this thesis.

First and foremost I praise and acknowledge **ALLAH**, the most beneficent and the most merciful. Secondly, my humblest gratitude to the Holy Prophet Muhammad (SAW) whose way of life has been a continuous guidance for me.

I would like to convey my gratitude to my supervisor, Dr. Fawad Khan, for his supervision and constant support. His invaluable help of constructive comments and suggestions throughout the experimental and thesis works are major contributions to the success of this research. Also, I would thank my committee members; Dr. Shahzaib Tahir, and Dr. Imran Makhdoom for their support and knowledge regarding this topic.

Last, but not the least, I am highly thankful to my parents, brothers and sisters. They have always stood by my dreams and aspirations and have been a great source of inspiration for me. I would like to thank them for all their care, love and support through my times of stress and excitement.

# Table of Contents

<b>THESIS ACCEPTANCE CETIFICATE .....</b>	<b>i</b>
<b>Declaration.....</b>	<b>ii</b>
<b>Dedication .....</b>	<b>iii</b>
<b>Abstract.....</b>	<b>iv</b>
<b>Acknowledgments .....</b>	<b>v</b>
<b>Chapter 1 .....</b>	<b>1</b>
<b>Introduction.....</b>	<b>1</b>
1.1 Background .....	1
1.2 Motivation / Justification for the Selection of the Topic .....	3
1.3 Problem Statement.....	4
1.4 Objectives .....	4
1.5 Thesis Contribution.....	4
1.6 Thesis Organization.....	5
<b>Chapter 2 .....</b>	<b>6</b>
<b>Preliminary Background and Related Word .....</b>	<b>6</b>
2.1 Introduction .....	6
2.2 Machine Learning.....	6
2.1.1 Types of Machine Learning and Algorithms .....	7
2.1.2 Classification Machine Learning Algorithms .....	8
2.3 Federated Learning.....	12
2.3.1. Federated Learning Challenges.....	14
2.3.2. Federated Learning Algorithms .....	14
2.4 Related Work .....	15
2.5 Privacy-Preserving.....	18
2.6 Advantages of Privacy-Preserving.....	21
2.7 Area of Application .....	21
<b>Chapter 3 .....</b>	<b>22</b>
<b>Methodology .....</b>	<b>22</b>
3.1 Introduction .....	22

3.2	Datasets .....	23
	Heart Disease Dataset.....	23
	Breast Cancer Dataset.....	23
3.3	Participants .....	24
3.4	Paillier Encryption Scheme .....	24
3.5	Machine Learning Algorithm.....	26
3.6	Traditional Machine Learning (Centralized) .....	26
3.7	Distributed On-Site Learning (Independent Learning).....	27
3.8	Federated Learning (Without Paillier Encryption) .....	28
3.9	Federated Learning Using Paillier Encryption.....	29
<b>Chapter 4 .....</b>		<b>31</b>
<b>Security and Performance Analysis .....</b>		<b>31</b>
4.1	Performance Measure Indices .....	31
	Accuracy:.....	31
	Precision:.....	31
	Recall:.....	32
	F1-Score: .....	32
4.2	Confusion Matrix.....	32
4.3	Breast Cancer Dataset Results .....	33
	4.3.1. Parameters and Values .....	33
	4.3.2. Breast Cancer Heatmap .....	33
	4.3.3. Distributed on-site learning (Independent Learning): .....	34
	4.3.4. Federated Learning (with paillier encryption) .....	35
	4.3.5. Federated Learning (without paillier encryption).....	37
	4.3.6. Centralized Machine Learning (Traditional ML).....	39
4.4	Heart Disease Dataset Results: .....	41
	4.4.1. Parameters and Results .....	41
	4.4.2. Heart Disease Heatmap .....	41
	4.4.3. Distributed on-site learning (Independent Learning) .....	42
	4.4.4. Federated Learning (with paillier encryption) .....	42
	4.4.5. Federated Learning (without paillier encryption).....	44
	4.4.6. Centralized Machine Learning (Traditional Machine Learning).....	46
4.5	Comparison: .....	48



<b>Chapter 5 .....</b>	<b>51</b>
<b>Conclusion .....</b>	<b>51</b>
<b>References .....</b>	<b>52</b>

# List of Figures

Figure 1 Types of Machine Learning .....	7
Figure 2 Logistic Regression .....	9
Figure 3 Support Vector Machine (SVM) .....	10
Figure 4 Decision Tree.....	11
Figure 5 Random Forest.....	12
Figure 6 Federated Learning Architecture .....	14
Figure 7 Traditional Machine Learning (Centralized) .....	27
Figure 8 Distributed On-Site Learning (Independent Learning).....	28
Figure 9 Federated Learning (Without Paillier Encryption) .....	29
Figure 10 Federated Learning Using Paillier Encryption .....	30
Figure 11 Heat Map of Breast Cancer Dataset.....	34
Figure 12 Confusion Matrix of FL With paillier encryption .....	37
Figure 13 Confusion Matrix of FL Without paillier encryption .....	39
Figure 14 Confusion Matrix of Centralized ML .....	41
Figure 15 Heatmap of Heart Disease .....	42
Figure 16 Confusion Matrix of FL With paillier encryption .....	44
Figure 17 Confusion Matrix of FL Without paillier encryption .....	46
Figure 18 Confusion Matrix of Centralized ML .....	48
Figure 19 Accuracy Results Comparison of Breast Cancer Dataset .....	49
Figure 20 Accuracy Results Comparison of Heart Disease Dataset .....	50

# List of Tables

Table 1 Confusion Matrix and its relationship with Classification Report.....	33
Table 2 Classification Report of FL With paillier encryption.....	36
Table 3 Confusion Matrix of FL With paillier encryption.....	36
Table 4 Classification Report of FL Without paillier encryption.....	38
Table 5 Confusion Matrix of FL Without paillier encryption.....	38
Table 6 Classification Report of Centralized ML.....	40
Table 7 Confusion Matrix of Centralized ML.....	40
Table 8 Classification Report of FL With paillier encryption.....	44
Table 9 Confusion Matrix of FL With paillier encryption.....	44
Table 10 Classification Report of FL Without paillier encryption.....	46
Table 11 Confusion Matrix of FL Without paillier encryption.....	46
Table 12 Classification Report of Centralized ML.....	48
Table 13 Confusion Matrix of Centralized ML.....	48
Table 14 Accuracy Comparison of two datasets.....	49

## Introduction

### 1.1 Background

The IT field has been experiencing rapid development, thousands even millions of IoT devices were developed and still working on IoT devices and improve their security and accuracy. IoT taking the advantage of faster 5G/6G internet. IoT devices have a crucial role in various sectors, including healthcare, transportation, mobile apps, defence, and cybersecurity. In today's era smart devices are used in the field of healthcare. These smart devices are used to monitor the health of patients 24/7 like heartbeat, fitness, BP (Blood Pressure), etc. These IoT devices generate a vast variety of data, but in healthcare, the generated data is related to patients. This data contains the private and sensitive information of the patient. Today maximum of hospitals owns an electronic health record (EHR). Most of the patient data are saved on a computer or saved on the network and doctors check the patient's reports remotely. But the data is not available publicly, it contains some constraints to accessing this data. As we know that the data on the network in plaintext or without any login credentials is not safe.[1]

Healthcare data is one of the most highly sensitive data in terms of data privacy and security concerns. In the world lot of hospitals and clinics have different medical departments. On daily basis, these hospitals generate and save a lot of data related to the patient's disease. This type of data is very sensitive. Hospital or clinic owners do not want their data to leave their premises and the hospital also wants that the computer generates correct results about the patient disease on the bases of parameters. Due to patient's data security and privacy concerns the hospitals are not in the favor of sharing data on a cloud or on a third party. Various laws have been established to safeguard the privacy and security of individuals' personal data. So, it is difficult to store data in a centralized location. Due to data security concerns and data privacy, different rules and regulations are made in the world such as:

- California Consumer Privacy Act (CCPA)
- General Data Protection Regulation (GDPR)
- Personal Data Protection Act (PDP)
- Consumer Privacy Bill of Rights (CPBR)
- Cybersecurity Law of the People's Republic (CLPR)

These regulatory authorities have been formed to protect user's privacy and security of personal data [2].

To protect the sensitive information from unauthorized user, we need some technological improvement. In real world, there are lot of big data generated on daily bases from different sources. These data are in hundreds of gigabytes, and it required large storage devices [4]. Artificial Intelligence and Machine learning (ML) is a technique that is used to predict the result on the base of given data. They can learn from the environment. In machine learning, it requires the data to be leave its premises and reside on a centralized location, which is a major security concern for the users. Typically, there are two ways to train model using machine learning. Initially uploading the data on centralized location, it takes lots of time and resources such as hard disk, bandwidth, etc. Second is, instead of uploading data to the centralized location we need to deploy the machine learning model on each site [3].

Federated Learning (FL) represents an innovative machine learning approach employed to address concerns related to user privacy and security. An algorithm is trained using FL, a machine learning technique, across numerous distributed devices. FL work on decentralized heterogeneous computing devices. It differs from traditional machine learning in several respects, such as time and resource efficiency. FL is like the distributed machine learning and is a technique of machine learning that improve privacy and security concerns. In FL, model is brought to the data instead of sending data to server for model training.

In FL, one is a server that contain the main model. The server shares the model with clients and clients train the local model collaboratively on the bases of data. This technique protects the user to be transferred data and it minimizes the privacy issues. After training the

local model, these local models generate weights and were sent to the server and server update the main model by using aggregating methodology on local model and server generate new weights of global model and again shares it with clients in a secure way.

Nowadays encryption is a common practice of everyone. No one wants to share data in plaintext form. In federated learning, we can encrypt the local model using Homomorphic Encryption (HE) to minimize data leakage issues. Homomorphic Encryption supports to apply different arbitrary computations on encrypted data. HE allows us to perform the arithmetic computation on encrypted text instead of decrypting text. But its better to encrypt user data before sending to centralized location. Encryption required more time to encrypt the plain text into cipher text.

## **1.2 Motivation / Justification for the Selection of the Topic**

There are lot of machine learning algorithms that train the model and predict the target. However, the data privacy and confidentiality are the main barrier to the adoption of traditional machine learning. Quality machine learning required quality training dataset and sometimes it is difficult to acquire.

Traditional machine learning requires that all the data should be available on a single centralized server in order to train the model while the Federated Learning (FL) operates without direct access to user's raw data, thereby enhancing data security and privacy measures.

So, we compare federated learning with traditional machine learning to check the data security, the accuracy of the model training and computation cost. Because as malicious activities on social media and online plat forms increases, day-by-day and traditional machine learning is lacking in providing any security and privacy.

FL presents a solution to this challenge by enabling multiple entities to collectively train a single machine learning model, all the while keeping their individual training data

undisclosed. In FL, raw data is not sent to the main server; instead, the local model is trained, and the resulting weight is sent to the main model for further training.

### **1.3 Problem Statement**

Federated learning emerged from the existing machine learning to evaluate the equivalent trained model without sharing the data. However, FL can be attained by doing performance trade-offs. Quantifying these approaches with respect to privacy and performance trade-offs is important to employ these in various applied domains.

### **1.4 Objectives**

The main objectives of this thesis are:

- Compare and contrast Federated Learning with traditional machine learning
- Ensuring Privacy while performing out-sourced computations on data
- Performing in-depth literature review for above mentioned goal
- Validate the results of the above-mentioned goal using any dataset.

### **1.5 Thesis Contribution**

In the 21st century, we find ourselves in the age of machine learning, where this technology finds application in nearly every facet of global existence. All hospitals want to get the correct prediction of the provided data of a specific disease, for this, we need to train a machine learning algorithm, training needs a large amount of correct and real-life datasets to train a model. The effectiveness of machine learning models in the medical field might be constrained if they are trained solely on a single dataset or data originating from a specific medical facility. Model training on a single dataset or medical site cannot produce the appropriate degree of accuracy due to the dearth of datasets. Because healthcare departments are not uploading their data to a centralized location to train a model like cloud. They have some security concerns about patients' data. So Federated learning with homomorphic will reduce the security concerns of the hospitals.

Federated learning is a machine learning method that enables non-affiliated hospitals to leverage the collective knowledge of multiple institutions' rich datasets without centralizing the data in one location. This method efficiently deals with essential considerations such as safeguarding data privacy, ensuring data security, upholding data access rights, and making use of diverse data sources [7].

By adopting federated learning, each hospital can collaborate in building a shared model without disclosing their raw data to a central entity. This ensures data privacy, as sensitive patient information remains locally stored within each institution, reducing the risk of privacy breaches. Additionally, the decentralized nature of federated learning enhances data security, as there is no single point of vulnerability for potential attacks.

FL technique will be helpful in all fields of life in terms of security, where electronic data are generated. FL can be implemented in hospitals, banking sector, IT department, Cyber Security, etc.

## **1.6 Thesis Organization**

There are six chapters within this thesis. List of chapters used in this thesis is given below:

- Chapter 2 is literature reviewed in this thesis. It focuses on discussion about machine learning and its different types of algorithms related to classification. It also focuses on federated learning, security challenges in federated learning and privacy preserving techniques.
- Chapter 3 contains the methodology that we will use in our thesis. It focuses on the selection of dataset, number of participants and homomorphic encryption scheme.
- Chapter 4 is security and performance analysis. In this analyze the different results and compare all these results. We use some performance measure indices to measure the accuracy of the algorithm.
- Chapter 5 serves as the conclusion, signifying the end of the document. Within this chapter, the conclusion is presented along with potential areas for future work.



# Preliminary Background and Related Word

## 2.1 Introduction

This chapter is related to literature review. In this we explained the preliminary background and related work. In this chapter we explained important topics in detail. We explained machine learning and its classification algorithms, machine learning and its algorithms, and secure encryption methodologies.

## 2.2 Machine Learning

Machine learning falls under the umbrella of Artificial Intelligence (AI). In machine learning computers assign a task to complete it and machine learning code learn it from its experiences and try to complete the task. Machine learning refers to learning on its own without writing lengthy code to complete a task. Machine learning emphasizes code that gets large dataset and trains itself on it using different machine learning algorithms. Machine learning learn itself from its experiences. More experience will give us more accurate results. After training algorithm, we used the same algorithm is used for making decision, predictions or forecasting based on data. There are different examples in our real life for which machine learning is used to predict cancer disease from different medical reports. Machine learning is used in wide verity of fields like robotic, business, computer games, google map, healthcare, online fraud detection, pattern recognition etc. [13]

## 2.1.1 Types of Machine Learning and Algorithms

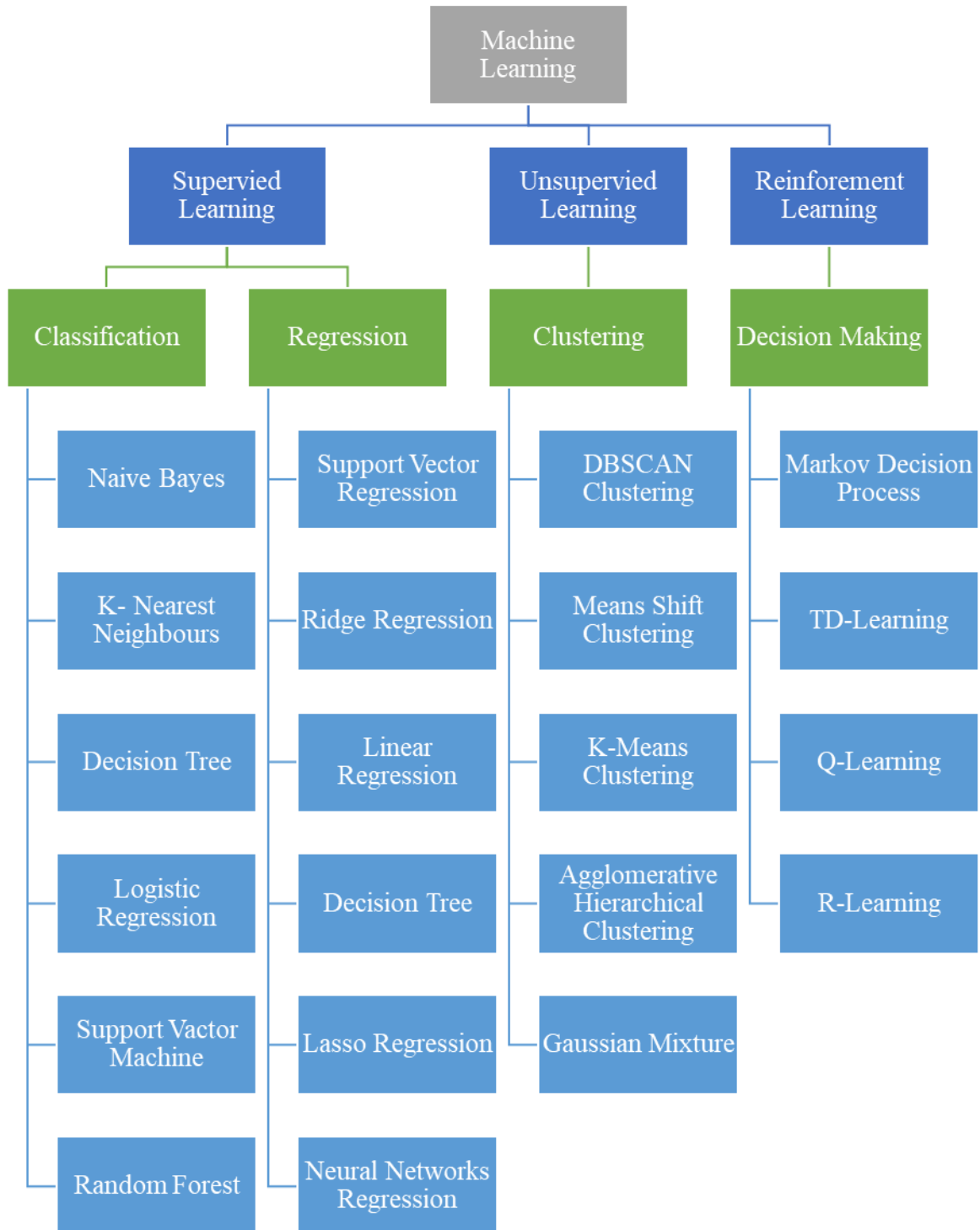


Figure 1 Types of Machine Learning

## 2.1.2 Classification Machine Learning Algorithms

### Logistic Regression

Logistic Regression (LR) falls under the category of supervised machine learning and is employed to address problems yielding binary outcomes. It works on the probability that the chance of an event occurrence or not based on the input. It is mostly used in those data sets which results are based on two results like yes/no, true/false, 1/0, etc. For example, the probability of a tumor is malignant or benign, or a patient have heart disease or not, or a received email is a spam or not. Logistic regression is a statistical technique utilized to analyze the relationship between variables that are dependent and independent. The logistic regression model assumes a logistic or sigmoidal relationship between the predictor variables and the outcome variable. The logistic function, also referred to as the sigmoid function, maps real-valued numbers to values within the range of 0 to 1 as shown in figure 2. This mathematical formula finds widespread application across diverse disciplines, such as statistics, machine learning, and neural networks. Its primary application is in modeling binary outcomes and introducing non-linearity [13][14].

#### Sigmoid Function:

$$f(x) = \frac{1}{1 + e^{-x}}$$

- $f(x)$  is the output value, confined to the range of 0 to 1.
- The numerical approximation of the base of the natural logarithm, denoted as 'e', is approximately 2.71828.
- The  $x$  is the input value, a real number.

#### Advantages:

LR has following advantages:

- Simply implementation
- Computational efficiency
- Train effectively
- Ease of regulation

- Efficient for large dataset
- For input feature no scaling is required

**Disadvantages:**

LR has following disadvantages:

- No ability to solve non-linear problems
- Susceptible to over fitting
- Susceptible to outliers

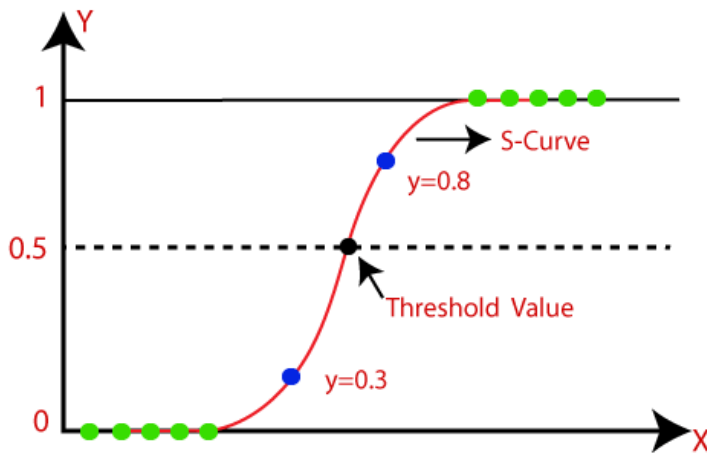


Figure 2 Logistic Regression

**Support Vector Machine (SVM)**

The SVM algorithm finds application in both classification and regression tasks. It is working on hyperplane. It finds the best suitable hyperlane that is the decision boundary between different classes that have different set of objects or points. SVM aims at classifying the objects based on examples in the training data set. The distance between classes is known as margin. The point on the margin is called a support vector as shown in figure 3. The kernel is a technique employed by SVM to handle data that cannot be linearly separated. The data is transformed into a space with a higher number of dimensions, enabling linear distinction [13][14].

**Advantages:**

SVM has following advantages:

- Manage linear and non-linear data
- Efficient for small and large dataset
- Less probability of over fitting

- Scale up with high dimensional data
- Support multiple classes

**Disadvantages:**

SVM has following disadvantages:

- Computationally expensive
- Large dataset effect its performance
- Difficult to select the kernel function
- Do not work when all data is noisy

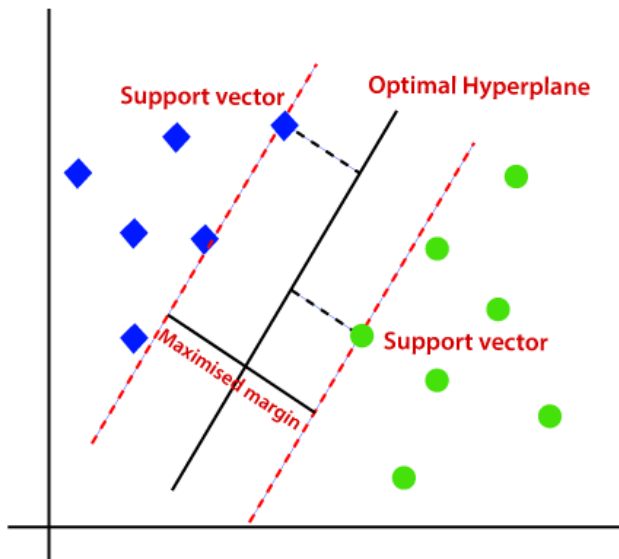


Figure 3 Support Vector Machine (SVM)

**Decision Tree**

The decision tree algorithm is employed for addressing regression and classification issues, achieved through iterative data division using specific criteria. Decision tree based on recursively splitting data. It builds hierarchical tree like a structure. The decisions are the leaves of tree and splitting data into nodes or branches as shown in figure 4. The target variable is predicted using a decision tree based on the decision on the bases of data features. In classification, decision tree dependent variable results are in discrete form (yes/no, 0/1, etc.) and in regression, decision tree dependent variable results are in continuous form [13].

**Advantages:**

Decision tree has following advantages:

- Can be used for classification and regression
- Can fill missing value in data
- Ease in interpretation
- Can Overcome the over fitting problem
- Deal with both numerical and categorical data
- They can handle both numerical and categorical features.
- They are computationally efficient during prediction.

### Disadvantages:

Decision tree has following disadvantages:

- It is unstable
- Difficult to manage large tree
- They can easily overfit the training data, capturing noise and outliers.
- sensitive to small changes in the data
- They may not generalize well to unseen data if the tree structure is too complex.

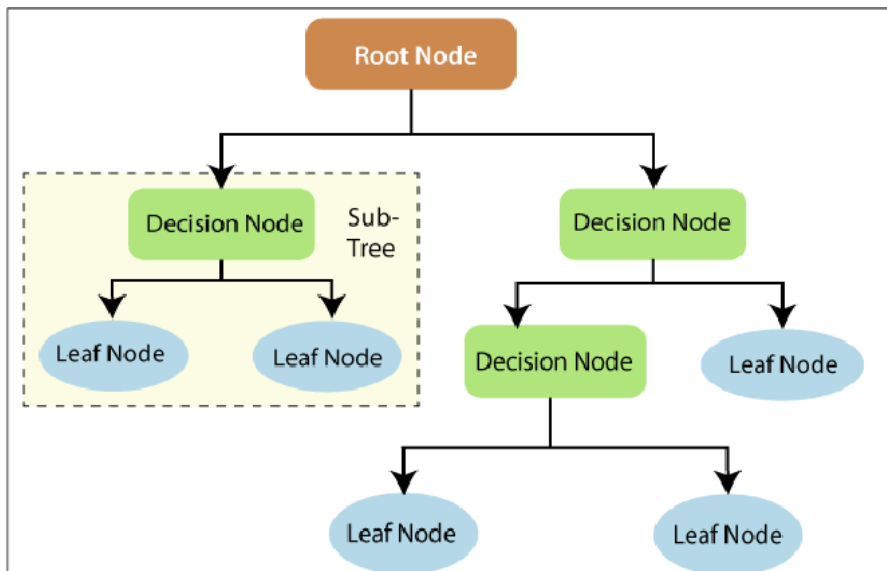


Figure 4 Decision Tree

### Random Forest

The Random Forest algorithm is utilized for both classification and regression tasks, and it consists of a collection of trees. It is an ensemble model approach that combines multiple classifiers to make accurate predictions. The more number of tree will have more

accurate results. It is like a forest containing many trees. It takes random data and creates a bunch of trees as shown in figure 5. In classification, random forest dependent variable results are in discrete form (yes/no, 0/1, etc.) and in regression, random forest dependent variable results are in continuous form [13].

### Advantages:

Random Forest has following advantages:

- Can be used for classification and regression
- Solve over fitting problems
- Efficient for huge dataset

### Disadvantages:

Random Forest has following disadvantages:

- Need more time for training
- High complexity

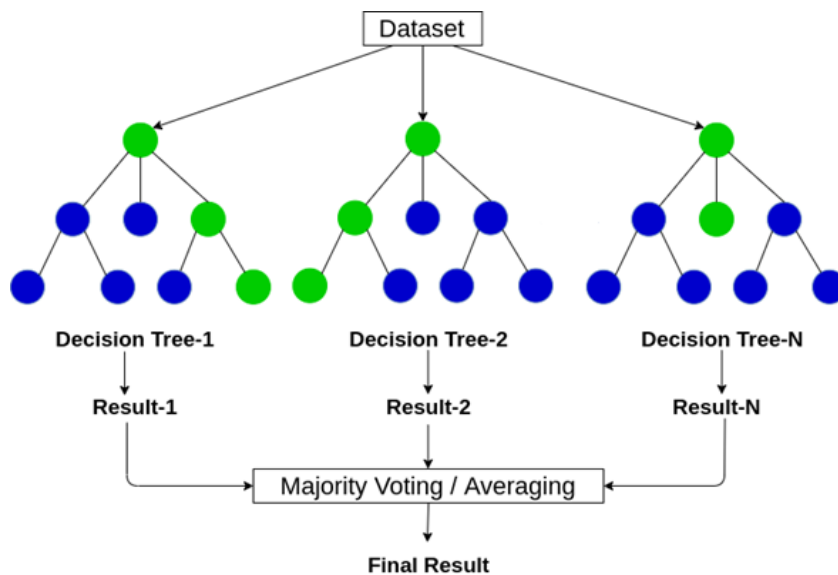


Figure 5 Random Forest

## 2.3 Federated Learning

Federated learning is a novel approach in machine learning that enable the model to the train itself using decentralized data without requiring to send or transfer data to the centralized server as shown in figure 6. FL train a common share model across distributed devices. It allows multiple parties such as device or organization of same domain to

participate in training process. It collaboratively builds a machine learning model while keeping their data on local device.

Federated learning differs from traditional machine learning in different aspects such as time and resource efficiency. In traditional machine learning, it collects all data in a centralized location for model training. However, centralized approach has lot of concerns in participant mind. They think that this approach may lead to user's data privacy, security, and the potential for sensitive information to be exposed. Federated learning approach address these challenges and mitigate the user's privacy and security concerns. FL allowing data to remain on local device while contributing to model training process [9].

The fundamental idea of federated learning encompasses the subsequent stages:

1. **Initialization:** A central server or authority generates a global model. This model is usually pre-trained on a large dataset to provide a starting point.
2. **Distribution:** The global model parameters is sent to participating devices or nodes in a network. Each device has its own local dataset that is representative of the broader population.
3. **Local Training:** On their respective devices, each node trains the global model parameters using their local data. Devices or nodes perform training locally sharing the data or any sensitive information with the central server or other nodes.
4. **Model Aggregation:** After local training, the nodes or device send only the updated model parameters (not the data) back to the centralized server.
5. **Model Aggregation and Update:** The central server combines or aggregates the model parameters received from all the nodes and incorporates them into the global model. This process can involve techniques like averaging or weighted aggregation to account for the varying quality or quantity of data across nodes.
6. **Iterative Process:** The process described in steps 2 to 5 is iteratively repeated either for a predefined number of times or until specific convergence criteria are satisfied. The global model is continuously improved by incorporating knowledge from all participating nodes.



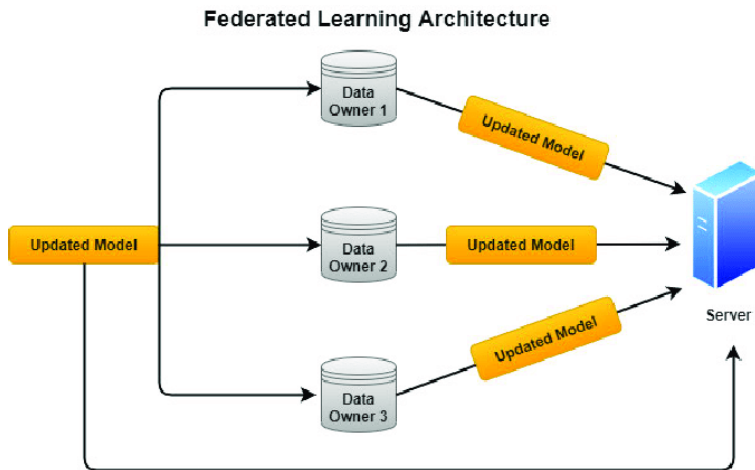


Figure 6 Federated Learning Architecture

### 2.3.1. Federated Learning Challenges

Federated learning is a secure methodology to train your algorithm without sharing your data with the centralized server. But FL still faces some challenges that affect the performance of algorithms prediction. Some of these challenged are:[19][25]

- Non-Independent and Non-Identical Distributed data
- Unbalance data
- Massively distributed data
- Unreliable data communication
- Limited device memory
- Poisoning attack

### 2.3.2. Federated Learning Algorithms

Federated learning algorithms are used on server. There are different types of federated algorithms that are used for averaging the local model on server side. Aggregation of local model is a necessary step in federated learning. FL enhances both the security and accuracy of the model. After aggregating the model, the updated model again sends to the clients and clients us the updated model.

The aggregation algorithm plays an important role in federated learning. It gather all local updates and combines these updated local models of the participants. Following are several well-known aggregation algorithms commonly used in federated learning. [18][5][6].

- FedAvg
- FedMA
- FedNAS
- FedGKT
- SMC-Avg
- FedProx
- FedSGD
- Scaffold
- Tensor Factorization

Federated Averaging (FedAvg) is a pioneering and extensively employed federated learning algorithm initially introduced by Google. Mostly researchers used FedAvg algorithm for aggregation. Mostly research papers are explained in detail about FedAvg algorithm. FedAvg is raised by google to help in joining several models in one global mode [17][18].

## **2.4 Related Work**

In literature there are many related work are done in term of federated learning but still there are some research gap these work. In federated learning, the main focus is on the accuracy and data security. Different authors works using different machine learning and federated learning algorithm. Some of the related work are explain as below.

Asad, Muhammad, Ahmed Moustafa, and Takayuki Ito [4] proposed a solution to secure the user's sensitive data. They test two different datasets (MNIST and CIFAR-10) against different scenarios and compare classical Machine Learning (centralized and distributed ML) and FL. They evaluate the convergence of these three models on the bases of three scenarios:

1. 50 participants and 100 rounds on the above datasets
2. 50 participants and 200 rounds on the above datasets
3. Assess the convergence considering the impact of participants and 100 rounds on the above datasets [Participants  $p = \{20, 40, 60, 80, 100\}$ ]

In scenario one and two, when applying centralized machine learning to the MNIST dataset, accuracy rates of 65% and 73% were achieved, respectively. For the CIFAR-10 dataset, the accuracy rates for scenario one and two were 54% and 62%, respectively.

In scenario one and two, distributed machine learning was employed on the MNIST dataset, resulting in accuracy rates of 72% and 78%, respectively. For the CIFAR-10 dataset, the accuracy rates for scenario one and two were 67% and 72%, respectively.

In scenario one and two, federated learning was utilized on the MNIST dataset, yielding accuracy rates of 92% and 97%, respectively. For the CIFAR-10 dataset, the accuracy rates for scenario one and two were 86% and 94%, respectively.

In the third scenario, the machine learning algorithms failed to achieve the desired level of convergence when dealing with datasets containing a small number of participants. However, as the number of participants increased, there was a notable improvement in the performance of both datasets, indicating a positive correlation between participant count and algorithm effectiveness.

Liu, Ji, Jizhou Huang, Yang Zhou [2] identify that FL works differently from traditional centralized machine learning. Initially, federated learning prohibits communication with raw data, whereas traditional machine learning permits such communication. FL allows work on distributed heterogeneous source devices, while the traditional ML relies on a single server. FL pays attention to user security and privacy and gives the advantage of encryption to ensure data privacy and security of user data, while the traditional ML pays little attention to these security issues.

Lo, Sin Kit, Qinghua Lu, Liming Zhu [8] mention federated learning life cycle. It consists of eight stages: client selection, model distribution, training, transmitted, aggregation, evaluation, deployment, and monitor.

P Varalakshmi, K Narmadha [3] use federated learning technology to predict the desired outcomes. They assess the precision of three distinct machine learning algorithms: Logistic Regression (LR), Support Vector Machine (SVM), and Perceptron, in comparison to conventional centralized machine learning. They used two different data sets that are digits dataset and UCI Obesity Dataset. The graphs were mentioned in the paper and show that the accuracy of each algorithm for federated learning is higher than traditional centralized machine learning algorithm.

Sinha, Nidhi, Teena Jangid [12] applied a variety of seven machine learning classification algorithms to forecast Cardiovascular Disease (CVD). The seven different classification models are Logistic Regression (LR), Support Vector Machine (SVM), Adda Boost, XG-Boost, K-Nearest Neighbour (KNN), Naive Bayes, and one simple artificial neural network. The author compares all these algorithms according to the accuracy of algorithm. After comparing results of all these seven algorithms SVM got higher accuracy which was 92.31% and KNN got lowest accuracy which was 70.33%.

The survey [9] discuss that the cost and latency is still a big problem in traditional machine learning. Both problems are difficult to solve because data are located on distributed locations and are in big size. This creates serious problem in traditional machine learning in communication and computation. If data are distributed in different locations, it can overwhelm the limited bandwidth in communication.

There are three types of federated learning that are mentioned in different publications. These are Vertical Federated Learning (VFL), Horizontal Federated Learning (HFL), and federated transfer learning [5][9]. In Horizontal federated learning, where features space is same but are different in terms of dataset. For instance, there are four different hospitals are in four different locations. Each hospital shares the same feature of the patients generated by the medical equipment and all hospital send it to server and train the global model. In Vertical

federated learning, where dataset can be similar but have different feature space. For example, a hospital and a health insurance company share the same user dataset. Hospitals are dealing with patients' diseases and their history, whereas insurance companies deal with medical bills. Federated transfer learning is used to utilize the data from different sources to train the model. Horizontal federated learning is commonly used in healthcare to train the model.

There are security potential attacks in federated learning that make FL vulnerable against these attacks, such as poisoning attack, inference attack, backdoor attack, malicious server and communication bottleneck [10][11].

FL is still resisted against the inference attack. Inference attack, that is, when data is used to train the model then an adversary infers the information without any prior knowledge by analyzing the large amount of data. An attacker illegally gains knowledge about the model and reconstructs the training data. The researcher proposed a method that he first reconstruct the medical data through a variational autoencoder (VAE) and to gain higher security they add some noise factor to resist inference attack [10].

There are some open-source FL systems that are mentioned in publications, e.g. PrivacyFL, TensorowFL, and Pysyft are now intensively used by both research communities, e.g., healthcare, and computer visions [2][5][6].

When the local model sends weights to the main server then it needs to aggregate all weights that are received. The aggregation step is very important to achieve higher security and reduce users' privacy concerns. So different aggregation algorithms are used in federated learning. Most researchers used the FedAvg algorithm for aggregation [5] [6].

## **2.5 Privacy-Preserving**

Our first thinking is that federated learning provides adequate privacy and security. User's concerns related to federated learning revolved around the privacy of the user data.

After implementing FL novel security challenges are raised. Transmission of model or weights can still reveal sensitive information.

### **Threats:**

Existing privacy preserving algorithms can still reveal user data. In [20] mention clearly how an attacker can leak information of client during clients training data. In this the attacker can infer the existence of exact data points in training i.e. specific locations. Suppose that are K participants that collaboratively train the model so there can be a participant who can an adversary. His goal is to infer information from the training data process. In this adversary download the updated model at each iteration. There can be a chance that the adversary may be a malicious participant and send the fake or bogus data to the server. At the end the sever will predict the wrong results. Federated learning setting can raise many vulnerabilities and threats. Some of malicious actors can be:[19]

- **Malicious Server:** A malicious server can inspect the user data. Sever can be honest-but curious. Sometime malicious server gathers information but does not alter information, but he can also temper the model.
- **Insider adversary:** This type of adversary acts as a participant and infers the information.
- **Outsider adversary:** When communication between client and server. The adversary and eavesdrop the channel.

There are some techniques that are used for FL privacy preserving. These are:

- **Secure Multiparty Computation (SMPC)** is a cryptographic method that enables multiple parties to jointly compute a shared function using their private inputs, all while preserving the privacy and security of each participant's sensitive data. The primary objective of SMPC is to safeguard the confidentiality of the individual inputs throughout the computation procedure.

In a standard situation, a number of entities (commonly denoted as "players" or "participants") possess individual private inputs and seek to collectively calculate a function using these inputs. However, they do not want to disclose their inputs to

one another. SMPC enables them to achieve this goal through cryptographic protocols [19] [24].

- **Differential Privacy (DP)** Differential privacy's fundamental concept is to inject noise into query results in a manner that ensures the inclusion or exclusion of any individual's data has minimal impact on the overall outcome of the query. This noise "blurs" the results and makes it difficult to infer sensitive information about any individual, even if an adversary has significant background knowledge or auxiliary information [19].
- **Homomorphic Encryption (HE)** is an advanced cryptographic technique that allows computations to be performed on encrypted data without the need to decrypt it. This property is known as homomorphism, and it enables privacy-preserving data processing and analysis. With homomorphic encryption, sensitive information remains encrypted throughout the entire computation process, protecting the confidentiality of the data and the anonymity of the individuals concerned [28]. There are three types of homomorphic encryption: Fully Homomorphic Encryption (FHE), Somewhat Homomorphic Encryption (SHE) and Partially Homomorphic Encryption (PHE).
- **Data anonymization** Data anonymization is a privacy-enhancing technique used to protect the identities of individuals in a dataset by removing or obfuscating direct or indirect identifiers. The primary goal of data anonymization is to make it difficult or practically impossible to link specific data records to the individuals they represent, while still preserving the utility of the data for analysis, research, or other purposes. Data anonymization hides or removes sensitive information from data before publishing [25].

Anonymization is particularly important when dealing with sensitive or personal data, as it helps to comply with privacy regulations and protect individuals' privacy rights. By anonymizing data, organizations can share or publish datasets for various purposes, such as research, without revealing sensitive information about the individuals in the dataset.

## **2.6 Advantages of Privacy-Preserving**

Following are some advantages:

- Minimize the security concerns of data breach
- Train the model efficiently and in less time
- Get higher accuracy after training the model
- Privacy preserving inference over the trained model

## **2.7 Area of Application**

- IT Industry
- Telecommunication Sector
- Banking Sector
- Healthcare
- Business Analytics
- Autonomous Driving



# Methodology

## 3.1 Introduction

This chapter holds significant importance as it outlines the systematic strategy and methodologies employed for data collection and analysis. This chapter provides a comprehensive understanding of the research process and the specific methods employed to address the research questions or objectives. By presenting a clear and detailed account of the chosen methodologies, the chapter enables readers to evaluate the reliability, validity, and generalizability of the study's findings. This chapter includes various stages of thesis, which include participants, dataset, machine learning algorithms, federate learning models and data analysis. This study has the following research questions:

- How can we protect client data?
- How can clients participate in the training process?
- Will the prediction accuracy up to the mark of machine learning?

This research is based on predicting the classified results based on data. In the modern era data privacy in the most serious concern for users or companies, most especially in healthcare sector. The record of a patient is very important data, and without patient permission any hospital cannot share the medical data of any patient with any other entity.

Machine learning is the technique that used different algorithm for predicting or forecasting based on given data. Traditional machine learning used the centralized training approach, where all data should reside on a centralized location or server. The data is available in plaintext form which is a big security concern for sensitive organizations or hospitals. Without using machine learning algorithms or sharing data, hospitals or clinics cannot achieve the high accuracy prediction of disease.

In this chapter we use the Federated Learning methodology. FL is the type of machine learning. But in FL the client will not share its data with any entity and nor give access to own data. FL aims to train model over the decentralized distributed dataset. Data will reside in its premises and will only train the shared model and will update the global model. In FL there are two participants, one is the model aggregation server and the other is participated clients on distributed locations.

## 3.2 Datasets

This research is related to classification algorithm. So, we are looking for different datasets that have predicted columns in the form of 1 and 0. So we download some datasets from different sources.

### Heart Disease Dataset

This database consists of 76 attributes, although in all published experiments, researchers have used only a subset of 14 attributes. Notably, machine learning researchers have exclusively utilized the Cleveland database. The "objective" category signifies the existence of heart disease within the patient. The dataset was obtained from the *UCI Machine Learning Repository* [15]. The data set has 1025 entries, and it contains 14 different data columns, and the last 14<sup>th</sup> one column is the target column. Heart disease dataset attribute names and its units are mentioned in table 1. In the target attribute, a value of one indicates the presence of heart disease in the person, while a value of zero indicates the absence of heart disease.

### Breast Cancer Dataset

The dataset can be accessed through the UCI Machine Learning Repository [16]. This dataset is acquired from the sklearn dataset website. We use this dataset in our python code using “`from sklearn.datasets import load_breast_cancer`” library. This dataset contains 569 records. The diagnosis column has two entries one is malignant (M=malignant) and other is benign (B=benign). In the diagnosis attribute, the letter M indicates that the person has breast

cancer, while the letter B indicates that the person does not have breast cancer. We will remove the last column i.e. the 23<sup>rd</sup> one (Unnamed: 32) because we don't need this attribute.

### 3.3 Participants

After selecting the dataset, we must know about our clients. Clients can be a person, hospital, clinic, etc. In this we will check how many clients will participate in model training process and we also know about the resources through which user will update the model. Resources means either the processing speed of model training, speed of internet to upload model on main centralized server, etc. In our case we assume that all the clients will be honest and will not upload fake data. The number of participants will increase the accuracy of prediction. Each client will update the model in a specified time.

As in our case we don't have real time data of hospital because these data are very sensitive, and they have privacy concerns to share the data with any unknown entity. Therefore, we select different dataset from a trusted source, details about dataset are already explained. So, after selecting the dataset we divided our data set into three equal parts, each part will represent a participant.

### 3.4 Paillier Encryption Scheme

Paillier homomorphic encryption is a cryptographic technique that allows specific computations to be executed on encrypted data without the need for decryption. It is simpler and supports only one kind of computation. Pascal Paillier introduced the Paillier homomorphic encryption scheme in 1999, which falls under the realm of public key cryptography.

The main property that makes Paillier encryption homomorphic is its additive homomorphism, which allows two encrypted values to be combined into a new encryption of the sum of the original values. Using Paillier homomorphic encryption, it is possible to compute an encryption of the sum ( $a + b$ ) given only the encrypted values  $E(a)$  and  $E(b)$ , without having access to the plaintext values  $a$  and  $b$  [26][27].

The Paillier encryption scheme involves the following key components:

**Key Generation:**

- Select two  $p$  and  $q$  large prime numbers, such that  $p \neq q$ .
- Compute  $n = p \times q$ , where  $n$  is a composite number used as the public modulus.
- Compute  $\lambda(n) = lcm(p - 1, q - 1)$ ,  
the least common multiple of  $(p - 1)$  and  $(q - 1)$ .
- Choose a random integer  $g$  such that  $1 < g < n^2$  and  $g^n \bmod n^2$ .
- Ensure  $n$  divides the order of  $g$  by checking the existence of the following modular multiplicative inverse:  
$$\mu = (L(g^\lambda \bmod n^2))^{-1} \bmod n$$
 where function  $L$  is defined as (Lagrange function)  
$$L(u) = (u - 1)/n$$
- The public key is  $(n, g)$ , and the private key is  $\lambda(n)$ .

**Encryption:**

- To encrypt a plaintext message  $m$  ( $0 \leq m < n$ ), the sender generates a random  $r$  ( $0 \leq r < n$ ) and computes the ciphertext  $c$  as follows:
- $c = g^m \times r^n \bmod n^2$

**Decryption:**

- To decrypt the ciphertext  $c$  and obtain the plaintext message  $m$ , the receiver uses the private key  $\lambda(n)$ :
- $m = L(c^{\lambda(n)} \bmod n^2) * \mu \bmod n$   
where  $L(x) = (x - 1)/n$  and  $\mu = (L(g^{\lambda(n)} \bmod n^2))^{-1} \bmod n$

**Homomorphic properties:**

Additive homomorphism:  $E(a) * E(b) \bmod n^2$  results in  $E(a + b) \bmod n^2$ .

Scalar multiplication:  $E(a)^k \bmod n^2$  results in  $E(a * k) \bmod n^2$  for any integer  $k$ .

Paillier encryption is primarily used for privacy-preserving computations in scenarios like secure multiparty computation, privacy-preserving data analysis, and secure voting systems. The Paillier cryptosystem facilitates collaborative computation on encrypted data among multiple parties, ensuring that they can work together without disclosing individual

data points to one another. However, it's essential to be cautious about the potential performance trade-offs as homomorphic encryption can be computationally expensive compared to regular operations on plaintext data.

### **3.5 Machine Learning Algorithm**

Our research is based on predicting the final output of the given data. Our outcome will be based on two possible outcomes. Therefore we will use the binary classification algorithm. So, we will classify our final result into number of class or groups. Such as yes or no, true or false, 1 or 0, spam or not. For example predicting the heart disease of a patient, in this our final result will be in the form of yes or no, yes means have heart disease and no means haven't heart disease.

There are different binary classification algorithms. But some well-known algorithms are:

- Logistic regression (LR)
- Support Vector Machine
- Decision tree

A majority of researchers utilize one of these algorithms in their research paper. These algorithms are easy to implement. More details about these algorithms are already explained earlier. So, we will use these algorithms in our research and will compare their accuracy in terms of federated learning.

### **3.6 Traditional Machine Learning (Centralized)**

Centralized Machine Learning is a traditional machine learning approach where data from multiple sources is collected, aggregated, and processed in a central location. Machine learning refers to learning on its own without writing lengthy code to complete a task. Machine learning emphasizes code that gets large dataset and train itself on it using different machine learning algorithms. A large number of datasets and more experience will give us more accurate results. As large numbers of dataset will require more bandwidth to send data

on server and required more computational resources on both ends' users and server. More experience will give us more accurate results.

In this approach, a central server or a cluster of servers trains a machine learning model using the data from various sources. After training algorithm, we used the same algorithm is used for making decision, predictions or forecasting based on data.

The central server receives data from all the sources as shown in fig 7 and uses it to train a machine learning model. The trained model is then shared with all the sources to use for inference on their data. In this way, centralized machine learning enables organizations to train models on large datasets that are distributed across multiple locations. This machine learning technique has lot of security concerns because all the data are shifted to the centralized server, which is a major security concern for users.

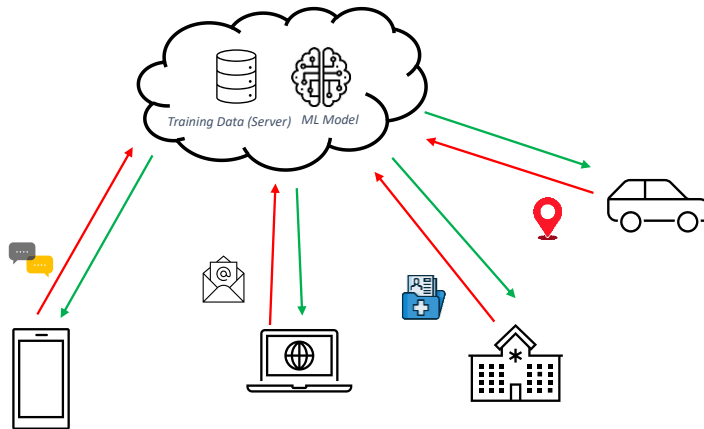


Figure 7 Traditional Machine Learning (Centralized)

### 3.7 Distributed On-Site Learning (Independent Learning)

Distributed on-site machine learning (ML) is gaining popularity due to the growing concerns over the risks of centralized data storage. On-site ML allows for the training, prediction, and inference of models based on live-streaming data, rather than sending data to a centralized cloud. This approach ensures that the data remains on the local devices, thereby preserving privacy.

In on-site machine learning (ML), a pre-trained or general ML model is distributed to devices via a server. Subsequently, each device tailors the model through local data-driven training. This process allows devices to conduct predictions specific to their data, engage in inferences for testing samples, and gain insights into the data generation process. The concept of on-device intelligence has found success across diverse applications, including but not limited to skin cancer detection, medical utilities, intelligent classrooms, and services aided by neural networks.

However, the downside of on-site ML is that the generated local models are limited to the user's experience without benefiting from peer's data as show in fig 8. Federated learning (FL) has been proposed to overcome this limitation by allowing users' computations to be federated while preserving privacy. In FL, multiple devices collaborate in training a shared model, and each device contributes to the model's improvement without sharing its data with others. FL provides a resolution to the constraints of on-site machine learning, all while guaranteeing the confidentiality of data.

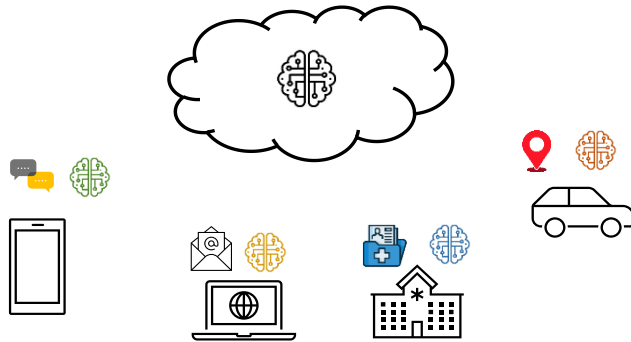


Figure 8 Distributed On-Site Learning (Independent Learning)

### 3.8 Federated Learning (Without Paillier Encryption)

Federated learning is a secure emerging methodology. In federated learning the user has complete autonomy over its own data, which increases the privacy protection of data owners. Federated learning is a distributed machine learning approach that enables training models on data that is distributed across multiple devices or edge nodes without requiring the data to be centrally collected on a server. Within the framework of federated learning,

the training procedure occurs on the client devices, while the central server is responsible for aggregating solely the model updates, as depicted in figure 9. It allows multiple parties such as devices or organizations of same domain to participate in the training process. This allows for the privacy-preserving training of machine learning models without the need to share sensitive data with a central entity.

Federated learning offers numerous benefits compared to conventional centralized machine learning methods. It permits model training using locally stored data, eliminating the necessity of transferring data to a central server. This approach mitigates the potential for data breaches and safeguards user privacy effectively. It also allows for the training of models in low-resource environments, such as mobile devices or edge nodes, where the network bandwidth and computing power are limited. Ultimately, this approach can enhance the generalization capabilities of machine learning models by encompassing the variety present in localized data, thus preventing the models from becoming overly tailored to a centralized dataset and avoiding overfitting.

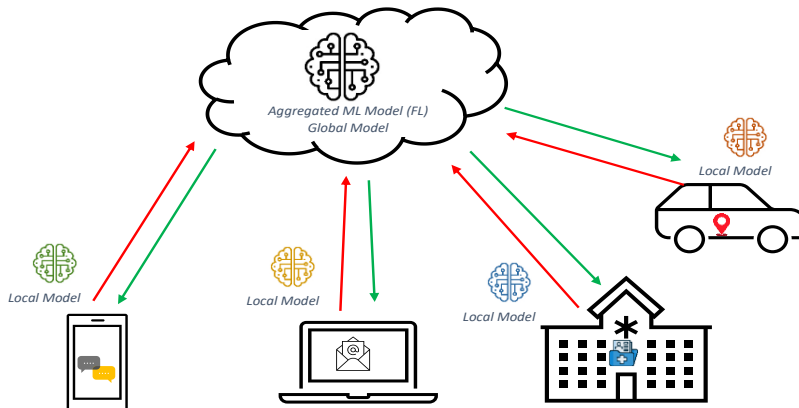


Figure 9 Federated Learning (Without Paillier Encryption)

### 3.9 Federated Learning Using Paillier Encryption

Federated learning is a secure emerging methodology as compared to traditional machine learning. It trains the model on data that is distributed on multiple devices or edge nodes without requiring the data to be centrally collected on a server. But after using federated learning technique, there are still it vulnerable against some attacks like



inference attack. If the main server is vulnerable or not honest then it has the possibility that the hacker can get back data from the weights that are share with the main server using inference attack.

Homomorphic encryption is a secure cryptosystem in which we can apply different computations on the ciphertext. There is no need to convert ciphertext into plaintext. Federated Learning using Paillier encryption is like the previous method, but the difference is that we send encrypted weights to the server as shown in fig 10. Paillier cryptosystem belongs to Partially Homomorphic Encryption (PHE). Paillier cryptosystem uses the additive property. In As we implemented these techniques in our code to increase the security of data. So that an unauthorized person cannot access or read sensitive information. As we know that after using the federated learning there are some threats to participants' leakage of private data. There can be different actors that can be involved in the training process, that can be a malicious server, insider adversary or outside adversary.

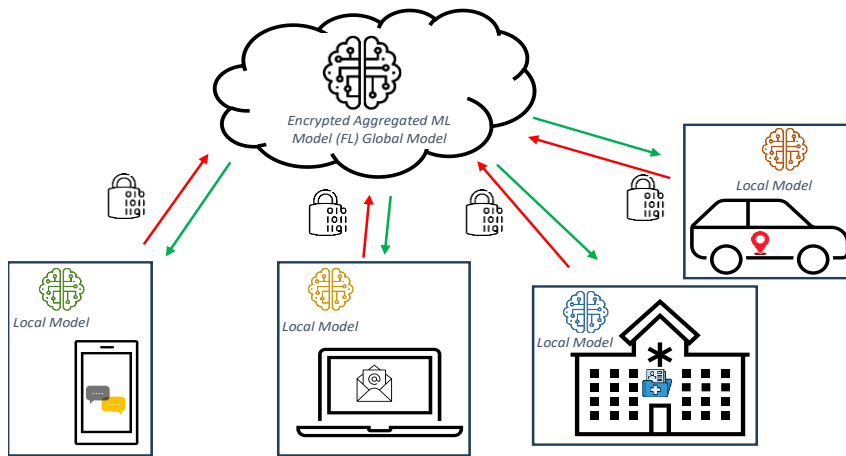


Figure 10 Federated Learning Using Paillier Encryption

# Security and Performance Analysis

## 4.1 Performance Measure Indices

The effectiveness of a model or algorithm is measured using performance indicators. Classification reports are generated to measure the performance and effectiveness of algorithms. Classification report contains accuracy, precision, f1-score, support, recall. The following are the formulas classification report. These are formulas are related with the confusion matrix in the table are as follows: [21] [22] [23]

- False Negative (FN): Instances that are predicted as negative but are actually positive (Type II error) [21].
- False Positive (FP): Instances that are predicted as positive but are actually negative (Type I error) [21].
- True Negative (TN): Instances that are correctly predicted as negative [21].
- True Positive (TP): Instances that are correctly predicted as positive [21].

### **Accuracy:**

In the context of a confusion matrix, accuracy is a performance metric that measures the overall correctness of a classifier's predictions. It is the ratio of correctly classified instances (both true positives and true negatives) to the total number of instances in the dataset [21]. Accuracy can be obtain using following formula:

$$Accuracy = \frac{(TP + TN)}{(TN + FP + TP + FN)}$$

### **Precision:**

Precision is a percentage that gauges the accuracy of positive predictions generated by a model. It is calculated by dividing the number of accurately predicted positive cases by the total count of positive predictions, yielding a measure of the model's ability to make correct positive identifications. It is also obtained using confusion matrix [21]. Precision can be obtained by using following formula:

$$Precision = \frac{TP}{(TP + FP)}$$

**Recall:**

Recall, expressed as a percentage, evaluates the model's capacity to identify all true positive cases among the total actual positive instances within the dataset. It is calculated by dividing the number of accurately predicted positive instances by the overall count of actual positive instances. A higher recall value signifies the model's effectiveness in capturing the majority of positive instances, thereby reducing instances of false negatives [21]. Recall can be obtained by using following formula:

$$Recall \text{ or } Sensitivity = \frac{TP}{(TP + FN)}$$

**F1-Score:**

The F1-score, also known as the F1 measure or F1 score, is a performance metric commonly used in binary classification tasks. It balances the trade-off between precision (the ability of the model to correctly identify positive instances) and recall (the ability of the model to capture all positive instances). The F1-score is the harmonic mean of precision and recall and is calculated using the following formula [21]:

$$F1 - Score = \frac{2(Precision \times Recall)}{(Precision + Recall)}$$

## 4.2 Confusion Matrix

A confusion matrix is a table used in machine learning to visualize the performance of a classification model on a set of data. It helps to understand how well the model's predictions align with the actual class labels. The confusion matrix is typically used for binary classification problems, but it can be extended to multi-class problems as well. The matrix is organized into four quadrants. Confusion matrix are represented in table 1, actual values represented on left side (horizontal) and predicted values are represented on the top of the table (vertical) [22].

		Predicted Values		
		Positive (P) +	Negative (N) -	
Actual Values	Positive (P) +	True Positive (TP)	False Negative (FN)	<b>Recall or Sensitivity</b> $= \frac{TP}{(TP + FN)}$
	Negative (N) -	False Positive (FP)	True Negative (TN)	<b>Specificity</b> $= \frac{TN}{(TN + FP)}$
		<b>Precision</b> $= \frac{TP}{(TP + FP)}$	<b>Negative Predictive value</b> $= \frac{TN}{(TN + FN)}$	<b>Accuracy</b> $= \frac{(TP + TN)}{(TP + TN + FP + FN)}$

Table 1 Confusion Matrix and its relationship with Classification Report

### 4.3 Breast Cancer Dataset Results

Breast cancer dataset is selected to predict the accuracy using different methods. First set the parameters to predict the accuracy of the model.

#### 4.3.1. Parameters and Values

Dataset = Breast Cancer

Clients =3

Key length=1024

Iterations=150

Learning Rate=0.05

Total no of Columns: 32

Entries in dataset (Rows): 569

#### 4.3.2. Breast Cancer Heatmap

A heatmap is a visual representation of data in a two-dimensional format, where distinct colors are utilized to depict various values along the x-axis and y-axis. It is used to



#### 4.3.4. Federated Learning (with paillier encryption)

The following are the results of the Federated learning using paillier encryption for breast cancer dataset. In this scenario, the encrypted data is transmitted to the centralized server. Paillier is a cryptographic scheme of homomorphic encryption. It attains more security as compared to other methodology, but its accuracy is less as compared to other and it takes more time for computation. The main property that makes Paillier encryption homomorphic is its additive homomorphism. Paillier encryption scheme makes our sensitive information more secure, but one of the major drawbacks is that it is computationally intensive and take more time (slow) in encryption. The accuracy of this is the number of correct predictions (or sum of all diagonal values of confusion matrix) divided by total number of predictions which becomes 94.40%.

$$Accuracy = \frac{135}{143} * 100 = 94.4056$$

Section 4.1 provides an explanation of precision, followed by the calculation. So, the precision of 1 is higher than 0 precision.

$$0 Precision = \frac{50}{56} * 100 = 89.2857$$

$$1 Precision = \frac{85}{87} * 100 = 97.7011$$

Section 4.1 provides an explanation of recall, followed by the calculation. So, the recall of 0 is higher than recall of 1.

$$0 Recall = \frac{50}{52} * 100 = 96.1538$$

$$1 Recall = \frac{85}{91} * 100 = 93.4066$$

Section 4.1 provides an explanation of f1-score, followed by the calculation. So, F1-Score of 1 is higher than F1-Score of 0.

$$0 F1 - Score = \frac{2(89.2857 * 96.1538)}{(89.2857 + 96.1538)} = 92.5926$$

$$1 F1 - Score = \frac{2(97.7011 * 93.4066)}{(97.7011 + 93.4066)} = 95.5056$$

Support refers to the count of real instances belonging to each class within the dataset. So, support of 0 is 52 and support of 1 is 91.

### Model Accuracy:

Average accuracy (federated training): 0.9441

Time taken to run federated training: 23.54388427734375

### CLASSIFICATION REPORT:

A classification report is a comprehensive summary of the performance of a classification model, often generated using metrics like precision, recall, F1-score, and support. Below table 2 represents all the performance measure indices for Federated learning using paillier encryption in tabular form that are calculated above.

	<b>0</b>	<b>1</b>	<b>Accuracy</b>	<b>Macro Avg</b>	<b>Weighted Avg</b>
Prec	0.892857	0.977011	0.944056	0.934934	0.946410
Rec	0.961538	0.934066	0.944056	0.947802	0.944056
F1-score	0.925926	0.955056	0.944056	0.940491	0.944463
Support	52.000000	91.000000	0.944056	143.000000	143.000000

Table 2 Classification Report of FL With paillier encryption

### Confusion Matrix:

Actual Values	<b>0</b>	50	2
	<b>1</b>	6	85
		<b>0</b>	<b>1</b>
		Predicted Values	

Table 3 Confusion Matrix of FL With paillier encryption

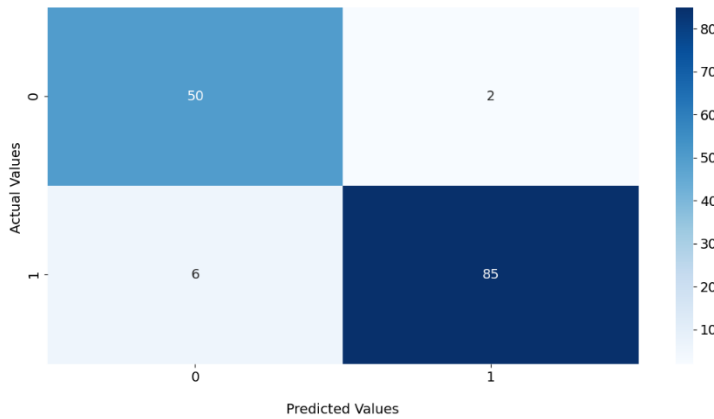


Figure 12 Confusion Matrix of FL With paillier encryption

### 4.3.5. Federated Learning (without paillier encryption)

The following are the results of the Federated learning without using paillier encryption for breast cancer dataset. In this we send the weights to the centralized server. This methodology is less secure compared to previous, because we send the weights in plaintext form. The accuracy of this is the number of correct predictions (or sum of all diagonal values of confusion matrix) divided by total number of predictions which becomes 95.10%.

$$Accuracy = \frac{136}{143} * 100 = 95.1049$$

Section 4.1 provides an explanation of precision, followed by the calculation. So, the precision of 1 is higher than 0 precision.

$$0 \text{ Precision} = \frac{51}{57} * 100 = 89.4737$$

$$1 \text{ Precision} = \frac{85}{86} * 100 = 98.8372$$

Section 4.1 provides an explanation of recall, followed by the calculation. So, the recall of 0 is higher than recall of 1.

$$0 \text{ Recall} = \frac{51}{52} * 100 = 98.0769$$



$$1 \text{ Recall} = \frac{85}{91} * 100 = 93.4066$$

Section 4.1 provides an explanation of f1-score, followed by the calculation. So, F1-Score of 1 is higher than F1-Score of 0.

$$0 \text{ F1 - Score} = \frac{2(89.4737 * 98.0769)}{(89.4737 + 98.0769)} = 93.5780$$

$$1 \text{ F1 - Score} = \frac{2(98.8372 * 93.4066)}{(98.8372 + 93.4066)} = 96.0452$$

Support refers to the count of real instances belonging to each class within the dataset. So, support of 0 is 52 and support of 1 is 91.

**Model Accuracy:**

Average accuracy (federated training): 0.9510

Time taken to run federated training: 4.620150327682495

**CLASSIFICATION REPORT:**

A classification report is a comprehensive summary of the performance of a classification model, often generated using metrics like precision, recall, F1-score, and support. Below table 4 represents all the performance measure indices for Federated learning without using paillier encryption in tabular form that are calculated above.

	<b>0</b>	<b>1</b>	<b>Accuracy</b>	<b>Macro Avg</b>	<b>Weighted Avg</b>
Prec	0.894737	0.988372	0.951049	0.941554	0.954323
Rec	0.980769	0.934066	0.951049	0.957418	0.951049
F1-score	0.935780	0.960452	0.951049	0.948116	0.951480
Support	52.000000	91.000000	0.951049	143.000000	143.000000

Table 4 Classification Report of FL Without paillier encryption

**Confusion Matrix:**

Actual Values	<b>0</b>	51	1
	<b>1</b>	6	85
		<b>0</b>	<b>1</b>
		Predicted Values	

Table 5 Confusion Matrix of FL Without paillier encryption

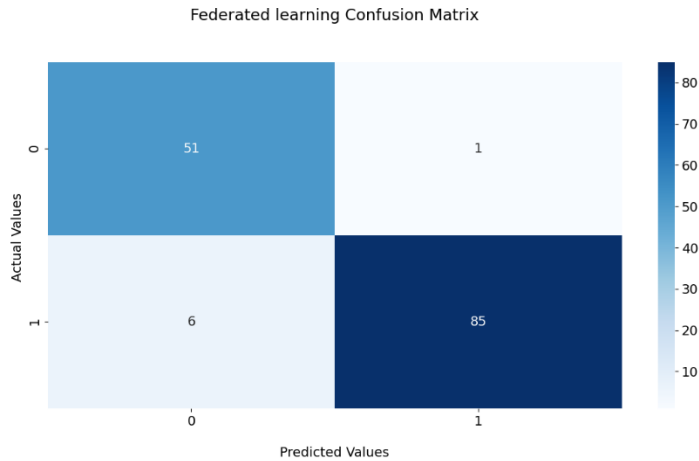


Figure 13 Confusion Matrix of FL Without paillier encryption

### 4.3.6. Centralized Machine Learning (Traditional ML)

Below are the outcomes obtained from applying centralized machine learning to the breast cancer dataset. In this we send the original data to the centralized server. This methodology is less secure compared to previous, because we send the weights in plaintext form. The accuracy of this is the number of correct predictions (or sum of all diagonal values of confusion matrix) divided by total number of predictions which becomes 95.80%.

$$Accuracy = \frac{137}{143} * 100 = 95.8042$$

Section 4.1 provides an explanation of precision, followed by the calculation. So, the precision of 1 is higher than 0 precision.

$$0 \text{ Precision} = \frac{50}{54} * 100 = 92.5926$$

$$1 \text{ Precision} = \frac{87}{89} * 100 = 97.7528$$

Section 4.1 provides an explanation of recall, followed by the calculation. So, the recall of 0 is higher than recall of 1.

$$0 \text{ Recall} = \frac{50}{52} * 100 = 96.1538$$

$$1 \text{ Recall} = \frac{87}{91} * 100 = 95.6044$$

Section 4.1 provides an explanation of f1-score, followed by the calculation. So, F1-Score of 1 is higher than F1-Score of 0.

$$0 \text{ F1 - Score} = \frac{2(92.5926 * 96.1538)}{(92.5926 + 96.1538)} = 94.3396$$

$$1 \text{ F1 - Score} = \frac{2(97.7528 * 95.6044)}{(97.7528 + 95.6044)} = 96.6667$$

Support refers to the count of real instances belonging to each class within the dataset. So, support of 0 is 52 and support of 1 is 91.

**Model Accuracy:**

Scikit-learn Logistic Regression Model

Accuracy: 0.9580

Time taken to run: 0.048969268798828125

**CLASSIFICATION REPORT:**

A classification report is a comprehensive summary of the performance of a classification model, often generated using metrics like precision, recall, F1-score, and support. Below table 6 represents all the performance measure indices for centralized machine learning in tabular form that are calculated above.

	<b>0</b>	<b>1</b>	<b>Accuracy</b>	<b>Macro Avg</b>	<b>Weighted Avg</b>
Prec	0.925926	0.977528	0.958042	0.951727	0.958764
Rec	0.961538	0.956044	0.958042	0.958791	0.958042
F1-score	0.943396	0.966667	0.958042	0.955031	0.958205
Support	52.000000	91.000000	0.958042	143.000000	143.000000

Table 6 Classification Report of Centralized ML

**Confusion Matrix:**

Actual Values	<b>0</b>	50	2
	<b>1</b>	4	87
		<b>0</b>	<b>1</b>
		Predicted Values	

Table 7 Confusion Matrix of Centralized ML

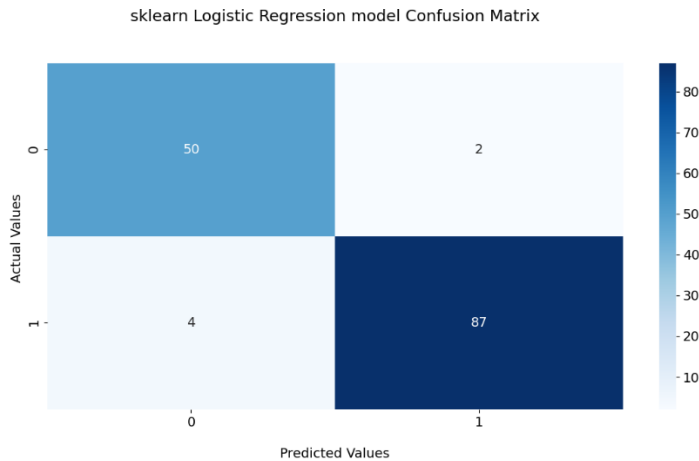


Figure 14 Confusion Matrix of Centralized ML

## 4.4 Heart Disease Dataset Results:

### 4.4.1. Parameters and Results

Dataset = Heart Disease

Clients = 3

Key Length = 1024

Iterations = 150

Learning Rate = 0.05

Total no of Columns: 14

Entries in dataset (Rows): 1025

### 4.4.2. Heart Disease Heatmap

A heatmap is a visual representation of data in a two-dimensional format. It is also called correlation heatmap. It is used to show the relationship between two variables. It is also called correlation heatmap. The range of the correlation heatmap extends from -1 to 1. In this range, a value of -1 denotes a state of perfect negative correlation, 0 signifies the absence of correlation, and a value of 1 indicates a state of perfect positive correlation. The value indicates the strength between two variables. It uses different colors to represent different values across the x-axis and y-axis. It is used to show the relationship between two variables. Columns are plotted on x-axis and y-axis. In below fig 15 minimum value of correlation is -0.4 and maximum is 1. Heatmap of heart disease dataset are shown in fig 15.

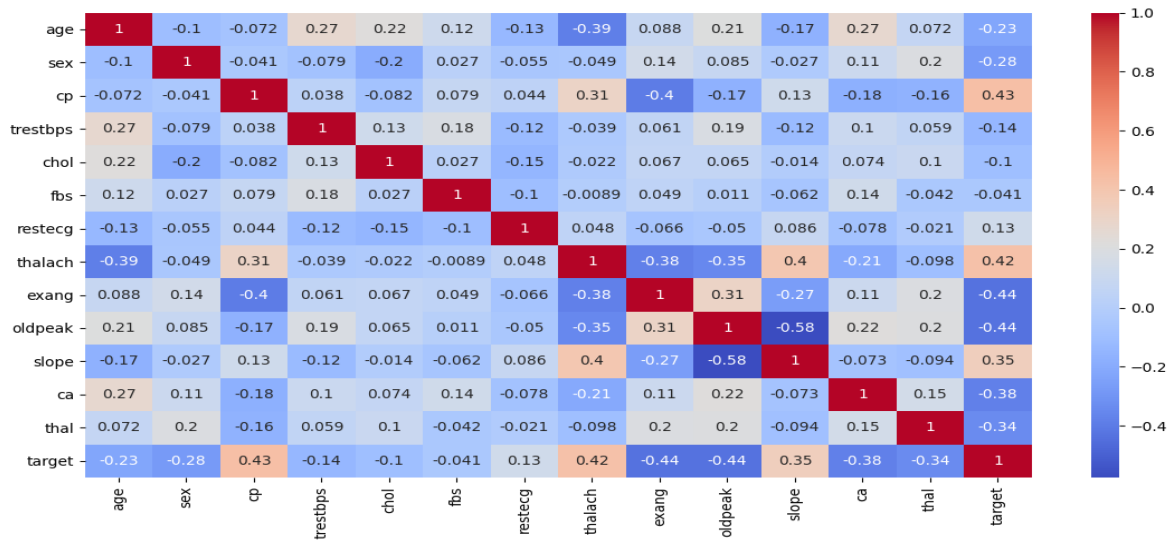


Figure 15 Heatmap of Heart Disease

#### 4.4.3. Distributed on-site learning (Independent Learning)

The following are the results of independent machine learning (on-site learning) for heart disease dataset. There are three clients that train their own model separately. The accuracy of client 1, client 2 and client 3 are 71%, 80% and 83% respectively.

Client 0 Accuracy: 0.71

Client 1 Accuracy: 0.80

Client 2 Accuracy: 0.83

Average accuracy (independent training): 0.7803

Time taken to run independent training: 0.04909634590148926

#### 4.4.4. Federated Learning (with paillier encryption)

The following are the results of the Federated learning using paillier encryption for heart disease dataset. The accuracy of this is the number of correct predictions (or sum of all diagonal values of confusion matrix) divided by total number of predictions which becomes 79.54%.

$$Accuracy = \frac{245}{308} * 100 = 79.5455$$

Section 4.1 provides an explanation of precision, followed by the calculation. So, the precision of 0 is higher than 1 precision.

$$0 \text{ Precision} = \frac{117}{144} * 100 = 81.2500$$

$$1 \text{ Precision} = \frac{128}{164} * 100 = 78.0488$$

Section 4.1 provides an explanation of recall, followed by the calculation. So, the recall of 1 is higher than recall of 0.

$$0 \text{ Recall} = \frac{117}{153} * 100 = 76.4706$$

$$1 \text{ Recall} = \frac{128}{155} * 100 = 82.5806$$

Section 4.1 provides an explanation of f1-score, followed by the calculation. So, F1-Score of 1 is higher than F1-Score of 0.

$$0 \text{ F1 - Score} = \frac{2(81.2500 * 76.4706)}{(81.2500 + 76.4706)} = 78.7879$$

$$1 \text{ F1 - Score} = \frac{2(78.0488 * 82.5806)}{(78.0488 + 82.5806)} = 80.2508$$

Support refers to the count of real instances belonging to each class within the dataset. So, support of 0 is 153 and support of 1 is 155.

### **Model Accuracy:**

Average accuracy (federated training): 0.7955

Time taken to run federated training: 16.066375970840454

### **CLASSIFICATION REPORT:**

A classification report is a comprehensive summary of the performance of a classification model, often generated using metrics like precision, recall, F1-score, and support. Below table 8 represents all the performance measure indices for Federated learning using paillier encryption in tabular form that are calculated above.

	0	1	Accuracy	Macro Avg	Weighted Avg
Prec	0.812500	0.780488	0.795455	0.796494	0.796390
Rec	0.764706	0.825806	0.795455	0.795256	0.795455
F1-score	0.787879	0.802508	0.795455	0.795193	0.795241
Support	153.000000	155.000000	0.795455	308.000000	308.000000

Table 8 Classification Report of FL With paillier encryption

### Confusion Matrix:

Actual Values	<b>0</b>	117	36
	<b>1</b>	27	128
		<b>0</b>	<b>1</b>
		Predicted Values	

Table 9 Confusion Matrix of FL With paillier encryption

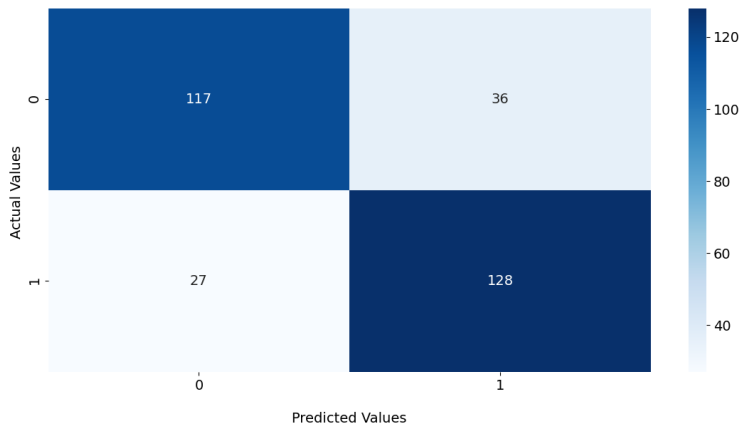


Figure 16 Confusion Matrix of FL With paillier encryption

#### 4.4.5. Federated Learning (without paillier encryption)

The following are the results of the Federated learning for breast heart disease. In this we send the weights to the centralized server in plaintext. The accuracy of this is the number of correct predictions (or sum of all diagonal values of confusion matrix) divided by total number of predictions which becomes 83.12%.

$$Accuracy = \frac{256}{308} * 100 = 83.1169$$

Section 4.1 provides an explanation of precision, followed by the calculation. So, the precision of 0 is higher than 1 precision.

$$0 \text{ Precision} = \frac{120}{139} * 100 = 86.3309$$

$$1 \text{ Precision} = \frac{136}{169} * 100 = 80.4734$$

Section 4.1 provides an explanation of recall, followed by the calculation. So, the recall of 1 is higher than recall of 0.

$$0 \text{ Recall} = \frac{120}{153} * 100 = 78.4314$$

$$1 \text{ Recall} = \frac{136}{155} * 100 = 87.7419$$

Section 4.1 provides an explanation of f1-score, followed by the calculation. So, F1-Score of 1 is higher than F1-Score of 0.

$$0 \text{ F1 - Score} = \frac{2(86.3309 * 78.4314)}{(86.3309 + 78.4314)} = 82.1918$$

$$1 \text{ F1 - Score} = \frac{2(80.4734 * 87.7419)}{(80.4734 + 87.7419)} = 83.9506$$

Support refers to the count of real instances belonging to each class within the dataset. So, support of 0 is 153 and support of 1 is 155.

### **Model Accuracy:**

Average accuracy (federated training): 0.8312

Time taken to run federated training: 2.2126266956329346

### **CLASSIFICATION REPORT:**

A classification report is a comprehensive summary of the performance of a classification model, often generated using metrics like precision, recall, F1-score, and support. Below table 10 represents all the performance measure indices for Federated learning without using paillier encryption in tabular form that are calculated above.



	<b>0</b>	<b>1</b>	<b>Accuracy</b>	<b>Macro Avg</b>	<b>Weighted Avg</b>
Prec	0.863309	0.804734	0.831169	0.834022	0.833831
Rec	0.784314	0.877419	0.831169	0.830867	0.831169
F1-score	0.821918	0.839506	0.831169	0.830712	0.830769
Support	153.000000	155.000000	0.831169	308.000000	308.000000

Table 10 Classification Report of FL Without paillier encryption

### Confusion Matrix:

Actual Values	<b>0</b>	120	33
	<b>1</b>	19	136
		<b>0</b>	<b>1</b>
		Predicted Values	

Table 11 Confusion Matrix of FL Without paillier encryption

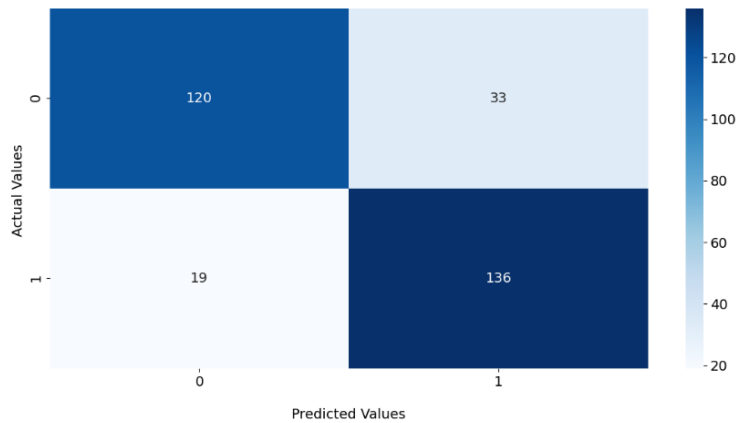


Figure 17 Confusion Matrix of FL Without paillier encryption

#### 4.4.6. Centralized Machine Learning (Traditional Machine Learning)

Results of the centralized machine learning for heart disease dataset are given below using different formulas. In this we send the original data to the centralized server. The accuracy of this is the number of correct predictions (or sum of all diagonal values of confusion matrix) divided by total number of predictions which becomes 85.71%.

$$Accuracy = \frac{264}{308} * 100 = 85.7143$$

Section 4.1 provides an explanation of precision, followed by the calculation. So, the precision of 0 is higher than 1 precision.

$$0 \text{ Precision} = \frac{124}{139} * 100 = 89.2086$$

$$1 \text{ Precision} = \frac{140}{169} * 100 = 82.8402$$

Section 4.1 provides an explanation of recall, followed by the calculation. So, the recall of 1 is higher than recall of 0.

$$0 \text{ Recall} = \frac{124}{153} * 100 = 81.0458$$

$$1 \text{ Recall} = \frac{140}{155} * 100 = 90.3226$$

Section 4.1 provides an explanation of f1-score, followed by the calculation. So, F1-Score of 1 is higher than F1-Score of 0.

$$0 \text{ F1 - Score} = \frac{2(89.2086 * 81.0458)}{(89.2086 + 81.0458)} = 84.9315$$

$$1 \text{ F1 - Score} = \frac{2(82.8402 * 90.3226)}{(82.8402 + 90.3226)} = 86.4198$$

Support refers to the count of real instances belonging to each class within the dataset. So, support of 0 is 153 and support of 1 is 155.

### **Model Accuracy:**

Scikit-learn Logistic Regression Model

Accuracy: 0.8571

Time taken to run: 0.03397989273071289

### **CLASSIFICATION REPORT:**

A classification report is a comprehensive summary of the performance of a classification model, often generated using metrics like precision, recall, F1-score, and

support. Below table 12 represents all the performance measure indices for centralized machine learning in tabular form that are calculated above.

	<b>0</b>	<b>1</b>	<b>Accuracy</b>	<b>Macro Avg</b>	<b>Weighted Avg</b>
Prec	0.892086	0.828402	0.857143	0.860244	0.860038
Rec	0.810458	0.903226	0.857143	0.856842	0.857143
F1-score	0.849315	0.864198	0.857143	0.856756	0.856805
Support	153.000000	155.000000	0.857143	308.000000	308.000000

Table 12 Classification Report of Centralized ML

### Confusion Matrix:

Actual Values	<b>0</b>	124	29
	<b>1</b>	15	140
		<b>0</b>	<b>1</b>
		Predicted Values	

Table 13 Confusion Matrix of Centralized ML

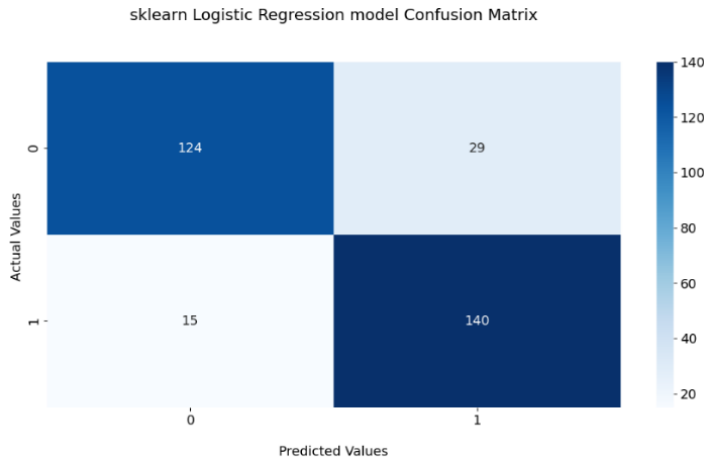


Figure 18 Confusion Matrix of Centralized ML

## 4.5 Comparison:

The following table 14 shows the comparison between two different datasets. In the following table centralized machine learning achieves highest accuracy whereas when we use privacy preserving paillier homomorphic scheme, it decreases the accuracy of the model as compared to other methodology due to computationally overhead. Homomorphic encryption requires more computational resources and time than conventional encryption. Graphical representation of both datasets is shown in figure 19 and 20.

	Breast Cancer Dataset		Heart Disease Dataset	
	Accuracy	Time	Accuracy	Time
Distributed on-site learning (Independent Learning)	0.9534	0.05696249008178711	0.7803	0.049096345901
Federated Learning (with paillier encryption)	0.9441	23.54388427734375	0.7955	16.066375970840454
Federated Learning (without paillier encryption)	0.9510	4.620150327682495	0.8312	2.2126266956329346
Centralized Machine Learning (Traditional Machine Learning)	0.9580	0.048969268798828125	0.8571	0.03397989273071289

Table 14 Accuracy Comparison of two datasets

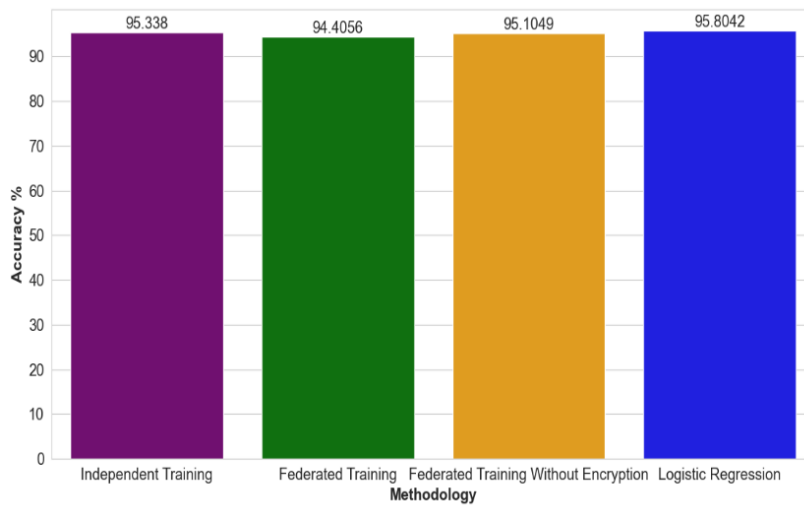


Figure 19 Accuracy Results Comparison of Breast Cancer Dataset

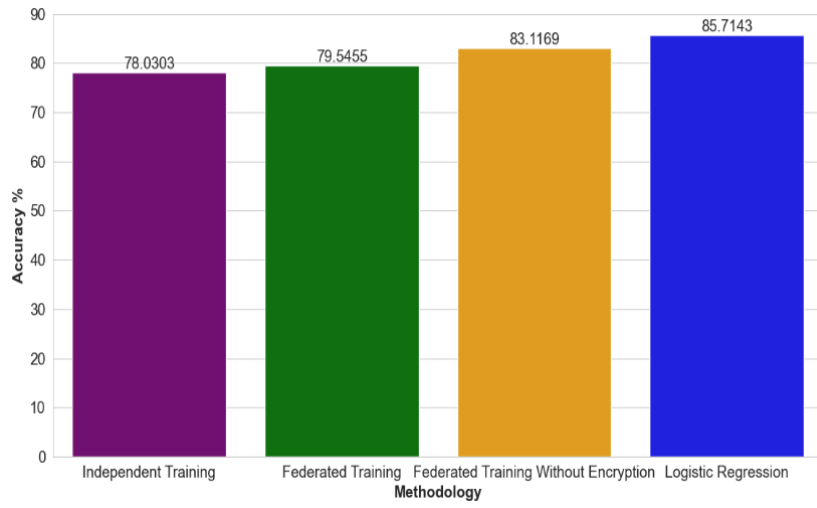


Figure 20 Accuracy Results Comparison of Heart Disease Dataset

# Conclusion

In this thesis we discussed different types of classified algorithms and implemented only one logistic regression algorithm using federated learning. Federated learning is an emerging secure methodology to train your algorithms, but it still has some security drawbacks. My proposed methodology used homomorphic encryption with federated learning for privacy preservation. We implement this methodology for two different datasets. After implementation we identify that the federated learning without paillier encryption gives us more accuracy as compared to federated learning with paillier encryption for both datasets. Paillier encryption takes more time as compared to simple federated learning. So, to gain higher accuracy with little bit security, simple federated learning will be suitable and for gaining higher security we need to use federated learning with homomorphic encryption.

## References

- [1] Zhang, Tuo, Lei Gao, Chaoyang He, Mi Zhang, Bhaskar Krishnamachari, and A. Salman Avestimehr. "Federated Learning for the Internet of Things: Applications, Challenges, and Opportunities." *IEEE Internet of Things Magazine* 5, no. 1 (2022): 24-29.
- [2] Liu, Ji, Jizhou Huang, Yang Zhou, Xuhong Li, Shilei Ji, Haoyi Xiong, and Dejing Dou. "From distributed machine learning to federated learning: A survey." *Knowledge and Information Systems* (2022): 1-33.
- [3] P Varalakshmi, K Narmadha, B Niveditha, A Akshaya, S K Sarah, "An Efficient Reliable Federated Learning Technology", *2022 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES)*, pp.1-5, 2022.
- [4] Asad, Muhammad, Ahmed Moustafa, and Takayuki Ito. "Federated Learning Versus Classical Machine Learning: A Convergence Comparison." *arXiv preprint arXiv:2107.10976* (2021).
- [5] Rauniyar, Ashish, Desta Haileselassie Hagos, Debesh Jha, Jan Erik Håkegård, Ulas Bagci, Danda B. Rawat, and Vladimir Vlassov. "Federated Learning for Medical Applications: A Taxonomy, Current Trends, Challenges, and Future Research Directions." *arXiv preprint arXiv:2208.03392* (2022).
- [6] Mothukuri, Virraaji; Parizi, Reza M.; Pouriye, Seyedamin; Huang, Yan; Dehghantaha, Ali; and Srivastava, Gautam, "A survey on security and privacy of federated learning" (2021). *Faculty Publications*. 5650.
- [7] Ng D, Lan X, Yao MM, Chan WP, Feng M. Federated learning: a collaborative effort to achieve better medical imaging models for individual sites that have small labelled datasets. *Quant Imaging Med Surg*. 2021 Feb;11(2):852-857. doi: 10.21037/qims-20-595. PMID: 33532283; PMCID: PMC7779924.

- [8] Lo, Sin Kit, Qinghua Lu, Liming Zhu, Hye-young Paik, Xiwei Xu, and Chen Wang. "Architectural patterns for the design of federated learning systems." *Journal of Systems and Software* 191 (2022): 111357.
- [9] M. Aledhari, R. Razzak, R. M. Parizi and F. Saeed, "Federated Learning: A Survey on Enabling Technologies, Protocols, and Applications," in *IEEE Access*, vol. 8, pp. 140699-140725, 2020, doi: 10.1109/ACCESS.2020.3013541.
- [10] X. Wang, J. Hu, H. Lin, W. Liu, H. Moon and M. J. Piran, "Federated Learning-Empowered Disease Diagnosis Mechanism in the Internet of Medical Things: From the Privacy-Preservation Perspective," in *IEEE Transactions on Industrial Informatics*, 2022, doi: 10.1109/TII.2022.3210597.
- [11] Nguyen, Dinh C., Quoc-Viet Pham, Pubudu N. Pathirana, Ming Ding, Aruna Seneviratne, Zihuai Lin, Octavia Dobre, and Won-Joo Hwang. "Federated learning for smart healthcare: A survey." *ACM Computing Surveys (CSUR)* 55, no. 3 (2022): 1-37.
- [12] Sinha, Nidhi, Teena Jangid, Amit M. Joshi, and Saraju P. Mohanty. "iCardo: A Machine Learning Based Smart Healthcare Framework for Cardiovascular Disease Prediction." *arXiv preprint arXiv:2212.08022* (2022).
- [13] S. Ray, "A Quick Review of Machine Learning Algorithms," *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, Faridabad, India, 2019, pp. 35-39, doi: 10.1109/COMITCon.2019.8862451.
- [14] Ibrahim, Ibrahim, and Adnan Abdulazeez. "The role of machine learning algorithms for diagnosing diseases." *Journal of Applied Science and Technology Trends* 2, no. 01 (2021): 10-19.
- [15] <https://archive.ics.uci.edu/ml/datasets/heart+disease>
- [16] [https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+\(diagnostic\)](https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic))
- [17] Wang, Haijin, Caomingzhe Si, and Junhua Zhao. "A federated learning framework for non intrusive load monitoring." *arXiv preprint arXiv:2104.01618* (2021).



- [18] Joshi, Madhura, Ankit Pal, and Malaikannan Sankarasubbu. "Federated learning for healthcare domain-pipeline, applications and challenges." *ACM Transactions on Computing for Healthcare* 3, no. 4 (2022): 1-36.
- [19] AbdulRahman, Sawsan, Hanine Tout, Hakima Ould-Slimane, Azzam Mourad, Chamseddine Talhi, and Mohsen Guizani. "A survey on federated learning: The journey from centralized to distributed on-site learning and beyond." *IEEE Internet of Things Journal* 8, no. 7 (2020): 5476-5497.
- [20] Melis, Luca, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. "Exploiting unintended feature leakage in collaborative learning." In *2019 IEEE symposium on security and privacy (SP)*, pp. 691-706. IEEE, 2019.
- [21] Ghosh, Pronab, Sami Azam, Mirjam Jonkman, Asif Karim, FM Javed Mehedi Shamrat, Eva Ignatious, Shahana Shultana, Abhijith Reddy Beeravolu, and Friso De Boer. "Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques." *IEEE Access* 9 (2021): 19304-19326.
- [22] Boukhatem, Chaimaa, Heba Yahia Youssef, and Ali Bou Nassif. "Heart disease prediction using machine learning." In *2022 Advances in Science and Engineering Technology International Conferences (ASET)*, pp. 1-6. IEEE, 2022.
- [23] F. M. Javed Mehedi Shamrat, P. Ghosh, M. H. Sadek, M. A. Kazi and S. Shultana, "Implementation of Machine Learning Algorithms to Detect the Prognosis Rate of Kidney Disease," 2020 IEEE International Conference for Innovation in Technology (INOCON), Bangluru, India, 2020, pp. 1-7, doi: 10.1109/INOCON50539.2020.9298026.
- [24] Bharati, Subrato, M. Mondal, Prajoy Podder, and V. B. Prasath. "Federated learning: Applications, challenges and future directions." *International Journal of Hybrid Intelligent Systems* 18, no. 1-2 (2022): 19-35.
- [25] Wen, Jie, Zhixia Zhang, Yang Lan, Zhihua Cui, Jianghui Cai, and Wensheng Zhang. "A survey on federated learning: challenges and

applications." *International Journal of Machine Learning and Cybernetics* 14, no. 2 (2023): 513-535.

- [26] T. Sridokmai and S. Prakancharoen, "The homomorphic other property of Paillier cryptosystem," 2015 International Conference on Science and Technology (TICST), Pathum Thani, Thailand, 2015, pp. 356-359, doi: 10.1109/TICST.2015.7369385.
- [27] R. Sendhil and A. Amuthan, "A Descriptive Study on Homomorphic Encryption Schemes for Enhancing Security in Fog Computing," 2020 International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 2020, pp. 738-743, doi: 10.1109/ICOSEC49089.2020.9215422.
- [28] M. Mohan, M. K. K. Devi and V. J. Prakash, "Homomorphic encryption-state of the art," 2017 International Conference on Intelligent Computing and Control (I2C2), Coimbatore, India, 2017, pp. 1-6, doi: 10.1109/I2C2.2017.8321774.
- [29] Vayadande, Kuldeep, Rohan Golawar, Sarwesh Khairnar, Arnav Dhiwar, Sarthak Wakchoure, Sumit Bhoite, and Darpan Khadke. "Heart Disease Prediction using Machine Learning and Deep Learning Algorithms." In *2022 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES)*, pp. 393-401. IEEE, 2022.