# Federated Learning for Healthcare Data Privacy

## 22AIE213


**TEAM 6**

CH.SC.U4AIE23008 – **Cheennepalli Mohith Reddy**
CH.SC.U4AIE23014 – **Faiz Rahaman**
CH.SC.U4AIE23019 – **Harinderan Thirumurugan**



AMRITA VISHWA VIDYAPEETHAM | School of Engineering
AMRITAPURI | BENGALURU | CHENNAI | COIMBATORE

# LITERATURE REVIEW

| S.NO | PAPER TITLE | METHODOLOGY | MERITS | DEMERITS | RESEARCH GAP |
|---|---|---|---|---|---|
| 1 | **Federated Learning in Healthcare: Enhancing Patient Privacy and Data Security**<br><br>**AUTHORS:**<br>Isabella Rossi<br><br>**DOI Link :**<br>10.5815/ijigsp.2018.02.02 | • The study evaluates different federated learning algorithms to determine their effectiveness in healthcare applications.<br>• To assess the impact of federated learning in healthcare, the paper presents a case study on predicting patient outcomes across multiple hospitals.<br>• The study involves | • Since federated learning allows model training without sharing raw patient data, it significantly reduces the risk of data breaches and ensures compliance with strict regulations like HIPAA and GDPR. Each institution retains control over its data, minimizing exposure to security threats.<br>• By aggregating | • Federated learning requires multiple rounds of communication between local institutions and the central server, leading to increased computational costs and network bandwidth usage. This can slow down model convergence compared to traditional centralized learning. | • Although federated learning prevents raw data sharing, advanced adversarial attacks, such as model inversion and gradient leakage, can still extract sensitive patient information |

# LITERATURE REVIEW

| S.NO | PAPER TITLE | METHODOLOGY | MERITS | DEMERITS | RESEARCH GAP |
|---|---|---|---|---|---|
| 2 | **Privacy Preserving Federated Learning in Medical Imaging with Uncertainty Estimation**<br><br>**AUTHORS:**<br>Nikolas Koutsoubisa, Yasin Yilmaz, Ravi P. Ramachandrand , Matthew Schabathb , Ghulam Rasool<br><br>**DOI Link :** | • Traditional machine learning models in medical imaging require centralized datasets, but due to strict privacy regulations like **HIPAA** and **GDPR**, direct data sharing is restricted. FL allows multiple healthcare institutions to collaboratively train models while keeping patient data decentralized. | • By leveraging federated learning, the model trains on decentralized medical imaging data without transferring raw images, ensuring compliance with privacy regulations like **HIPAA** and **GDPR**.<br>• The integration of **Bayesian Neural Networks (BNNs)** and **Monte Carlo Dropout (MC Dropout)** helps quantify prediction | • Federated learning requires frequent communication between local hospitals and a central aggregator, increasing network bandwidth usage and computational costs.<br>• Medical imaging datasets vary significantly | • While techniques like differential privacy and homomorphic encryption improve security, they often degrade model accuracy.<br>• Medical imaging datasets differ across hospitals due to variations in equipment, patient |

# LITERATURE REVIEW

| S.NO | PAPER TITLE | METHODOLOGY | MERITS | DEMERITS | RESEARCH GAP |
|---|---|---|---|---|---|
| 3 | **Federated learning in healthcare using structured medical data**<br><br>**AUTHORS:**<br>Wonsuk Oh<br>, Girish N Nadkarni<br><br>**DOI Link :**<br>10.5815/ijigsp.2018.02.02 | • This study explores Federated Learning (FL) as a solution to train machine learning models across multiple healthcare institutions while maintaining patient privacy.<br>• Each hospital trains a machine learning model on its local structured medical data, including patient | • FL allows multiple hospitals to collaborate on AI model training without sharing patient records, ensuring compliance with privacy regulations like HIPAA and GDPR.<br>• Training on diverse structured datasets from multiple healthcare | • FL requires frequent exchange of model updates between institutions and the central server, increasing network traffic and processing time.<br>• Structured medical data varies across hospitals due to different EHR systems, clinical | • FL requires high computational power and network bandwidth. Further studies should focus on lightweight FL architectures for hospitals with limited computing infrastructure.<br>• While differential privacy and |

# LITERATURE REVIEW

| S.NO | PAPER TITLE | METHODOLOGY | MERITS | DEMERITS | RESEARCH GAP |
|---|---|---|---|---|---|
| 4 | **Securing Healthcare Data: Federated Learning for Privacy-Preserving AI in Medical Applications**<br><br>**AUTHORS:**<br>Venkatesh Popuri<br><br>**DOI Link :**<br>10.5815/ijigsp.2018.02.02 | • The study utilizes Federated Learning (FL) to train machine learning models on decentralized healthcare data while preserving patient privacy.<br>• Instead of sharing sensitive medical data, local models are trained at different hospitals or healthcare institutions, and only model updates | • The combination of homomorphic encryption, differential privacy, and blockchain ensures that patient data remains secure while allowing collaborative model training.<br>• FL enables multiple healthcare institutions to contribute to model training without centralized data storage, making | • Homomorphic encryption and differential privacy add significant computational costs, making FL models slower and resource-intensive.<br>• While blockchain enhances security, it can introduce challenges such as high storage requirements and | • There is a need to develop lightweight encryption and privacy-preserving methods that offer strong security without significantly increasing computational costs. |

# LITERATURE REVIEW

| S.NO | PAPER TITLE | METHODOLOGY | MERITS | DEMERITS | RESEARCH GAP |
|---|---|---|---|---|---|
| 5 | **The federated learning framework in healthcare employs a decentralized training approach, allowing multiple medical institutions to collaboratively train machine learning models without exposing raw patient data.**<br>**AUTHORS:**<br>Anand G Buddhikot, Nitin M. Kulkarni, | •Design a hand gesture interface for rear seat passengers to interact with an automotive infotainment system's media player.<br>•Use the Hi-Vision (Hi-Vi) algorithm for skin detection based on color spaces (RGB, HSV, YCbCr).<br>•Capture images of hand gestures using a webcam.<br>•Enhance contrast and intensity of images | • Provides a hands-free method for passengers to control the infotainment system, reducing distractions for the driver.<br>• Allows rear seat passengers to interact with the system without needing to rely on the driver.<br>• The Hi-Vi classifier | • The system may struggle with varying illumination and complex backgrounds, which can affect skin detection accuracy.<br>• Users may need to learn specific gestures to interact effectively with the system.<br>• The effectiveness of the system is | Despite significant advancements in automotive infotainment systems, the integration of hand gesture interfaces based on skin detection techniques still faces several challenges. Existing methods often struggle with varying lighting conditions, skin |

# LITERATURE REVIEW

| S.NO | PAPER TITLE | METHODOLOGY | MERITS | DEMERITS | RESEARCH GAP |
|---|---|---|---|---|---|
| 6 | **Securing Health Data on the Blockchain: A Differential Privacy and Federated Learning Framework, 2024**<br><br>**AUTHORS:**<br>Daniel Commey, Sena Hounsinou, Garth V. Crosby<br><br>**DOI Link :**<br>10.48550/arXiv.2405.11580 | •The study integrates Differential Privacy (DP) with Federated Learning (FL) to protect sensitive health data collected by IoT devices.<br>•It employs dynamic personalization and adaptive noise distribution strategies to balance privacy and data utility.<br>•Blockchain technology is utilized for secure and transparent aggregation and storage of model | • The framework enhances patient privacy while allowing for effective health data analytics.<br>• It successfully combines multiple advanced technologies (FL, DP, and blockchain) to create a comprehensive solution.<br>• The use of dynamic | • The framework's performance is evaluated on a dataset (SVHN) that is not directly related to healthcare, which may limit its applicability.<br>• The complexity of integrating multiple technologies may pose challenges in real-world implementation.<br>• The reliance on | There is a need for further exploration of the framework's effectiveness on actual healthcare datasets to validate its applicability.<br>The dynamic nature of healthcare data and the need for adaptive privacy mechanisms have not been fully addressed.<br>More comprehensive |

# LITERATURE REVIEW

| S.NO | PAPER TITLE | METHODOLOGY | MERITS | DEMERITS | RESEARCH GAP |
|---|---|---|---|---|---|
| 7 | **Securing Healthcare Data: Federated Learning for Privacy-Preserving AI in Medical Applications, 2023**<br><br>**AUTHORS:**<br>Venkatesh Popuri<br><br>**DOI Link :**<br>https://doi.org/10.37745/ijmt.2013/vol11n36482 | •Utilizes Federated Learning (FL) to train machine learning models on decentralized healthcare data, ensuring that sensitive patient information remains local and is not shared.<br>•Involves data preprocessing steps such as filtering out uncertainties, dimensionality reduction using techniques like AutoEncoders, and | • Enhances patient privacy by keeping sensitive data decentralized, reducing the risk of data breaches.<br>• Facilitates collaboration among multiple healthcare institutions, allowing them to improve model accuracy without compromising individual patient | • There are potential privacy risks associated with sharing model parameters, which could lead to information leakage if not properly managed.<br>• The implementation of privacy-preserving algorithms can be complex and may | • There is a need fo further investigation into vertical federated learning approaches, whic could enhance da utilization while preserving privacy<br>• More research is required to develo advanced privacy-preserving techniques that ca effectively mitigat risks without |

# LITERATURE REVIEW

| S.NO | PAPER TITLE | METHODOLOGY | MERITS | DEMERITS | RESEARCH GAP |
|---|---|---|---|---|---|
| 8 | **Survey of Medical Applications of Federated Learning, 2024**<br><br>**AUTHORS:**<br>Geunho Choi, Won Chul Cha, Se Uk Lee, Soo-Yong Shin<br><br>**DOI Link :**<br>https://doi.org/10.4258/hir.2024.30.1.3 | •Conducted a literature search using keywords related to federated learning in medical contexts on Google Scholar and PubMed, resulting in the selection of 58 relevant studies.<br>•Categorized the selected studies based on data types, target diseases, dataset openness, local models, and neural network models used.<br>•Examined issues related | • Provides enhanced privacy protection for sensitive medical data by allowing model training without sharing original data.<br>• Facilitates collaborative research across multiple institutions, enabling the | • Limited focus on diverse data types and diseases, with most studies concentrating on specific areas like cancer and COVID-19.<br>• Potential biases due to the small number of clients in medical federated learning, which can affect the generalizability of | • Underrepresentation of certain data types and diseases in federated learning studies, particularly in cardiovascular and neurological disorders.<br>• Need for more effective solutions to handle non-IID data, which can impact model |

# LITERATURE REVIEW

| S.NO | PAPER TITLE | METHODOLOGY | MERITS | DEMERITS | RESEARCH GAP |
|---|---|---|---|---|---|
| 9 | **Federated Learning Approach to Protect Healthcare Data over Big Data Scenario, 2021**<br><br>**AUTHORS:**<br>Gaurav Dhiman, Sapna Juneja, Hamidreza Mohafez, Ibrahim El-Bayoumy, Lokesh Kumar Sharma, Maryam Hadizadeh, Mohammad Aminul Islam, Wattana Viriyasitavat, Mayeen Uddin Khandaker<br><br>**DOI Link :**<br>https://doi.org/10.3390/su14052500 | • The paper discusses a federated learning framework that allows multiple medical institutions to collaboratively train machine learning models without sharing sensitive patient data.<br>• It emphasizes the use of privacy-preserving techniques such as differential privacy and secure multiparty computation to protect patient information | • Enhances patient privacy by ensuring that sensitive data remains within local institutions while still contributing to a shared learning model.<br>• Improves the accuracy of medical data analysis by leveraging diverse datasets from multiple sources without | • The complexity of implementing federated learning can be a barrier for smaller healthcare institutions with limited resources.<br>• There may be challenges in achieving consensus among different institutions regarding data sharing policies | • There is a need for more empirical studies to evaluate the real-world effectiveness of federated learning in diverse healthcare settings.<br>• Further exploration is required on how to balance privacy protection with the usability of |

# LITERATURE REVIEW

| S.NO | PAPER TITLE | METHODOLOGY | MERITS | DEMERITS | RESEARCH GAP |
|---|---|---|---|---|---|
| 10 | **A Comprehensive Survey on Federated Learning in the Healthcare Area: Concept and Applications, 2022**<br><br>**AUTHORS:**<br>Deepak Upreti, Eunmok Yang, Hyunil Kim, Changho Seo<br><br>**DOI Link :**<br>10.32604/cmes.2024.048932 | •The article conducts a systematic review of healthcare federated learning (FL), contrasting various approaches and their performance. It presents an overview of recent literature on FL, specifically its use in medical imaging, electronic health records, and disease prediction.<br>•Thestudy concentrates on the manner in which | • FL allows models to be trained without exposing sensitive patient data, enhancing data privacy.<br>• It enables healthcare organizations to distribute and develop diagnostic materials without violating the confidentiality of the patients.<br>• FL can handle large | • There remain unresolved technical challenges in applying FL effectively in healthcare, particularly with respect to data integration and standardization.<br>• More research needs to be done to explore FL applications in medical | There remain unresolved technical challenges in applying FL effectively in healthcare, particularly with respect to data integration and standardization. More research needs to be done to explore FL applications in medical imaging and real-time |

# Problem Identification

•**Privacy Issues of Data** – Centralized training using the sharing of patient medical records creates security issues and regulatory issues.

•**Decentralized Healthcare Data** – Healthcare data is spread in many hospitals and institutions such that centralized learning is impossible.

•**Computational Constraints** – Hospitals can have limited computational resources, and the learning process should be distributed and efficient.

•**Federated Learning Need** – A decentralized AI model can enable institutions to jointly train a model without sharing the data and keeping it local and secure.

# Problem Statement

- Heart disease is one of the biggest killers globally, and therefore, early diagnosis and correct prediction are necessary for treatment.
- Traditional centralized machine learning frameworks require data to be stored in one location, thus increasing privacy concerns and the risk of data breaches.
- The necessity for a privacy-preserving artificial intelligence model that can effectively predict heart disease without violating patient sensitive data is evident.

# Proposed Solution

- **Federated Learning for Private-Preserving Training** – Employ a decentralized AI strategy where various clients (hospitals) train the models locally without exposing patient data, preserving data privacy and security.

- **Deep Learning Model for Predicting Heart Disease** – Employ a neural network in TensorFlow to predict heart disease risk from patient medical attributes with great accuracy and effectiveness.

- **Federated Averaging to Improve Global Models** – Federated Averaging of model updates from several clients, enabling collaborative learning with data remaining local and secure.

# Methodology

- **Data Acquisition** – Used a heart disease dataset that contains 900 records with quantitative and qualitative features.
  **Data Preprocessing** – Implemented one-hot encoding, feature scaling, and dropped unnecessary columns to tune the model.
- **Client Data Partitioning** – Divide the data set into five sets in order to mimic various clients for Federated Learning.
- **Model Development** – A dense layer neural network was constructed with TensorFlow to facilitate binary classification.
- **Federated Training** – Trained client models individually and applied federated averaging to aggregate global model updates.
- **Model Evaluation** – Assessed model performance in terms of accuracy, precision, F1-score, and confusion matrix.

# Hardware and Software Requirements

**Hardware Requirements:**

- **Processor:** AMD Ryzen 5 / 7 or higher (Supports efficient model training)
- **RAM:** Minimum **8GB** (Recommended **16GB** for smooth execution)
- **Storage:** At least **5 GB free space** (For dataset, models, and logs)
- **GPU (Optional):** NVIDIA GTX 1650 / RTX 3060 or higher (For faster model training)

**Software Requirements:**

- **Operating System:** Windows 10 / 11 or Ubuntu 20.04+
- **Programming Language:** Python 3.8+
- **Development Environment:** VS Code (Preferred) / Jupyter Notebook
- **Libraries & Frameworks:** TensorFlow, NumPy, Pandas, Matplotlib, Seaborn, Joblib
- **Virtualization (Optional):** Docker (For scalable FL deployment)

# Datasets Used

- Heart Disease UCI

https://www.kaggle.com/code/rizwanrizwannazir/heart-disease-prediction-best-model-selection/input?select=heart_disease_uci.csv

This is a multivariate type of dataset which means providing or involving a variety of separate mathematical or statistical variables, multivariate numerical data analysis. It is composed of 14 attributes which are age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise-induced angina, oldpeak — ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, number of major vessels and Thalassemia. This database includes 76 attributes, but all published studies relate to the use of a subset of 14 of them. The Cleveland database is the only one used by ML researchers to date. One of the major tasks on this dataset is to predict based on the given attributes of a patient that whether that particular person has heart disease or not and other is the experimental task to diagnose and find out various insights from this dataset which could help in understanding the problem more.

# Methodology Diagram