Privacy Preserving Machine Learning in Healthcare for Pandemic Prediction using Genomic Data

Thesis Submitted
In Partial Fulfilment of the Requirements for the Degree of

MASTER OF TECHNOLOGY
in
Data Science
By

Riti Rathore
(2K23/DSC/26)

Under the Supervision of
Dr. Abhilasha Sharma
(Associate Professor)

To the
Department of Software Engineering
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-110042, India

May 2025

ACKNOWLEDGEMENT

I would like to express my deep appreciation to **Dr. Abhilasha Sharma**, Associate Professor at the Department of Software Engineering, Delhi Technological University, for her invaluable guidance and unwavering encouragement throughout this research. Her vast knowledge, motivation, expertise, and insightful feedback have been instrumental in every aspect of preparing this research plan.

I am also grateful to **Prof. Ruchika Malhotra**, Head of the Department, for her valuable insights, suggestions, and meticulous evaluation of my research work. Her expertise and scholarly guidance have significantly enhanced the quality of this thesis.

My heartfelt thanks go out to the esteemed faculty members of the Department of Software Engineering at Delhi Technological University. I extend my gratitude to my colleagues and friends for their unwavering support and encouragement during this challenging journey. I have had some friends that I am thankful to be around. They made me feel truly at home. I would like to thank **Miss Roshni Singh** whom I had such a great time. The last two years were great with such a lovely bunch of people around me. Their intellectual exchanges, constructive critiques, and camaraderie have enriched my research experience and made it truly fulfilling.

While it is impossible to name everyone individually, I want to acknowledge the collective efforts and contributions of all those who have been part of this journey. Their constant love, encouragement, and support have been indispensable in completing this MTech thesis.

Riti Rathore
(2K23/DSC/26)

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

CANDIDATE DECLARATION

I RITI RATHORE (2K23/DSC/26) hereby certify that the work which is being presented in the thesis entitled "**Privacy Preserving Machine Learning in Healthcare for Pandemic Prediction using Genomic Data"** in partial fulfillment of the requirements for the award of the Degree of Master of Technology submitted in the Department of Software Engineering, Delhi Technological University in an authentic record of my work carried out during the period from August 2023 to May 2025 under the supervision of Dr. Abhilasha Sharma.

The matter presented in the thesis has not been submitted by me for the award of any other degree of this or any other Institute.

Riti Rathore

This is to certify that the student has incorporated all the corrections suggested by the examiner in the thesis and that the statement made by the candidate is correct to the best of our knowledge.

Signature of Supervisor(s) **Signature of External Examiner**

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

CERTIFICATE BY THE SUPERVISOR

Certified that Riti Rathore (2K23/DSC/26) has carried out their project work presented in this thesis entitled "**Privacy Preserving Machine Learning in Healthcare for Pandemic Prediction using Genomic Data**" for the award of **Master of Technology** from the Department of Software Engineering, Delhi Technological University, Delhi under my supervision. The thesis embodies the results of original work, and studies are carried out by the student herself, and the contents of the thesis do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/Institution.

Abhilasha Sharma

Dr. Abhilasha Sharma
Associate Professor
Department of Software Engineering,
Delhi Technological University -Delhi, India

Date:

Privacy Preserving Machine Learning in Healthcare for Pandemic Prediction using Genomic Data

Riti Rathore

ABSTRACT

The demand for analyzing healthcare data can be attributed to the curiosity for personalized prediction, treatment, and monitoring. The rapid growth of healthcare information by the demand of healthcare requires better strategies for healthcare data analysis. However, while healthcare data analytics has been proposed to combination of information, network expertise, mixed in models that are trained on this private information; the use of late deep systems, improving an architecture time and complexity. Healthcare data is information about a patient's healthcare status. It includes various types of data such as structured and non-structured, or private and national healthcare data. The global movement of having health data for the public is producing many initiatives. While healthcare data analytics have demonstrated some promising results, there are still challenges, particularly in models trained in private data. Privacy Preserving Deep Learning techniques in the healthcare domain addressed the critical challenge of protecting the privacy of the patient and ensuring the judicious usage of data for models in machine learning. In this research, we have discussed comparative study of the key techniques which involve Federated Learning, Differential Privacy, Homomorphic Encryption, Secure Multi-party Computation and Synthetic Data Generation. These techniques will provide robust solutions for data-confidentiality and secure model training. This also discusses the amalgamation of these advanced technologies with regulatory compliance, which helps in emphasizing the potential of balancing innovation with ethical responsibility to transform healthcare.

In recent times, there has been a rapid spread of pandemics caused by rapidly mutating viruses, such as SARS-CoV-2 which has present significant challenges for healthcare systems worldwide. The global health crises like COVID-19 underscore the need for predictive models that support containment and resource management. Genomic data is very crucial in providing critical insights into viral evolution and the mechanics of dynamics. Genomic datasets contain information that requires such computational methods that protect privacy. We have used federated deep learning architecture using genomic data for the pandemic prediction. We have achieved both data privacy by identifying key genomic features and implementing federated learning and robust model performance. Our results help in demonstrating the effectiveness of the method proposed by offering a scalable solution for the monitoring of pandemics.

Keywords: Healthcare Data, Privacy Preserving, Machine Learning, Federated Learning (FL), Data encryption, Deep Learning, Genomic data, Genome Sequence, Differential Privacy.

TABLE OF CONTENT

LIST OF TABLE(S)

LIST OF FIGURE(S)

LIST OF ABBREVIATION(S)

ML	Machine Learning
COVID-19	Coronavirus Disease of 2019
DL	Deep Learning
FL	Federated Learning
HIPAA	Health Insurance Portability and Accountability Act
GDPR	General Data Protection Regulation
AI	Artificial Intelligence
SVM	Support Vector Machine
IID	Independent and Identically Distributed
DP	Differential Privacy
ReLU	Rectified Linear Unit
HE	Homomorphic Encryption
ECG	Electrocardiogram
PPML	Privacy Preserving Machine Learning
SMPC	Secure Multi-Party Computation
GAN	Generative Adversarial Networks
LSTM	Long Short-Term Memory
SPICE	Simulation Program with Integrated Circuit Emphasis
EHR	Electronic Health Record
CNN	Convolutional Neural Network
RNA	Ribonucleic acid
ACE2	Angiotensin-Converting Enzyme 2
MAFFT	Multiple Alignment using Fast Fourier Transform
MUSCLE	Multiple Sequence Comparison by Log-Expectation
PCA	Principal Component Analysis
LSTM	Long-Short Term Memory
AUC ROC	Area Under the Receiver Operating Characteristic Curve
MAE	Mean Absolute Error

CHAPTER 1

INTRODUCTION

1.1 Overview

In typical data-driven machine learning tasks, a model is trained on a sample of input-output pairs that have been collected and labeled in advance. Such freedom of data collection and labeling is not achievable in many healthcare applications due to strict regulations, ethical considerations, and the potential privacy loss associated with revealing confidential medical information. In the context of health-related ML model training, specialized approaches that ensure differential privacy or fully homomorphically cryptic techniques have been developed to facilitate the joint analysis of distributed data repositories, allowing cross-organization collaborations without violating privacy guidelines. In addition, horizontal or vertical differentially private distributed algorithms have been shown effective in model training for biomedical image data analysis, demonstrating utility in privacy-preserving multi-center collaborations. Considering widespread interest in privacy-preserving ML techniques for healthcare data among the computer science, biomedical, and health economics community, a field review on this fast-developing research area is both timely and critically needed.

Recent years have witnessed a surge of machine learning (ML) applications in healthcare, ranging from predictive algorithms for patient management to analysis tools for next-generation sequencing data. Large-scale healthcare databases, including electronic health records, must be widely and democratically accessible for these advancements to occur. There is a potential privacy risk of disclosing medical information to unapproved parties has also grown in importance, as evidenced by the sharp rise in medical data breaches over the last ten years. Consequently, a critical challenge of using healthcare data for ML applications is how to unlock the full potential of large-scale healthcare datasets while enabling the privacy protection of patients and the secrecy of sensitive medical data. In this survey, we aim to shed light on the emerging research area termed privacy-preserving machine learning for healthcare information.

Genomic analysis plays a critical role in understanding viral mutations, resistance patterns and their spread. Deep learning techniques excel in extracting meaningful patterns from the high-dimensional and complex genomic data. The COVID-19 pandemic created the necessity for reliable predictive models to effectively manage and control such health crises. However, the sharing of genomic data between institutions raises significant ethical and privacy concerns.

A solution is offered by federated learning (FL) by enabling the decentralized training of machine learning models, allowing institutions to collaborate and develop predictive models without the need to exchange sensitive genomic data. With the help of genomic sequencing, we have predicted the pandemic using federated learning; it has ensured data privacy while identifying the critical features from the genomic dataset, and hence a deep learning model was built.

1.2 Problem Statement

The problem is exacerbated by the fact that large-scale hospitals and/or insurance companies have little legal framework forcing them to handle healthcare data in a safe and ethical way. Today, such companies freely exchange their patient healthcare data to increase their gains. The primary focus of this book is to present secure ML algorithms that can extract accurate models without violating patients' privacy.

The pervasive use of electronic healthcare data in hospitals, clinics, and mobile health applications has led to an increasing interest from the healthcare industry in leveraging machine learning (ML) technologies for better understanding patient data and improving healthcare outcomes. The use of modern ML algorithms presents new privacy challenges, however. The standard ML pipeline involves the collection of sensitive healthcare data in a server that applies ML algorithms to the data, and the resulting models are sent back to each of the hospitals or mobile devices for local predictions. As a result, adversaries that can access the collected healthcare data, even after the server has received the model, are able to perform reverse engineering to extract sensitive patient data for illicit purposes.

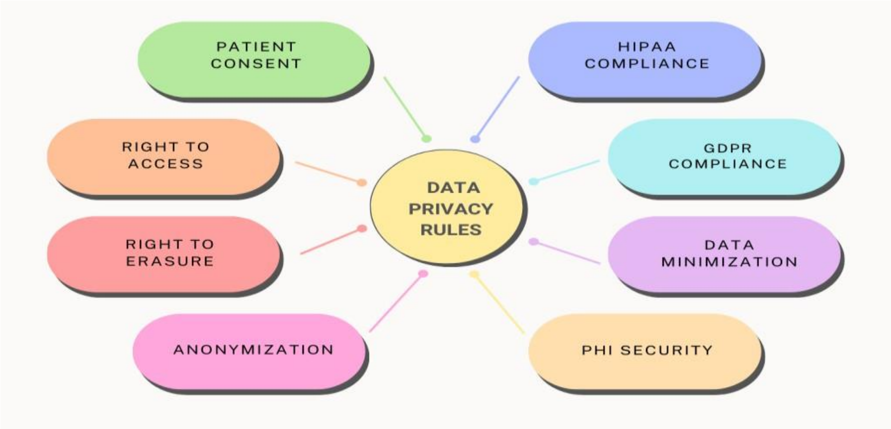


Figure 1.1. Comprehensive Healthcare Data Privacy Rules [1]

With the rise of the sensitive data in the world, it has created concerns over confidentiality of it. Even after a pandemic like COVID-19 opened the gateway for highly confidential medical data, which has increased the demand for privacy preservation of it. Hence, the federated learning algorithm has been widely used as a privacy-preserving machine learning technique in the domain of health and medical data.

CHAPTER 2

PRIVACY PRESERVING TECHNIQUES IN AI

Privacy laws, such as HIPAA [2], are pervasive and are designed to protect the rights of individuals while allowing them to seek medical treatment when available. It is important to accommodate the regulations explicitly when analyzing healthcare data. By allowing healthcare data models to be built in accordance with privacy regulations, the potential for model development will be increased without breaking the laws. The benefits of healthcare AI and ML are significant. Technologies can provide patient outcomes, value-based care, and healthcare population management across the healthcare space. Optimally designed and implemented privacy-preserving machine learning systems [3] that utilize AI algorithms can make accessing, processing, and analyzing healthcare information much simpler. More efficient information sharing can create access to complex health model data which will make training on new data more successful.

There are major challenges that need to be overcome to make healthcare AI and ML a reality [4]. One of the primary challenges is data privacy and protection from data breaches. Large-scale adoption and success of AI and ML healthcare systems are impeded by the lack of effective ways to accommodate privacy regulations. The most straightforward way to protect sensitive healthcare information is to reduce the amount of accessible data [5].

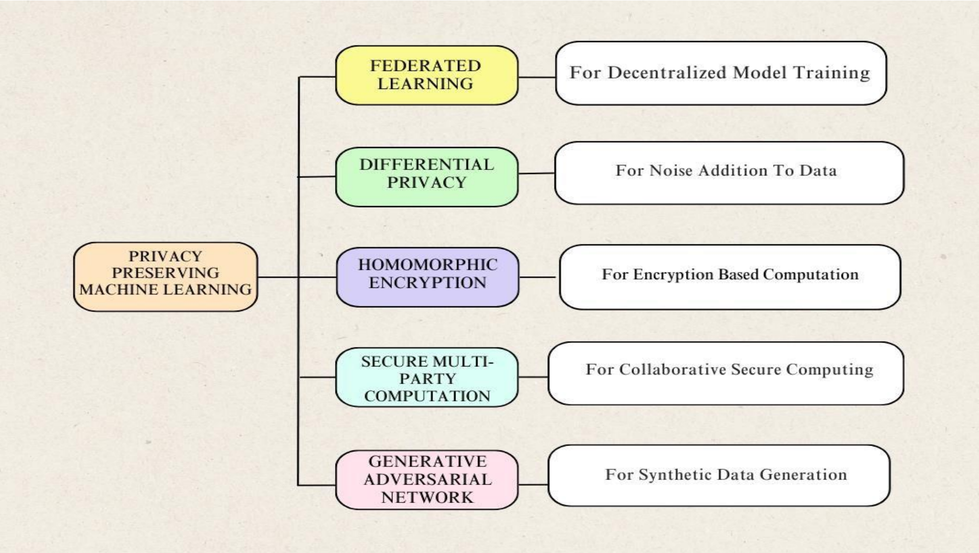

PRIVACY PRESERVING MACHINE LEARNING
FEDERATED LEARNING
For Decentralized Model Training
DIFFERENTIAL PRIVACY
For Noise Addition To Data
HOMOMORPHIC ENCRYPTION
For Encryption Based Computation
SECURE MULTI-PARTY COMPUTATION
For Collaborative Secure Computing
GENERATIVE ADVERSARIAL NETWORK
For Synthetic Data Generation

Figure 2.1. Privacy-Preserving Techniques with their use cases [6]

2.1 Differential Privacy

The most common recipe for "differentially-privatizing" is the training of different types of machine learning models, including linear or logistic regression, SVM (Support Vector Machines), and decision trees, is to utilize the Laplace mechanism. In this case, the training process is no longer simply looking for the most likely parameter setting given the data but rather draws a parameter setting from the conditional differential distribution of the parameters given the data. This can be achieved by adding appropriately scaled noise to the likelihood or loss functions used by the learning algorithms.

Differential privacy is a popular notion of privacy in machine learning and statistics that we mentioned earlier. Differential privacy requires, roughly, that a machine learning model yield near-equal prediction accuracy independent of the presence ("absorbing" an individual's data) or absence (without absorbing an individual's data) of any given individual's data. Differential privacy was designed to protect training data while releasing statistical information about the data that can be learned from the data, but while preventing re-identification of individuals in the data. It is able to achieve this by adding a certain amount of noise to the learning/parameter estimation process.

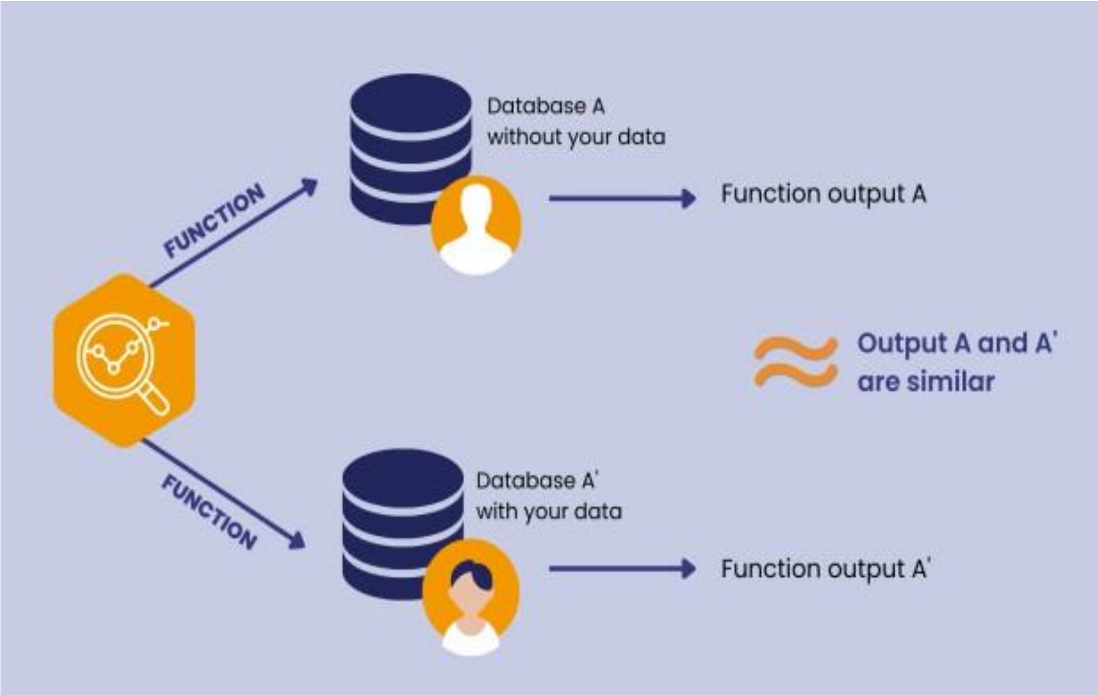


Figure 2.2. In a differentially private system, the output of a function does not depend on whether a record is present or absent [7]

2.2 Federated Learning

Several methods enable privacy-preserving machine learning and will be surveyed here, starting with Federated Learning (FL). Recent advances in FL are the automation of model architectures, hyper-parameters, weights used in ensembling, and adjusting, which all enable its application to a wide range of models. However, including hyper-parameters means that the local training error from each node is needed. Moreover, FL continues using explicit weight averaging per epoch, so it has additional complexity costs associated with ensemble model averaging over better untrusted (including adversarial) models. Note also that it involves a parameter broadcast and a model aggregate network operation together with two or more communication steps, an additional latency overhead that is irrelevant for reducing the direct exposure of sensitive data of local training used in learning remote models. These latencies are more significant in a decentralized telecom backbone context than in a device scenario, where the broadcast is between device and local infrastructure. Aside from the mentioned use-case of devices on device habitat, a federated learning type model would be particularly useful in a permuted leaf environment. Current implementations of FL, however, bypass the concealed topic of private learning of perturbation of non-IID training data at the nodes by employing 'trusted curation' based on consensus clustering algorithms, which prevents extending it to the broader topic of privacy preservation, including the learning stage that is the subject of this discussion.

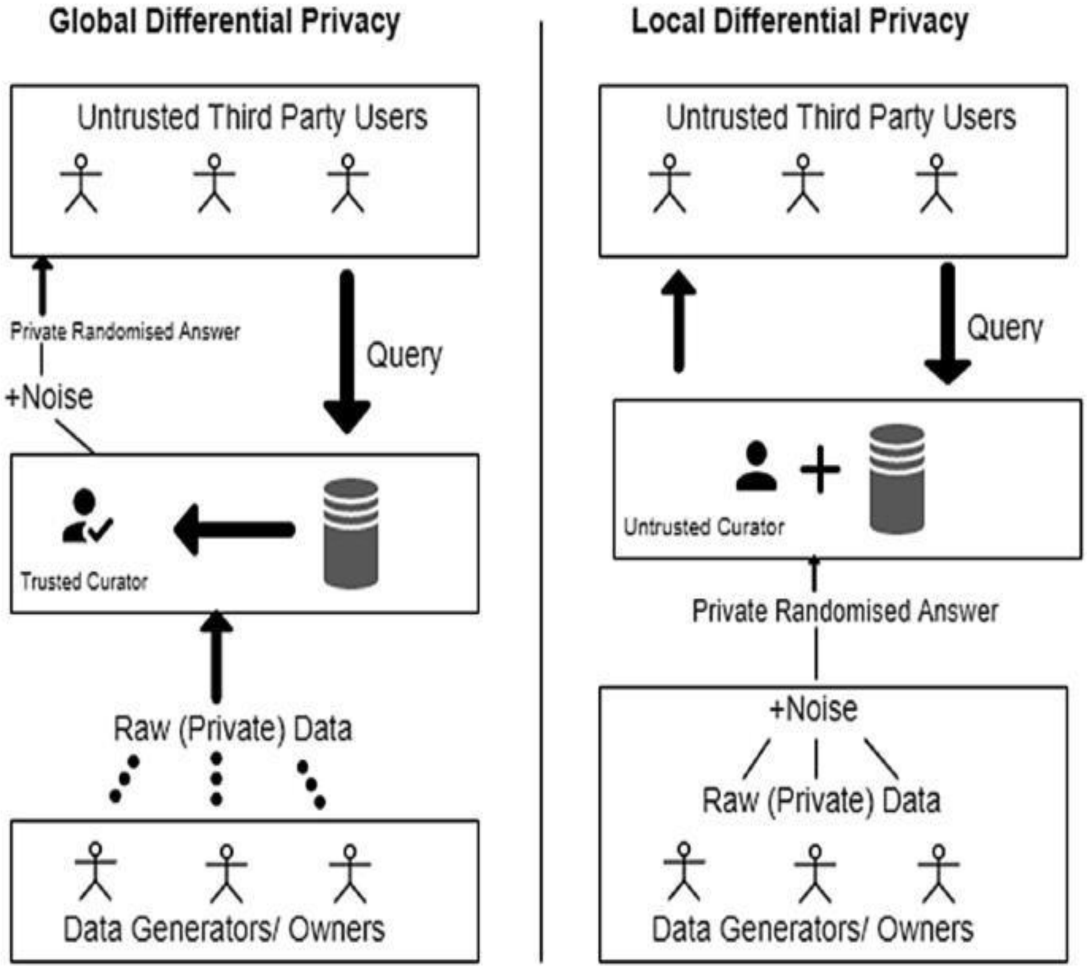

Global Differential Privacy
Untrusted Third Party Users
Private Randomised Answer
+Noise
Query
Trusted Curator
Raw (Private) Data
Data Generators/ Owners
Local Differential Privacy
Untrusted Third Party Users
Query
Untrusted Curator
Private Randomised Answer
+Noise
Raw (Private) Data
Data Generators/ Owners

Figure 2.3. Global Differential Privacy v/s Local Differential Privacy [8]

2.3 Homomorphic Encryption (HE)

A revolutionary cryptographic method called Homomorphic Encryption (HE) enables calculations to be done directly on the encrypted data without making use of decryption. This capability ensures that the confidentiality of sensitive medical data, such as patient records or electrocardiogram (ECG) signals, is preserved throughout the processing pipeline. There are several encryption techniques with great potential for the implementation of PPML. Homomorphic Encryption is one of the most well-known cryptographic methods that enable arithmetic operations over ciphertexts without the need for decryption or plain transformation of data. This encryption method is well suited for secure computation in cloud computing, and it is applicable for some specific and simple-to-complex solutions of secure data analysis.

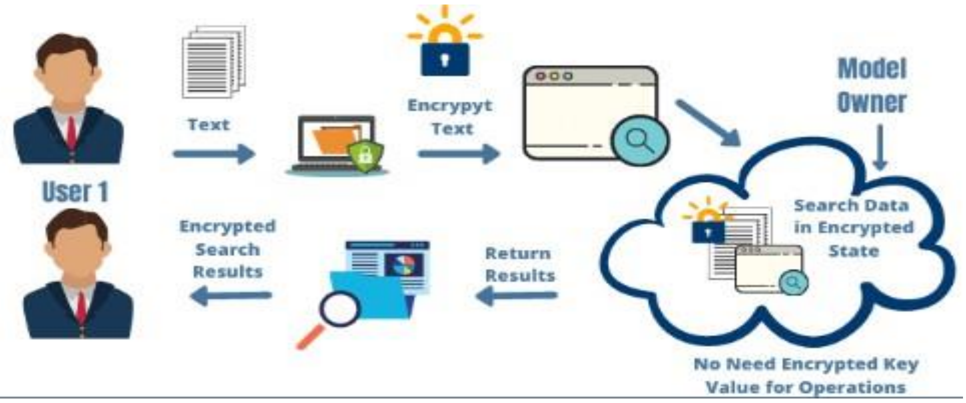

User 1
Text
Encrypyt Text
Model Owner
Search Data in Encrypted State
Encrypted Search Results
Return Results
No Need Encrypted Key Value for Operations

Figure 2.4. Block Diagram for Homomorphic Encryption [9]

In FL, HE is typically used to compute the weighted gradient for updating the global model with encrypted locally computed gradients on the client side. From the three types of Homomorphic Encryption schemes (partially and somewhat fully), RLWE (ring learning with errors) is known as a noise-rich type of encryption. While somewhat and partially homomorphic approaches have limitations on the numbers of operations or the magnitudes of the obtained outputs, these problems can be solved by using skillfully selected natural numbers, hence making the error grow exponentially in the encryption operations.

2.4 Secure Multi-Party Computation (SMPC)

There is another cryptographic method called Secure Multi-Party Computation (SMPC) that allows several parties to work together to calculate a function over their inputs while maintaining the privacy of those inputs. This method aggregates the inputs from all parties and helps in computing the function without decrypting/disclosing any information about their inputs other than the function's result. SMPC is typically categorized from semi-honest to malignant opponents, depending on the number of parties and the degree of security attempted. Yao's protocol, commonly known as garbled circuits, is the most widely used protocol for safe computing between semi-honest parties.

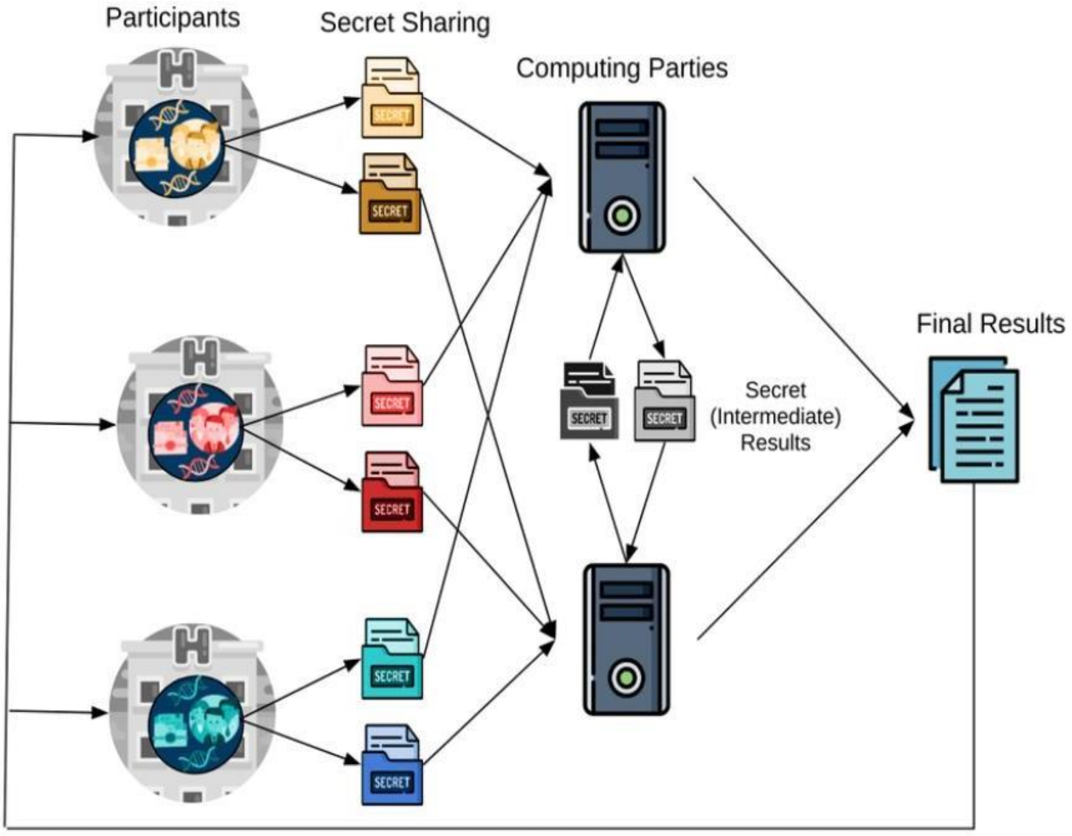


Figure 2.5. Block Diagram for SMPC [10]

2.5 Generative Adversarial Networks (GANs) for Synthetic Data

With the rise in technology, Generative Adversarial Networks (GANs) have been proved as a highly effective and most useful method to generate synthetic data that are closely resembled with the real datasets while privacy has been maintained. GANs are able to achieve this by making use of two interconnected and intertwined neural networks: a generator and a discriminator, which work together in the production of realistic synthetic data. In the domain of healthcare, GANs are making huge impact in the generation of high-quality synthetic medical data, such as imaging datasets, patient health records, or electrocardiogram (ECG) signals. The synthetic data being generated, can be taken into account for the deep model training without having the risk of sensitive patient data being exposed to any model.

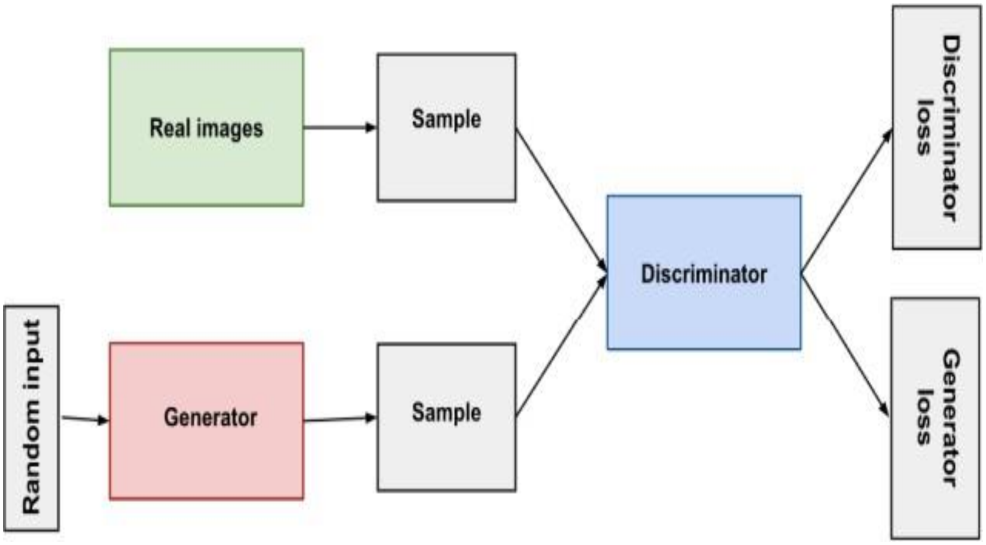


Figure 2.6. GAN System for Synthetic Data [11]

This deep learning method has various advantages and one of the major advantages of GANs in privacy-preserving is the ability to solve both privacy concerns and data scarcity at the same time. With the creation artificial/synthetic datasets that tries to replicate the statistical patterns of real-world data without directly copying individual cases, GANs has great significance in reducing the risk of privacy exposures. Furthermore, this quality of GANs makes them highly suitable for critical applications like analysis of ECG for arrhythmia detection, where large, varied and diverse datasets are crucial to develop most possibly accurate and robust machine learning models for its critical use. Hence, GANs makes balance in the need for privacy and the demand for reliable training data in healthcare research.

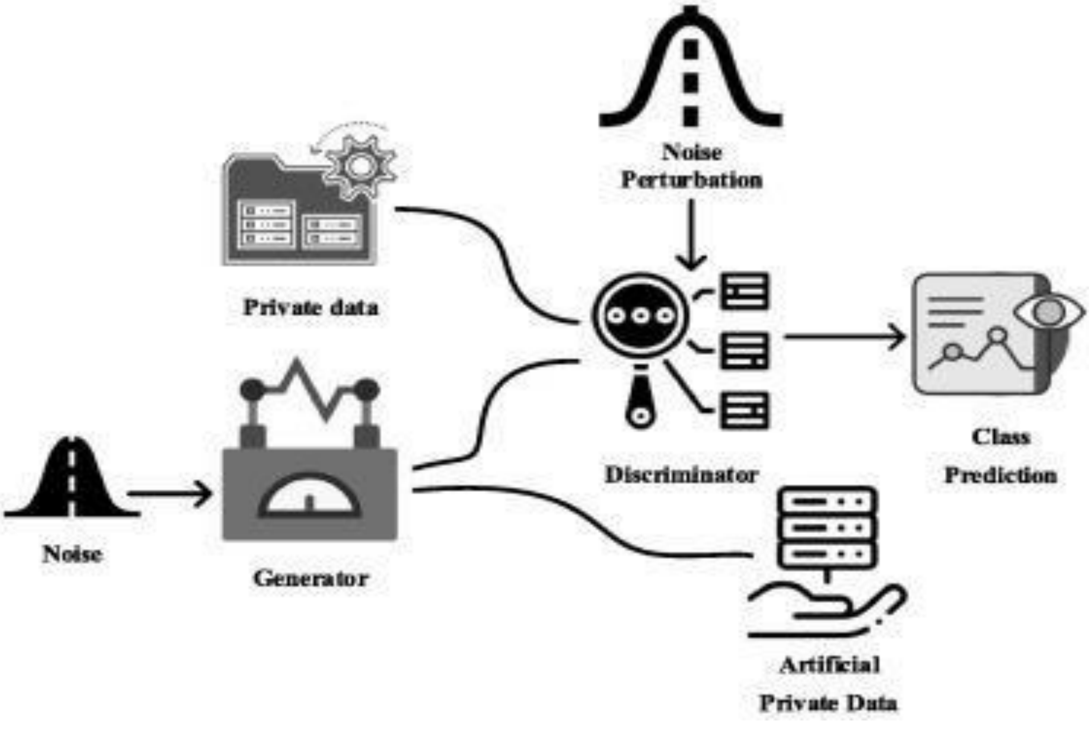


Figure 2.7. GAN Based Privacy Preserving Method [12]

2.6 Parameters in Privacy-preserving techniques

In the following table, there is a comparison of the privacy-preserving techniques in machine learning showing the wide variety of strategies, each having their own advantage and disadvantage.

Table 2.1. Comparison based on Accuracy, Privacy, Scalability and Latency

Technique	Accuracy Impact	Privacy Level	Scalability	Latency
Federated Learning	High	Moderate	High	Moderate
Differential Privacy	Moderate	High	High	Low
Homomorphic Encryption	Low	Very High	Low	High
Secure Multi-Party Computation	Moderate	Very High	Moderate	High
GANs for Synthetic Data	Moderate	High	High	Low

CHAPTER 3

LITERATURE REVIEW

3.1 Unsupervised Deep Learning:

Most used models for unsupervised deep learning methods in healthcare data analyses include autoencoders and restricted Boltzmann machines (RBMs).

In SPICE, after aggregating features are called patterns, patients' regular clusters using unsupervised learning methods such as K-mean, mixture model, ward clustering, etc. Pattern frequencies are then computed in the plaintext setting of a query. Such an approach then tries to make the query non-reusable as a separate approach that allows further queries after reallocating the used resource.

3.2 Unsupervised Machine Learning:

Unsupervised learning deals with problems where one is interested in understanding the structure of an unlabeled dataset, for instance in a way to identify subgroups or descriptive features in healthcare datasets such as EHRs or CNN images. Clustering methods, including k-means and classical methods for relay-based adversarial privacy, may be used to group visits of patients. In k-means, while the number of clusters must be specified, the L-Drawback based algorithms such as k-learning try to learn what is the most informative clusters count.
In the upcoming paragraphs, we will overview three general types of methods in machine learning, namely unsupervised learning, supervised learning, and deep learning for healthcare data analyses, respectively.

Similarly, using high-dimensional data in healthcare applications, Zhao et al. utilized matrix completion techniques combined with Gaussian process regression for multi-label prediction of comorbidities in oncology. The method was tested against clinical and histological data of over 400,000 patients suffering from 15 types of cancer. Another psychiatry-related work by Sun et al. assessed neurocognitive performance in depression using unsupervised machine learning techniques. Requirements are rising which propose alternate processing paradigms that protect patient privacy and do not need the user data at the processing end. To tackle these concerns, privacy-preserving machine learning techniques are being researched. The privacy-preserving framework confirms that the updated model does not precisely reveal contents present in the update.

Saria et al. analyzed Electronic Health Records (EHR) to predict the risk of adverse events and diseases using an unsupervised large margin learning method. In another study, using a cohort of 400 patients hospitalized with cancer, we developed a risk prediction model for septicemia using the method of generalized sequential pattern mining to extract temporal patterns from vital sign measurements. The authors made sure to employ methods which are interpretable, able to detect early warning signals, and require a minimum number of variables.

The proposed framework and its objective to enable privacy-preserving machine learning for healthcare data can broadly be categorized under two main research areas: machine learning and data mining in health informatics utilizing Electronic Health Records (EHRs), and applications of privacy-preserving techniques for real-world applications, including healthcare. We now elaborate on some research within these broad categories.

With the rise of the sensitive data in the world, it has created concerns over confidentiality of it. Even after a pandemic like COVID-19 opened the gateway for highly confidential medical data, which has increased the demand for privacy preservation of it. Hence, the federated learning algorithm has been widely used as a privacy-preserving machine learning technique in the domain of health and medical data. In this literature review, we have highlighted recent research that is going on genomic data, its analysis and the potential for predicting pandemics through deep learning and privacy preserving techniques.

3.3 Federated Learning in Healthcare:

- **McMahan et al. (2017)**: Federated learning is one of the most used techniques which allows the training of machine learning models in a segregated manner, where data is not shared across clients. This method has been highly advantageous in healthcare as it provides a way to develop predictive models that help in maintaining the privacy of patient data [13].
- **Hard et al. (2018)**: Federated learning is applied to clinical trials focused on privacy preservation with improvement in model performance [14].
- **Brisimi et al. (2018)**: The efficiency of federated learning has shown good and remarkable results in healthcare for the prediction of the outcomes while maintaining data security through differential privacy [15].

3.4 Genomic Data and Deep Learning for Pandemic Prediction & Federated Learning:

- **Suliman et al. (2020)**: Used SARS-CoV-2 genomic data to track mutations, highlighting the importance of spike protein mutations in transmissibility and infection severity [16].
- **Ying et al. (2017)**: Proposed differential privacy in federated learning to prevent leakage of sensitive data from model updates [17].

3.5 Performance Evaluation:

Following bar chart shows the performance evaluation of the techniques on the basis of various factors like Accuracy, Privacy, Scalability and Latency [18]. This comparative analysis highlights the strengths and weakness of different techniques of privacy-preserving.

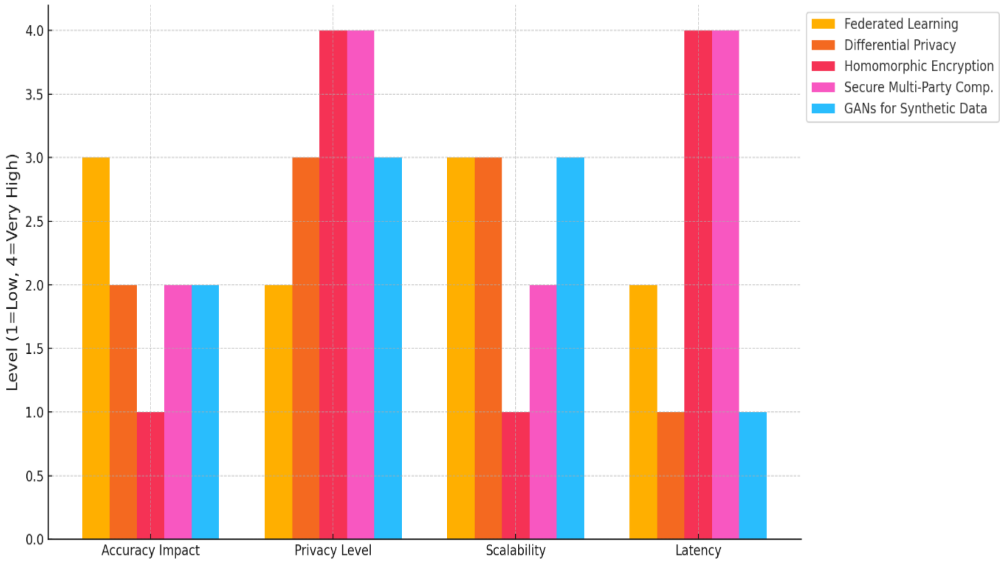


Figure 3.1. Comparison of Privacy-Preserving techniques on different parameters

In the case of federated learning, there is huge scalability, but it faces communication overhead issues. If there is no sharing of raw data, then this method proves to be excellent in the collaboration with model training. In differential privacy, we can add noise so that there is strong mathematical guarantee; and hence it necessitates a good balance between privacy and precision (when we are dealing with sensitive medical data like ECG signals [19]). Homomorphic encryption provides a high security approach by encrypting the calculations, but the real-time usage is very limited because of its high computational complexity. Secure Multi-Party Computation has the ability of distributing the computations among the parties, but it suffers from scalability issues in the case of huge datasets.

GANs (Generative Adversarial Networks) [20] addresses the problem of privacy issues and data scarcity and helps in efficiently producing synthetic data, but often it faces the overfitting problem and leakage problems. Each of the technique has a unique purpose and we have to choose a particular technique on the basis of use case, legal requirements and the limitations in computing. There's always a scope for the hybrid approaches that helps in improving the privacy and performance in the healthcare.

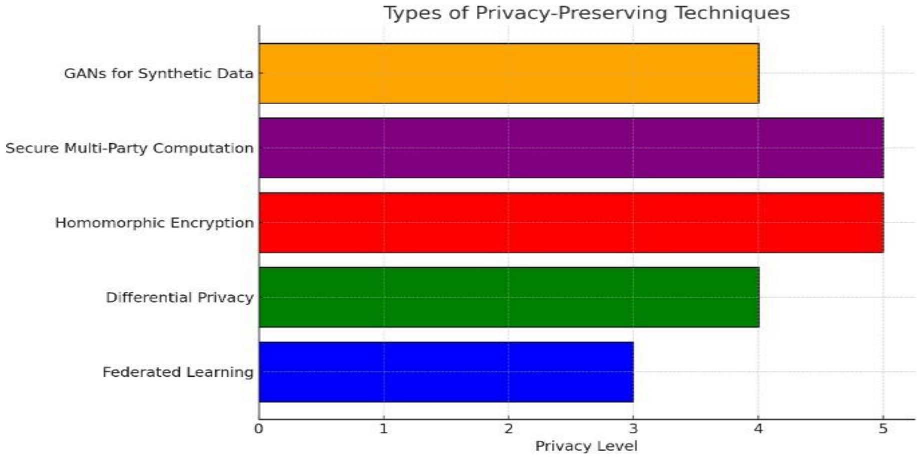


Figure 3.2. Privacy-Preserving Techniques on the scale of privacy level

Table 3.1. Related Work and Key Takeaways

Authors	**Paper Title**	**Scope of Paper**	**Key Takeaways**
Yunyun Cheng et al.	Machine Learning Techniques Applied to COVID-19 Prediction: A Systematic Review [21]	Systematic review of ML models for COVID-19 prediction	Hybrid models combining ML techniques improve prediction accuracy significantly.

Authors	**Paper Title**	**Scope of Paper**	**Key Takeaways**
Zohair Malki et al.	The COVID-19 pandemic: prediction study based on machine learning models [22]	Predicting COVID-19 spread using ML models	ML models can accurately predict COVID-19 spread; significant decline predicted.
Ramu, Agusthiyar et al.	A COVID-19 Prediction Based on Machine Learning Algorithms: A Literature Review [23]	Review of ML algorithms for predicting COVID-19 trajectory	ML models provide high accuracy in predicting COVID-19 cases.
Ashraf Ewis et al.	Machine Learning Models for COVID-19 Prediction and Privacy Preservation [24]	Combining ML models with privacy-preserving techniques for COVID-19 prediction	Effective integration of privacy-preserving techniques with ML models.
El-Sayed Atlam et al.	Predicting COVID-19 Spread Using Machine Learning Models [25]	Predicting the spread of COVID-19 using various ML models	ML models can forecast COVID-19 spread with high accuracy.
Mohamed M. Abdel-Daim et al.	Machine Learning Approaches for COVID-19 Prediction [26]	Application of ML approaches for predicting COVID-19 cases	ML approaches enhance prediction accuracy for COVID-19 cases.
Ibrahim Gad et al.	Privacy-Preserving Techniques in COVID-19 Prediction [27]	Review of privacy-preserving techniques in COVID-19 prediction	Privacy-preserving techniques are crucial for sensitive health data.
Guesh Dagnew et al.	Federated Learning for COVID-19 Prediction [28]	Federated learning models for predicting COVID-19 spread	Federated learning models maintain data privacy while predicting COVID-19 spread.

Authors	Paper Title	Scope of Paper	Key Takeaways
Osama A. Ghoneim et al.	Machine Learning and Privacy Preservation in COVID-19 Prediction [29]	Combining ML and privacy preservation for COVID-19 prediction	Effective combination of ML and privacy preservation techniques.

CHAPTER 4

PROPOSED ARCHITETURE

4.1 Nextstrain Database

Nextstrain is an open source and powerful platform which plays an important role in tracking the evolution of various pathogens. For the COVID-19 pandemic, SARS-CoV-2 virus was responsible; it is also included in Nextstrain. Following are the known and crucial features about the SARS-CoV-2 genomic data and its metadata:

- **Genomic Sequences**: Nextstrain analyzes and compiles genomic data from globally collected SARS-CoV-2 samples. To monitor the evolution of the virus continuously and its spread in various areas, all the sequences are updated regularly.
- **Metadata**: This platform also provides different methods for the visualization of phylogenetic trees, which helps in showing the genetic relationships between different virus strains along with genomic sequences. It also includes data such as the collection date and geographical location of each sample, which also tells about the temporal dependencies of the virus and its geographical spread. This metadata helps in the identification of the mutations and in tracking the emergence of new variants of the virus.
- **Nextclade**: This database also allows users to classify their sequences into specific clades of the virus and to compare them with the SARS-CoV-2 genome which was responsible for the pandemic in 2020, and also to identify most potential issues related to sequence quality of the genomic data.
- **Global and Regional Analysis**: This tool is powerful enough that it is updated daily based on the analyses, both regionally and globally. It also highlights the insights into the development of viruses with time and in various regions. It helps in interpreting the evolution of the virus within geographic boundaries and on more broad levels.

4.1.1. Phylogenetic Tree Rooting

A Phylogenetic tree [30] is basically helpful in representing the relationship of evolution between various biological entities in the graphical form. In simple words, it shows different species or viruses, how they are related, and who evolved from whom, over time.

The root in the phylogenetic tree is the starting point which shows the common ancestor of all the species or viruses. It represents the most ancient common ancestor from which all the other species or sequences in the tree have evolved. Finding the root helps scientists understand the direction of evolution and how different species have branched out over time.

In the context of pandemic prediction, rooting helps in tracking how the virus has evolved and spread from its original source. It helps in comparing genetic similarities and differences among species, helps in finding out evolutionary history.

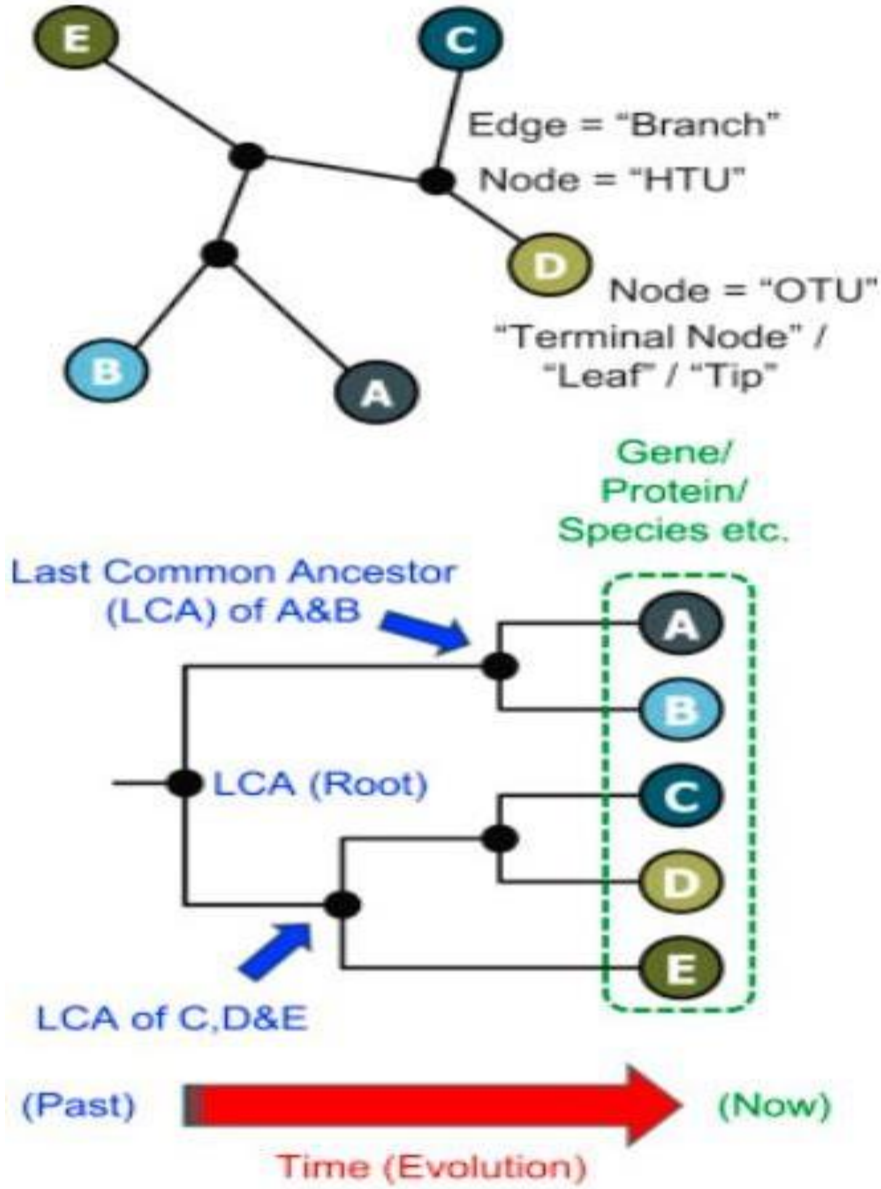


Figure 4.1. Phylogenetic Tree Rooting

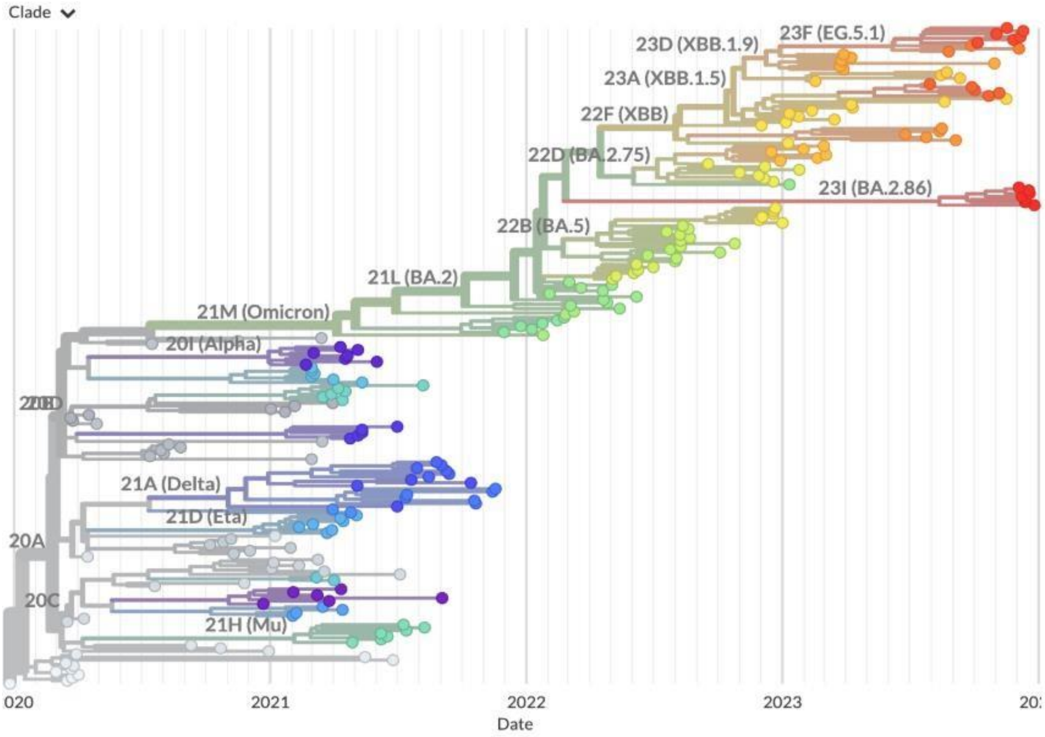


Figure 4.2. Used Auspice server to visualize phylogeny of around 200 sequences [31]

Hence, Nextstrain is beneficial for monitoring and recognizing the dynamics of SARS-CoV-2 in real time, which helps researchers and public health workers to stay informed about the mutation of the virus patterns and its spread across different areas.

4.1.2 Structure of SARS-CoV-2

SARS-CoV-2 is an RNA virus with a single-stranded, positive-sense RNA genome. It has genome size of ~29,900 bases. It also helps in encoding for structural, non-structural, and accessory proteins.

Table 4.1. Structural Proteins in SARS- CoV- 2

Protein	Function	Significance
Spike (S)	Binds to ACE2 receptor for cell entry	Target for vaccines, evolves rapidly
Envelope (E)	Helps in virus assembly and release	Structural stability
Membrane (M)	Maintains the shape of the virus	Most abundant
Nucleocapsid (N)	Binds and protects viral RNA	Important for diagnostics

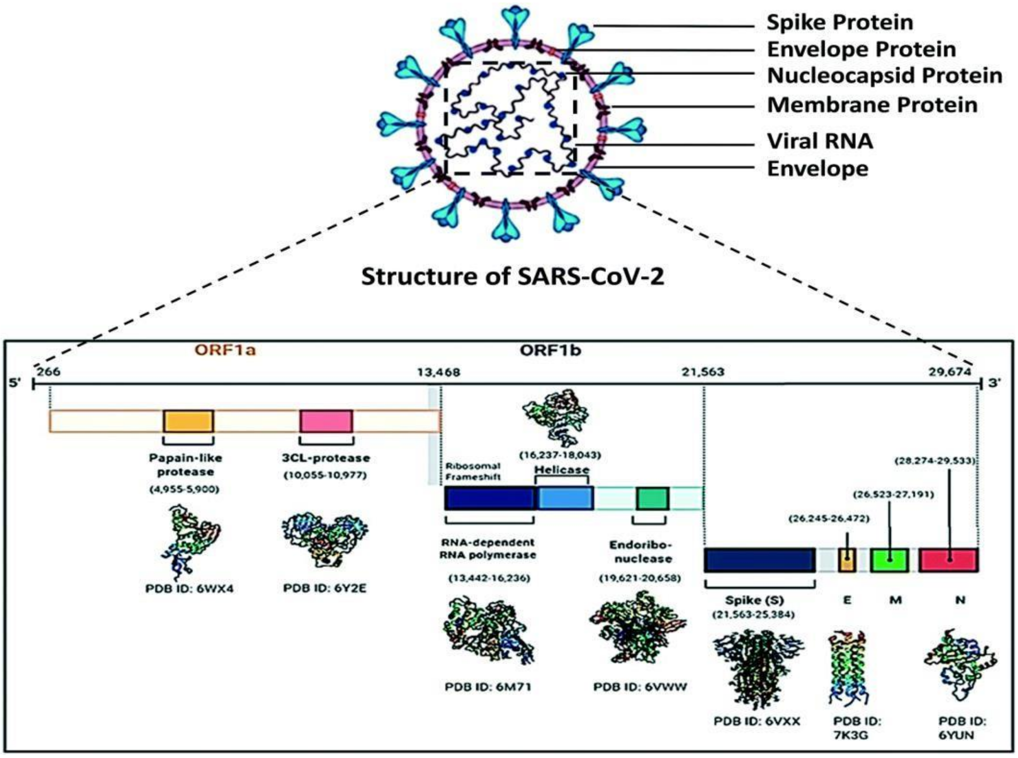


Figure 4.3. Genomic Structure of SARS-CoV-2 [32]

Nonstructural Proteins (nsps):

It is produced from ORF1a and ORF1b regions:

- Involved in replication, transcription, immune evasion
- Examples: RNA-dependent RNA polymerase (RdRp), proteases, helicases

5’ – ORF1a – ORF1b – S – E – M – N – 3’
↓
Non-structural Structural

This is a simplified layout of the nsps in this genome.

Table 4.2. Key Components of monitoring of viral evolution, spread, and mutation over time across different populations and geographic regions.

Component	Description
Genomic Sequencing	Viral RNA is sequenced from patient samples.
Lineage Identification	Identifies how closely related viruses are (e.g., Delta, Omicron).
Mutation Surveillance	Tracks changes in the virus's genetic code.

Component	Description
Phylogenetic Tree Construction	Visualizes how the virus evolves and spreads over time.
Spatiotemporal Analysis	Maps show how and when variants spread geographically.

4.2 Methodology

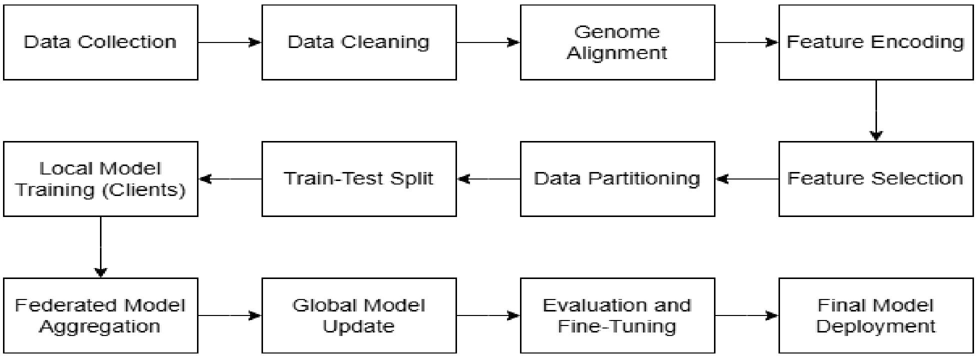


Fig.4.4. Detailed Steps of the Methodology of the Pandemic Prediction

4.2.1 Data Preparation and Preprocessing:

From the Nextstrain database, we have gathered the genomic sequences of the SARS-CoV-2, its metadata and all the other information related to it. For our study, the dataset we had consisted of the following:

- Genomic Sequences: Taken 10,000 SARS-CoV-2 samples to get viral RNA sequences.
- Metadata Attributes: Collection date of the sample, place where it occurred (geographical location), viral clade, variant labels (Alpha, Delta, Omicron) and the outcomes or results.
- For each genome sequence, we had approximately 29,000 base pairs (bp) for SARS-CoV-2. We used padding and truncation to make standard input sizes for CNN i.e. 30,000 base pairs.

The first and most basic step is to clean the dataset as it helps in ensuring the integrity of the data. Identification and removal of the insufficient data and the data which had some ambiguity was done to get the cleaned dataset. We have followed the three steps for preprocessing as discussed below:

1. *Genome Alignment.* There was a high need to get the standardized sequences and its preparation for the predictions and certain analysis, genome alignment needs to be done using open-source bioinformatics tools available. There are tools such as MAFFT (Multiple Alignment using Fast Fourier Transform) and MUSCLE (Multiple Sequence Comparison by Log-Expectation) which are usually used for genome alignment. Using such a technique helps in stabilizing the length of the genome and it becomes much easier to find the significant mutations in the data

provided.

In this research paper, we have used the MUSCLE tool to align all sequences, ensuring uniformity in the dataset and better preparatory analysis for subsequent data.

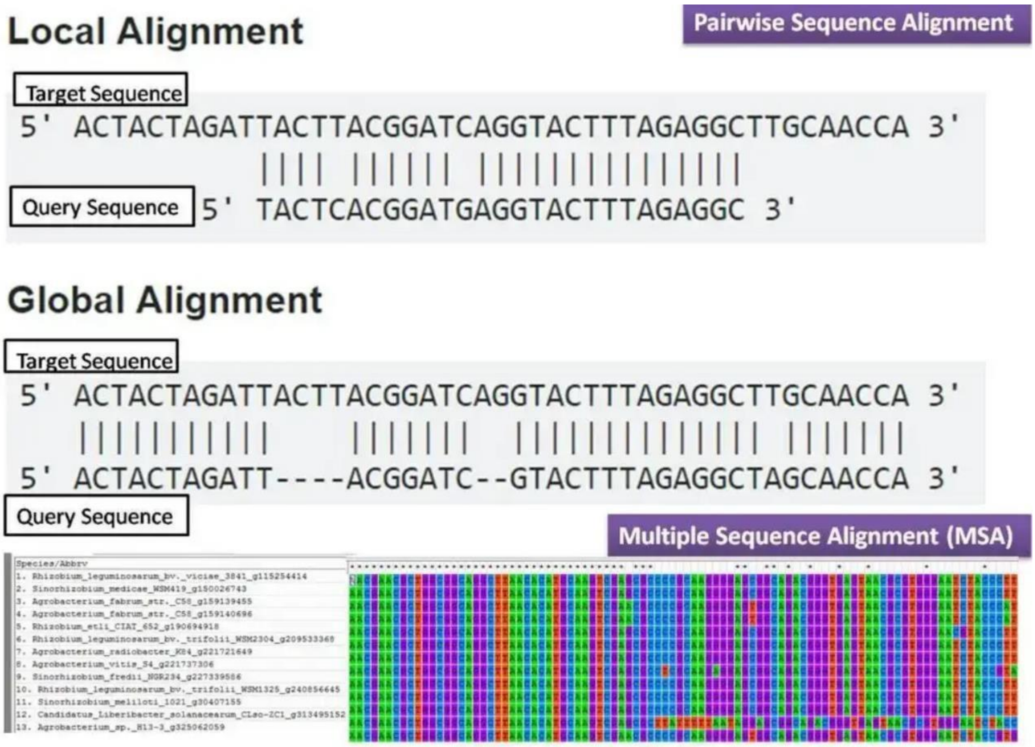

Local Alignment
Pairwise Sequence Alignment
Target Sequence
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
Query Sequence
5' TACTCACGGATGAGGTACTTTAGAGGC 3'
Global Alignment
Target Sequence
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
5' ACTACTAGATT----ACGGATC--GTACTTTAGAGGCTAGCAACCA 3'
Query Sequence
Multiple Sequence Alignment (MSA)
Species/Abbrv
1. Rhizobium_leguminosarum_bv._viciae_3841_g115254414
2. Sinorhizobium_medicae_WSM419_g150026743
3. Agrobacterium_fabrum_str._C58_g159139455
4. Agrobacterium_fabrum_str._C58_g159140696
5. Rhizobium_etli_CIAT_652_g190694918
6. Rhizobium_leguminosarum_bv._trifolii_WSM2304_g209533368
7. Agrobacterium_radiobacter_K84_g221721649
8. Agrobacterium_vitis_S4_g221737306
9. Sinorhizobium_fredii_NGR234_g227339586
10. Rhizobium_leguminosarum_bv._trifolii_WSM1325_g240856645
11. Sinorhizobium_meliloti_1021_g30407155
12. Candidatus_Liberibacter_solanacearum_CLso-ZC1_g313495152
13. Agrobacterium_sp._H13-3_g325062059

Figure 4.5. Types of Sequence Alignment [33]

2. *Feature Extraction*. We have now aligned all the sequences through the tool discussed above, now they are required to transform into a high-dimensional space with numerical features. We have methods like k-mer frequency representation [34] and one-hot encoding to capture various patterns in the dataset. While k-mer frequency is used to find the patterns in nucleotide in combination of length k, one-hot encoding is used when binary vector is there which is represented as A, T, G and C. This transformation of the unstructured datasets into structured datasets through these techniques has led to effective utilization of the ML models. These methods usually result in increasing the dimensionality, yet they are important in capturing the complexity that might be present in the genomic sequences in a machine-readable format.

   Here, we have used k-mer frequencies (k=3) to encode which has helped in capturing the trinucleotide patterns from each of the sequences. And then, one-hot encoding was applied that helped us to represent the position of each nucleotide, resulted in high-dimensional feature vectors and ensured dimensional consistency throughout the whole dataset taken.

   The k-mer counts for the sequences are -

   Counter ({'TAG': 7, 'GCT': 6, 'CTA': 6, 'AGC': 5, 'CGT': 3, 'GTA': 3, 'CGA': 2, 'GAT': 2, 'ATC': 2, 'TCG': 2, 'TAC': 2, 'ACG': 2, 'ATG': 1, 'TGC': 1, 'GCG': 1})

   The one hot encoded k-mers matrix is –

```
[[1. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
 [0. 1. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
 [0. 0. 1. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
 [0. 0. 0. 2. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
 [0. 0. 0. 0. 2. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
 [0. 0. 0. 0. 0. 2. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
 [0. 0. 0. 0. 0. 0. 2. 0. 0. 0. 0. 0. 0. 0. 0.]
 [0. 0. 0. 0. 0. 0. 0. 3. 0. 0. 0. 0. 0. 0. 0.]
 [0. 0. 0. 0. 0. 0. 0. 0. 3. 0. 0. 0. 0. 0. 0.]
 [0. 0. 0. 0. 0. 0. 0. 0. 0. 2. 0. 0. 0. 0. 0.]
 [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 2. 0. 0. 0. 0.]
 [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 7. 0. 0. 0.]
 [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 5. 0. 0.]
 [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 6. 0.]
 [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 6.]]
```

3. *Target Variable.* We have defined a target variable that helps in connecting the genomic features to predictable and quantifiable results. This could include the transmissibility rate of the virus or the severity of the infection (like mild, severe, or critical cases) if we want to predict the pandemic. Now, with the help of these target labels, supervised learning is made possible, enabling the model to use genomic inputs to forecast clinical outcomes or epidemic dynamics.

This procedure of data preprocessing bridges the gap between genomics and predictive modelling by guaranteeing a pipeline from raw sequences to actionable insights.

Strategies for Handling Missing Data and for noisy data –

Missing Nucleotides (N bases)	Imputed using k-mer nearest neighbors: Predict missing bases from similar sequences.
Gaps in Alignment	Gap-filling using consensus alignment (e.g., MAFFT for multiple sequence alignment).
Low-Quality Reads	Reads with >20% low-confidence bases are discarded or trimmed.
Artificial Mutations	Statistical mutation modeling is used to detect and correct outliers.

4.2.2 Data Partitioning for Federated Learning:

The dataset was partitioned to simulate five institutions (clients), each representing a unique healthcare or research organization with localized genomic data [35]. The details for each client were as follows:

- Client 1: Represented data from a specific geographic region with 1,400 sequences predominantly from early pandemic phases.

- Client 2: Included data from another region, emphasizing mutations observed in mid-pandemic waves.
- Client 3: Focused on sequences linked to severe outcomes, providing insights into high-risk mutations.
- Client 4: Contained diverse sequences collected across several regions to simulate a global dataset within a single institution.
- Client 5: Featured data highlighting regulatory region mutations impacting viral replication.

Each client independently trained the model using its data subsets (training, validation, and testing splits). This ensured that training occurred without cross-sharing raw genomic data. Since we have done the training independently for each client, this has allowed us to fine-tune our models on the localized data for each client, and also it has helped in preserving the data ownership hence it has helped in incorporating the FL framework.

- For each of the clients, we had 1,400 sequences and the metadata associated with it.
- We have distributed the dataset equally to help in simulating the real-world scenarios, where each of the clients manage data specific to their local regions/area.

While making the partitions, we ensured the following things:

- Distinct Data Allocation: Each client operated on a unique set of sequences, ensuring no duplication or overlap of data across clients.
- Independent Model Training: Clients conducted model training autonomously using only the data allocated to them. [36][37]

4.2.3 Training-Validation-Testing Splits:

We have divided the datasets for each of the clients as follows:

- **Training Set**: 70% of the sequences (1,400 sequences) for model training.
- **Validation Set**: 20% for hyperparameter tuning. [38]
- **Testing Set**: 10% to evaluate model performance locally before aggregation.

4.2.4 Feature Selection:

We have identified all the crucial genomic features through the evaluation of the biological importance and their statistical correlation with the target variables. Following are the features used for training-

a. Raw Genome Sequences – (A, T, G, C → binary vectors) using one-hot encoding and (k-mer embeddings) learned embedded representations.

b. Patterns in Mutation – Spike protein region in SARS-CoV-2 (denotes mutational hotspots).

c. Temporal Features – It helps in tracking the evolution over time (date of sequence collection).

d. Features based on geography – To understand the regional genomic variations that directly influences the severity in pandemic.

Steps for Feature Selection:

- Mutation Analysis: For the identification of the meaningful patterns from the genomic data, we have examined the mutation trends in the spike protein and regulatory regions.

- Statistical Validation: In order to get the bias free target outcomes, we have prioritized features showing high statistical relationships with the target outcomes.

- Dimensionality Reduction: Principal Component Analysis (PCA) was applied to select high-dimensional feature vectors, and hence optimizing performance of the model while preserving important information. [39]

4.2.5 Model Architecture and Federated Learning Implementation:

Model Architecture. In this work, the deep learning model [40] used has following layers to make the predictions of pandemic [41]:

1. **Input Layer**: After the preprocessing of the genomic feature vectors, these are passed into the input layer to encode the genomic data in a more efficient manner after we have applied k-mer frequency representation and one hot encoding.

2. **Convolutional Layers**: We have used these layers to extract the spatial patterns from encoded genomic data from the input layer. [42]

   Architecture Details:

   - **Conv1D Layer 1:**
     - **Filters:** 64
     - **Kernel Size:** 5 (captures short-range dependencies like small mutations)
     - **Activation:** ReLU
     - **Stride:** 1

   - **Conv1D Layer 2:**
     - **Filters:** 128
     - **Kernel Size:** 10 (captures larger motifs and mutation regions)
     - **Activation:** ReLU

   - **MaxPooling Layer:**
     - Pool Size: 2 (reduces feature size, prevents overfitting)

3. **Recurrent Layers (LSTM) [43]**: This layer is helpful in capturing the sequential dependencies in mutation patterns given by the stack of convolutional layers. With the help of this layers, it tracks the mutations took in the past and how it influences the strains in future.

Architecture Details:

- **LSTM Layer 1:**
  - **Units:** 128
  - **Return Sequences:** True (keeps sequential information for stacking)
  - **Dropout:** 0.3 (prevents overfitting)

- **LSTM Layer 2:**
  - **Units:** 64
  - **Return Sequences:** False (final feature extraction)

4. **Dense Layers**: This layer helps in mapping of the extracted features from the LSTM to high-level representations for further predictions. We have used this layer as it helps in combining the spatial features from CNN and sequential features from LSTM to make the final predictions.

   Architecture Details:

   - **Dense Layer 1:**
     - Units: 128
     - Activation: ReLU
   - **Dropout:** 0.3
   - **Dense Layer 2:**
     - Units: 64
     - Activation: ReLU

5. **Output Layer**: This is the final layer of the model which helps in making the predictions for infection severity and transmissibility from the genomic data.

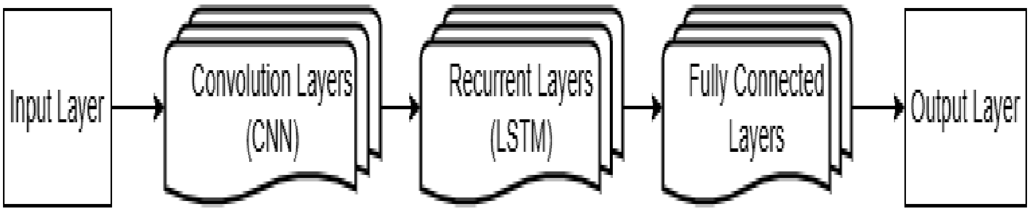
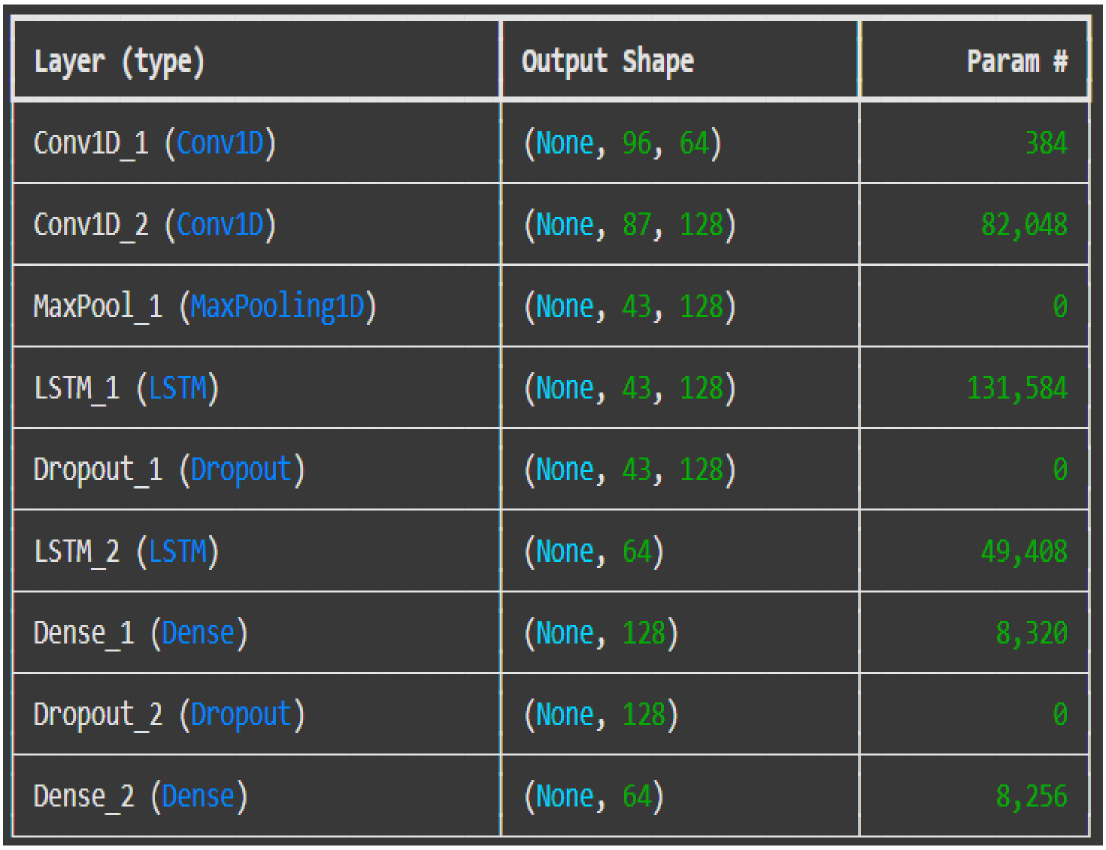


Fig.4.6. Layered Diagram of the Model Architecture

Table 4.3. Table for Model Architecture with output shapes



Layer (type)	Output Shape	Param #
Conv1D_1 (Conv1D)	(None, 96, 64)	384
Conv1D_2 (Conv1D)	(None, 87, 128)	82,048
MaxPool_1 (MaxPooling1D)	(None, 43, 128)	0
LSTM_1 (LSTM)	(None, 43, 128)	131,584
Dropout_1 (Dropout)	(None, 43, 128)	0
LSTM_2 (LSTM)	(None, 64)	49,408
Dense_1 (Dense)	(None, 128)	8,320
Dropout_2 (Dropout)	(None, 128)	0
Dense_2 (Dense)	(None, 64)	8,256

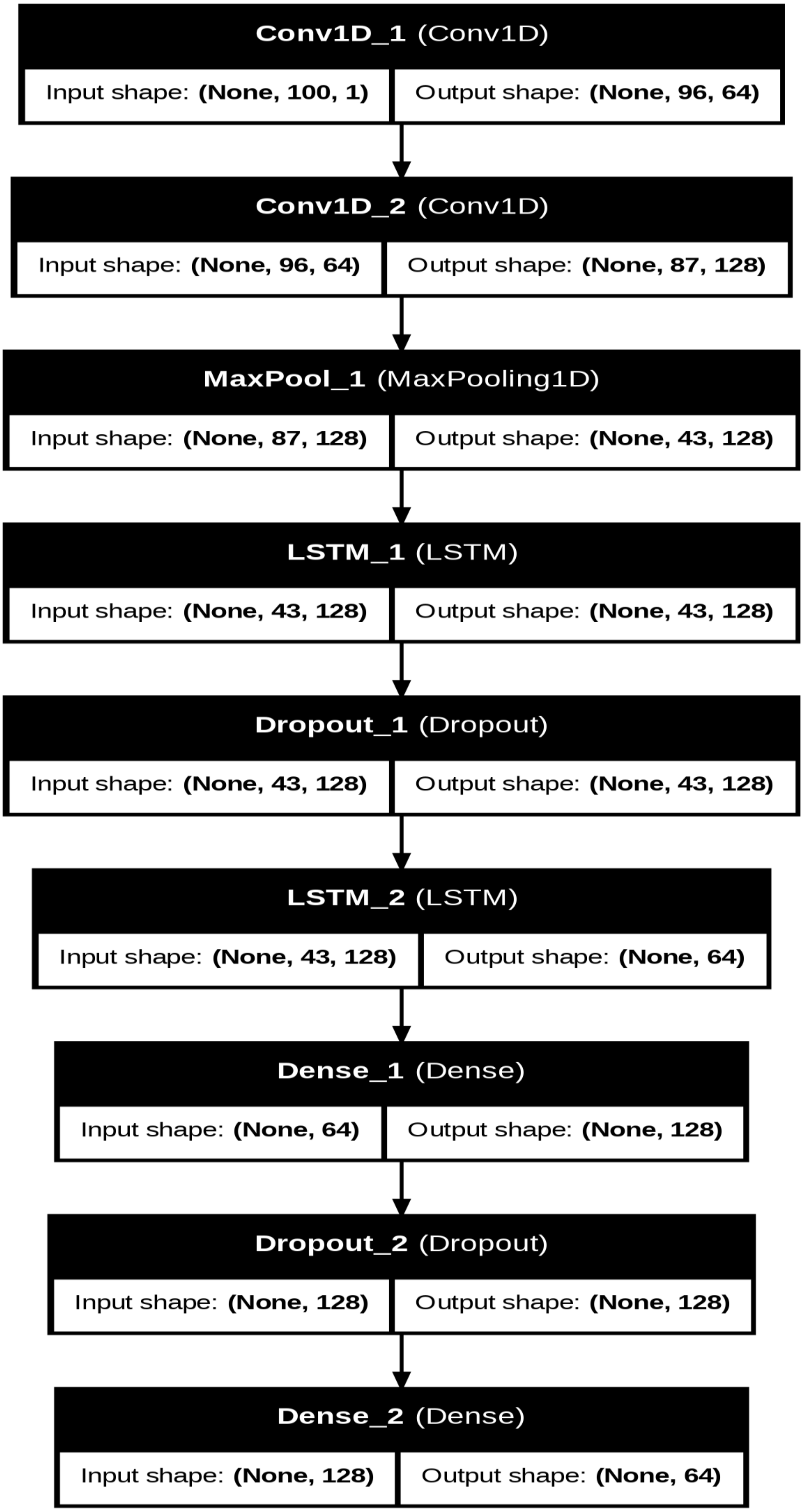

Conv1D_1 (Conv1D)
Input shape: (None, 100, 1)
Output shape: (None, 96, 64)
Conv1D_2 (Conv1D)
Input shape: (None, 96, 64)
Output shape: (None, 87, 128)
MaxPool_1 (MaxPooling1D)
Input shape: (None, 87, 128)
Output shape: (None, 43, 128)
LSTM_1 (LSTM)
Input shape: (None, 43, 128)
Output shape: (None, 43, 128)
Dropout_1 (Dropout)
Input shape: (None, 43, 128)
Output shape: (None, 43, 128)
LSTM_2 (LSTM)
Input shape: (None, 43, 128)
Output shape: (None, 64)
Dense_1 (Dense)
Input shape: (None, 64)
Output shape: (None, 128)
Dropout_2 (Dropout)
Input shape: (None, 128)
Output shape: (None, 128)
Dense_2 (Dense)
Input shape: (None, 128)
Output shape: (None, 64)

Figure 4.7. Model Architecture with input and output shapes

4.2.6 Federated Learning Workflow:

FL [44] was implemented to ensure privacy and enable decentralized collaboration:

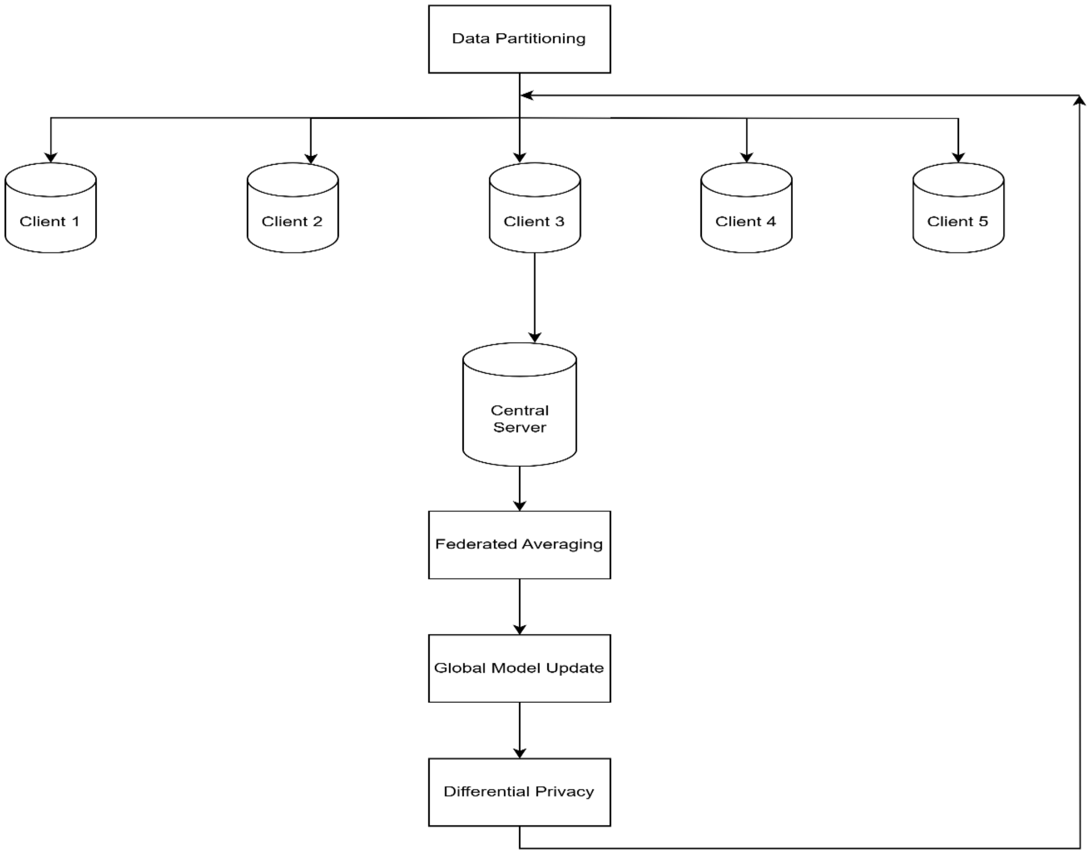


Fig.4.8. Diagrammatic insight of the Federated Learning Model

1. **Local Training**: Each client trained its model independently using the training and validation splits.

2. **Model Aggregation**: The central server aggregated client updates using Federated Averaging (FedAvg), combining the weights from each client proportionally. [45]

3. **Privacy Mechanisms**: Differential privacy was employed during aggregation, adding noise to prevent the reconstruction of sensitive data from updates.

CHAPTER 5

EXPERIMENTAL EVALUATION

5.1 Training Configuration:

Parameters for the FL framework were:

- **Federated Rounds**: 15
- **Local Epochs**: 5
- **Batch Size**: 32
- **Learning Rate**: 0.001

5.2 Validation Process:

For the validation, we have taken 20% of data – 2000 sequences. It consisted of the sequences from various virus strains across different regions. Each of the federated client validates its local model on separate validation set. We have used early stopping as the training stops if the validation loss does not improve for 5 consecutive rounds.

5.3 Testing Process:

For the testing, we have taken 10% of data – 1000 sequences. These sequences consist of unseen mutations to test the adaptability. So, each federated client tests the global model on its held-out test set. And then, the central server aggregates all the individual test results for the overall evaluation.

5.4 Evaluation Metrics:

Performance was assessed using:

1. **Accuracy**: Measured as the proportion of correct predictions on the test set.
2. **Privacy Loss**: Evaluated using differential privacy bounds.
3. **Communication Overhead**: Quantified as the total data transmitted during federated training.

5.5 Results and Observation

Performance Metrics**:**

Higher AUC-ROC (0.96) means better pandemic strain classification.
Lower MAE (0.11) indicates more precise mutation trajectory forecasting.

5.6 Key Findings:

- Spike protein mutations were the most predictive feature.
- FL demonstrated comparable accuracy to centralized models while significantly enhancing privacy.
- Differential privacy slightly reduced accuracy but ensured robust data protection.

Table 5.1. Comparison of Results and Observations with the implemented model v/s traditional methods for Pandemic Prediction

Metric	CNN-LSTM-FL-DP (Ours)	Traditional ML (Random Forest, SVMs)	CNN Only	LSTM Only
Accuracy	92%	78.1%	85.3%	87.2%
F1- Score	91.7	76.5	83.8	86
AUC-ROC	0.96	0.81	0.89	0.91
Precision	93.2%	75.9%	86%	88.1%
Recall	90.4%	77.2%	82.5%	85.9%
MAE (Mutation Forecasting)	0.11	0.23	0.17	0.14

Table 5.2. Comparison of Model Performance with Privacy Risks (Our Privacy Mechanism -Basic Differential Privacy)

Privacy Mechanism	**Average Model Accuracy (%)**	**Privacy Risk Level**	**Communication Cost (MB)**	**Computational Overhead**
No Privacy	87%	High	50	Low
Basic Differential Privacy	92%	Moderate	120	Moderate
Strong Differential Privacy	90%	Low	250	High
Homomorphic Encryption	91%	Very Low	500	Very High

The process of pandemic prediction discussed above helps in protecting the genomic data while ensuring the effectiveness of the data. Since, Federated Learning uses decentralized training with differential privacy helps in ensuring compliance with all the regulatory frameworks (HIPAA, GDPR). It also prevents the single-client domination in the training as it prevents malicious clients from affecting the model. So, it outperforms purely centralized models or standalone federated learning models as it balances the local training, privacy and generalization.

In terms of scalability, we can train this model using region-specific clients (Asia, Australia, Europe or America etc.) to handle geographic variations. We can also use model pruning and quantization to optimize the deployment process on low resource devices. Also, the data never leaves the local institutions or clients hence making it GDPR and HIPAA compliant. We can also design it to process different virus strains like Influenza, Ebola etc.

For the deployment, we can use cloud deployment like Google Cloud, Amazon Web Service (AWS) or local High-Performance Computing (HPC) clusters to train on large scale using federated learning. It can also be incorporated to run on hospital or labs with Tensorflow Lite or PyTorch Mobile to make predictions in real-time. The integration of this proposed model with the Nextstrain database and the Public Health Systems would result in highly impactful and would help in this social cause as it helps in tracking the pandemic in real time.

CHAPTER 6

CONCLUSION, LIMITATIONS AND FUTURE SCOPE

6.1 Conclusion

Privacy-preserving techniques are essential in the advancement of safe AI applications in healthcare and preserving private medical information, like genomic data, patient personal details or ECG signals etc[7]. This study emphasizes the distinctive benefits of several important methods discussed, which are Differential Privacy, Homomorphic Encryption, Secure Multi-Party Computation, Federated Learning, and Generative Adversarial Networks (GANs). Federated Learning and Differential Privacy balance privacy and performance, whereas Homomorphic Encryption and Secure Multi-Party Computation place security above computational efficiency. Although privacy preservation and its management are necessary to limit any data breaches, GANs are good at handling both privacy and data scarcity. These methods should be used on the basis of the particular use case which takes into account privacy related concerns, its computing requirements and regulatory compliance. Using strong hybrid frameworks could help in providing individual limit requirements, more safe and scalable solutions and hence effective AI solutions for various healthcare problems like arrhythmia detection.

Hybrid Privacy Frameworks: Creation of integrated frameworks that overcome the limitations of approaches by mixing federated learning, differential privacy, and homomorphic encryption to maximize scalability, data security, and its computational performance.

Enhanced Real-Time Processing: Creating more computationally efficient methods to support real-time applications and models without risking system responsiveness or its data privacy, such as continuous ECG monitoring for arrhythmia detection or genomic data or DNA data.

Synthetic Data Advancements: Enhancing GANs for the generation of more diversified, good-quality synthetic datasets that preserve privacy, makes more reliable model training, and solve data lacking issues.

Privacy-Aware Healthcare Systems: Creation of more scalable, in accordance with legal compliant healthcare applications that fuses privacy-preserving technology to meet standards like GDPR and HIPAA, ensuring data security across distributed networks.

Broader Applicability in Healthcare: To make sure of the safe AI-driven systems and its advancements in the medical industry, we can definitely extend the use of these techniques to a range of medical industry and healthcare, such as genomic data processing, medical imaging, and remote patient monitoring.

The privacy-preserving federated learning (FL) approach for pandemic prediction utilizing genomic data has been shown to be effective in this study. By using FL, sensitive genomic data may be kept secret while model training can be done collaboratively across institutions. In conclusion, federated learning has shown promising results in creating more scalable and privacy-preserving pandemic prediction models. While there are certain limitations in communication and accuracy trade-offs, there is huge potential in more secure and cooperative and collaborative research which makes it highly suitable for real-world applications in health care and genomic medicine and its prediction.

6.2 Challenges and Limitations

This study has shown that federated learning essentially balances prediction accuracy and privacy, making it more suitable for synergistic genomic research in pandemic prediction scenarios. Following are some challenges –

- High-dimensional data can lead to communication overhead.
- Computational costs are introduced by differential privacy mechanisms.

Table 6.1. Challenges in the real-world scenario with their potential solution

Challenge	Solutions
Integration of genomic data from diverse sources	Implement a data harmonization protocol to ensure consistency in sequencing formats, metadata structures, and labeling standards. Utilize ontology-based mapping for dataset alignment and cross-institutional compatibility.
Scalability of federated learning	Introduce hierarchical federated learning (HFL), grouping institutions with similar genomic data for localized model aggregation before global updates. Implement federated transfer learning to fine-tune pre-trained models for institutions with different data distributions.
Data heterogeneity across institutions	Apply personalized federated learning techniques like meta-learning or client-specific model fine-tuning to adapt to institutional variations. Use weighted averaging in federated aggregation (e.g. FedProx) to handle disparities in data distribution.

For the approach discussed in this work, we had the following current limitations-

- It needs multiple institution or clients to participate but have only limited adequate resources.
- If it has higher Differential Privacy noise levels, then the model accuracy gets reduced slightly.
- If there are unseen variants or new strain emerges then the model may not have an exact class label.

6.3 Future Scope

In the future scope, we will definitely work upon these challenges faced and also, we have recommended scope of improvements which includes:

- Incorporating advanced encryption techniques such as homomorphic encryption.
- Expanding the approach to real-world clinical genomic datasets.
- Optimizing communication protocols to minimize FL overhead.

CHAPTER 7

SOCIAL IMPACT

The social impact of Pandemic prediction using Nextstrain database is very significant and multifaceted. With the help of database systems like Nextstrain which updates the data locally and globally on a regular basis, it would be highly helpful in making real-time predictions.

Here we have discussed various social impacts of the model discussed in this thesis.

1. Prediction of the mutations or variant spread patterns early helps in enabling faster public health responses. It reduces the scalability and severity of pandemics through predictive modelling.
2. With federated learning, institutions or clients (hospitals, labs, governments) can contribute to global models without sharing raw data. It overcomes data-sharing barriers due to privacy, regulations (e.g., GDPR), or politics.
3. Privacy-preserving techniques (like federated learning, differential privacy, SMPC or GANs for synthetic data) align with ethical standards and helps in building public trust. People are more willing to allow use of their genomic and clinical data as only data required for training is used and all the personal or private information is not used for model training.
4. This model can be accommodated for real time analysis and its prediction. Real-time granular prediction models help allocate resources to regions before they become hotspots. It then helps in reducing health disparities, resulting in enhancement of pandemic preparedness globally.

By turning worldwide (globally and locally) healthcare surveillance into a proactive, judicious, and ethical responsibility framework, a privacy-preserved pandemic prediction system can help societies or areas better equip and prepare themselves to tackle biological dangers without compromising individual rights.

REFERENCES

[1] Green, M., & Stigall, S. (2017). Compliance challenges for healthcare providers under GDPR and HIPAA. Health Policy and Technology, 6(2), 141–149.

[2] McKnight, R., & Franko, O. (2016). HIPAA compliance for healthcare workers: Understanding the rules. *Journal of Medical Systems, 40*(6), 135.

[3] Nazish Khalid, Adnan Qayyum, Muhammad Bilal, Ala Al-Fuqaha, Junaid Qadir, Privacy-preserving artificial intelligence in healthcare: Techniques and applications, Computers in Biology and Medicine, Volume 158, 2023, 106848, ISSN 0010-4825, https://doi.org/10.1016/j.compbiomed.2023.106848.

[4] Soumia Zohra El Mestari, Gabriele Lenzini, Huseyin Demirci, Preserving data privacy in machine learning systems, Computers & Security, Volume 137, 2024, 103605, ISSN 0167-4048, https://doi.org/10.1016/j.cose.2023.103605.

[5] Zhao W, Luo X, Qiu T. Smart Healthcare. *Applied Sciences*. 2017; 7(11):1176. https://doi.org/10.3390/app7111176

[6] Y. Sun, G. G. Yen and Z. Yi, "Evolving Unsupervised Deep Neural Networks for Learning Meaningful Representations," in IEEE Transactions on Evolutionary Computation, vol. 23, no. 1, pp. 89-103, Feb. 2019, doi: 10.1109/TEVC.2018.2808689.

[7] Padula WV, Armstrong DG, Pronovost PJ, et alPredicting pressure injury risk in hospitalised patients using machine learning with electronic health records: a US multilevel cohort studyBMJ Open 2024;14:e082540. doi: 10.1136/bmjopen-2023-082540.

[8] Preserving Privacy in Data Analysis: An analysis of Differential Privacy Techniques: Exploring the benefits and limitations of Differential Privacy Mechanisms, Koopman, J. H. (Author). Jun 2023.

[9] Uday, J., Ghosh, M. (2022). Safeguarding GeoLocation for Social Media with Local Differential Privacy and L-Diversity. In: Rao, U.P., Patel, S.J., Raj, P., Visconti, A. (eds) Security, Privacy and Data Analytics. Lecture Notes in Electrical Engineering, vol 848. Springer, Singapore. https://doi.org/10.1007/978-981-16-9089-1_2.

[10] Benzekki, Kamal & El Fergougui, Abdeslam & El Belrhiti El Alaoui, Abdelbaki. (2016). A Secure Cloud Computing Architecture Using Homomorphic Encryption. International Journal of Advanced Computer Science and Applications. 7. 10.14569/IJACSA.2016.070241.

[11] Torkzadehmahani, Reihaneh & Nasirigerdeh, Reza & Blumenthal, David & Kacprowski, Tim & List, Markus & Matschinske, Julian & Späth, Julian & Wenke, Nina & Bihari, Béla & Frisch, Tobias & Hartebrodt, Anne & Hauschild, Anne-Christin & Heider, Dominik & Holzinger, Andreas & Hötzendorfer, Walter & Kastelitz, Markus & Mayer, Rudolf & Nogales, Cristian & Pustozerova, Anastasia & Baumbach, Jan. (2020). Privacy-preserving Artificial Intelligence Techniques in Biomedicine. 10.48550/arXiv.2007.11621.

[12] Xikun Jiang, Chaoyue Niu, Chenhao Ying, Fan Wu, Yuan Luo, Pricing GAN-based data generators under Rényi differential privacy, Information Sciences, Volume 602, 2022, Pages 57-74, ISSN 0020-0255, https://doi.org/10.1016/j.ins.2022.04.030.

[13] McMahan, B., Moore, E., Ramage, D., Hampson, S. &amp; Arcas, B.A.y.. (2017). “Communication-Efficient Learning of Deep Networks from Decentralized Data”,

Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, in Proceedings of Machine Learning Research.

[14] Hard, A., Rao, K., Mathews, R., Ramaswamy, S., Beaufays, F., Augenstein, S., ... & Ramage, D. (2018). Federated learning for mobile keyboard prediction. arXiv preprint arXiv:1811.03604.

[15] Brisimi TS, Chen R, Mela T, Olshevsky A, Paschalidis IC, Shi W. Federated learning of predictive models from federated Electronic Health Records. Int J Med Inform. 2018 Apr;112:59-67. doi: 10.1016/j.ijmedinf.2018.01.007. Epub 2018 Jan 12. PMID: 29500022; PMCID: PMC5836813.

[16] Suliman Khan, Ashaq Ali, Hongwei Shi, Rabeea Siddique, Shabana, Ghulam Nabi, Junjie Hu, Tiejun Wang, Men Dong, Wajid Zaman, Guang Han, COVID-19: Clinical aspects and therapeutics responses, Saudi Pharmaceutical Journal, Volume 28, Issue 8, 2020, Pages 1004-1008, ISSN 1319-0164, https://doi.org/10.1016/j.jsps.2020.06.022.

[17] Ying Lin, Ling-Yan Bao, Ze-Minghui Li, Shu-Zheng Si, Chao-Hsien Chu, Differential privacy protection over deep learning: An investigation of its impacted factors, Computers & Security, Volume 99, 2020, 102061, ISSN 0167-4048, https://doi.org/10.1016/j.cose.2020.102061.

[18] Bezerra, Daniel & Filho, Assis & Silva, Iago & Silvério Melo Dantas, Marrone & Barbosa, Gibson & Souza, Ricardo & Kelner, Judith & Sadok, Djamel. (2022). A Machine Learning-Based Optimization for End-to-End Latency in Tsn Networks. SSRN Electronic Journal. 10.2139/ssrn.4117311.

[19] Aziz, S., Ahmed, S. & Alouini, MS. ECG-based machine-learning algorithms for heartbeat classification. Sci Rep 11, 18738 (2021). https://doi.org/10.1038/s41598-021-97118-5.

[20] E. Debie, N. Moustafa and M. T. Whitty, "A Privacy-Preserving Generative Adversarial Network Method for Securing EEG Brain Signals," 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 2020, pp. 1-8, doi: 10.1109/IJCNN48605.2020.9206683.

[21] Cheng, Y.; Cheng, R.; Xu, T.; Tan, X.; Bai, Y. Machine Learning Techniques Applied to COVID-19 Prediction: A Systematic Literature Review. Bioengineering 2025, 12, 514. https://doi.org/10.3390/bioengineering12050514.

[22] Malki, Z., Atlam, ES., Ewis, A. et al. The COVID-19 pandemic: prediction study based on machine learning models. Environ Sci Pollut Res 28, 40496–40506 (2021). https://doi.org/10.1007/s11356-021-13824-7.

[23] Ramu, Agusthiyar & Saranya, S. (2022). A Covid-19 Prediction based on Machine Learning Algorithms – A Literature Review. Ymer Digital. 21. 692-698. 10.37896/YMER21.06/70.

[24] Amna Faisal, NZ Jhanjhi, Humaira Ashraf, et al. A Comprehensive Review of Machine Learning Models: Principles, Applications, and Optimal Model Selection. *TechRxiv*. March 24, 2025.

[25] Zohair Malki, El-Sayed Atlam, Aboul Ella Hassanien, Guesh Dagnew, Mostafa A. Elhosseini, Ibrahim Gad,

[26] Association between weather data and COVID-19 pandemic predicting mortality rate: Machine learning approaches, Chaos, Solitons & Fractals, Volume 138,2020, 110137, ISSN 0960-0779, https://doi.org/10.1016/j.chaos.2020.110137.

[27] M. Nabil, A. Sherif, M. Mahmoud, W. Alsmary and M. Alsabaan, "Privacy-Preserving

Non-Participatory Surveillance System for COVID-19-Like Pandemics," in *IEEE Access*, vol. 9, pp. 79911-79926, 2021, doi: 10.1109/ACCESS.2021.3082910.

[28] Malki, Z., Atlam, ES., Ewis, A. *et al.* ARIMA models for predicting the end of COVID-19 pandemic and the risk of second rebound. *Neural Comput & Applic* **33**, 2929–2948 (2021). https://doi.org/10.1007/s00521-020-05434-0.

[29] M. A. Rahman, M. S. Hossain, M. S. Islam, N. A. Alrajeh and G. Muhammad, "Secure and Provenance Enhanced Internet of Health Things Framework: A Blockchain Managed Federated Learning Approach," in *IEEE Access*, vol. 8, pp. 205071-205087, 2020, doi: 10.1109/ACCESS.2020.3037474.

[30] Richard J. Edwards, Phylogenetic Tree Rooting, Editor(s): Shoba Ranganathan, Michael Gribskov, Kenta Nakai, Christian Schönbach, Encyclopedia of Bioinformatics and Computational Biology, Academic Press, 2019, Pages 727-735, ISBN 9780128114322, https://doi.org/10.1016/B978-0-12-809633-8.20258-6.

[31] Hadfield et al., Nextstrain: real-time tracking of pathogen evolution, Bioinformatics (2018).

[32] Ghosh A, Kar PK, Gautam A, Gupta R, Singh R, Chakravarti R, et al. An insight into SARS-CoV-2 structures, pathogenesis, target hunting for drug development and vaccine initiatives. *RSC Med Chem* (2022) 13:647–75. doi: 10.1039/D2MD00009A

[33] Naznin, F., Sarker, R., & Essam, D. (2012). Progressive Alignment Method Using Genetic Algorithm for Multiple Sequence Alignment. IEEE Transactions on Evolutionary Computation, 16(5), 615–631. doi:10.1109/tevc.2011.2162849

[34] Kurtz, S., Narechania, A., Stein, J.C. *et al.* A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. *BMC Genomics* **9**, 517 (2008). https://doi.org/10.1186/1471-2164-9-517.

[35] Runhua Xu, Nathalie Baracaldo, Yi Zhou, Ali Anwar, James Joshi, and Heiko Ludwig. 2021. FedV: Privacy-Preserving Federated Learning over Vertically Partitioned Data. In Proceedings of the 14th ACM Workshop on Artificial Intelligence and Security (AISec '21). Association for Computing Machinery, New York, NY, USA, 181–192. https://doi.org/10.1145/3474369.3486872.

[36] J. Zhang, S. Guo, J. Guo, D. Zeng, J. Zhou and A. Y. Zomaya, "Towards Data-Independent Knowledge Transfer in Model-Heterogeneous Federated Learning," in *IEEE Transactions on Computers*, vol. 72, no. 10, pp. 2888-2901, Oct. 2023, doi: 10.1109/TC.2023.3272801.

[37] H. Wang, Z. Kaplan, D. Niu and B. Li, "Optimizing Federated Learning on Non-IID Data with Reinforcement Learning," *IEEE INFOCOM 2020 - IEEE Conference on Computer Communications*, Toronto, ON, Canada, 2020, pp. 1698-1707, doi: 10.1109/INFOCOM41043.2020.9155494.

[38] Lee, S.; Kim, J.; Kang, H.; Kang, D.-Y.; Park, J. Genetic Algorithm Based Deep Learning Neural Network Structure and Hyperparameter Optimization. *Appl. Sci.* **2021**, *11*, 744. https://doi.org/10.3390/app11020744.

[39] P. Cristea, V. Mladenov, G. Tsenov, R. Tuduce and S. Petrakieva, "Application of neural networks, PCA and feature extraction for prediction of nucleotide sequences by using genomic signals," *2008 9th Symposium on Neural Network Applications in Electrical Engineering*, Belgrade, Serbia, 2008, pp. 83-88, doi: 10.1109/NEUREL.2008.4685575.

[40] Ahmad Waleed Salehi, Preety Baglat, Gaurav Gupta, Review on machine and deep learning models for the detection and prediction of Coronavirus, Materials Today:

Proceedings, Volume 33, Part 7, 2020, Pages 3896-3901, ISSN 2214-7853, https://doi.org/10.1016/j.matpr.2020.06.245.

[41] Ajagbe, S.A., Adigun, M.O. Deep learning techniques for detection and prediction of pandemic diseases: a systematic literature review. *Multimed Tools Appl* **83**, 5893–5927 (2024). https://doi.org/10.1007/s11042-023-15805-z.

[42] Kugunavar, S., Prabhakar, C.J. Convolutional neural networks for the diagnosis and prognosis of the coronavirus disease pandemic. *Vis. Comput. Ind. Biomed. Art* **4**, 12 (2021). https://doi.org/10.1186/s42492-021-00078-w.

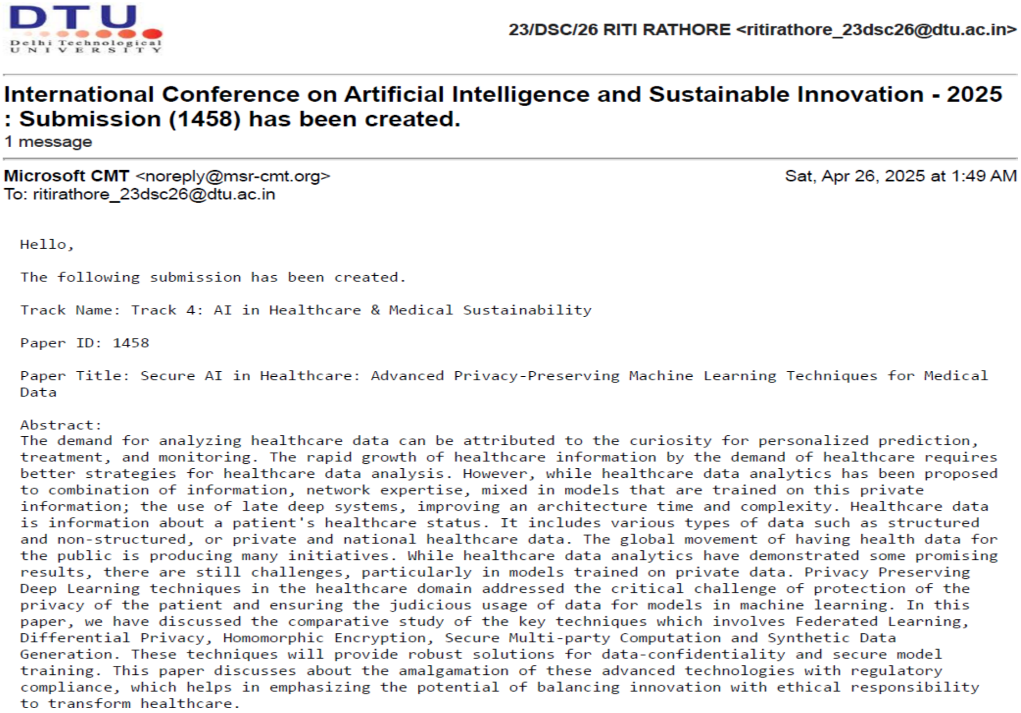
[43] Shastri, S., Singh, K., Kumar, S. *et al.* Deep-LSTM ensemble framework to forecast Covid-19: an insight to the global pandemic. *Int. j. inf. tecnol.* **13**, 1291–1301 (2021). https://doi.org/10.1007/s41870-020-00571-0.

[44] A. Qayyum, K. Ahmad, M. A. Ahsan, A. Al-Fuqaha and J. Qadir, "Collaborative Federated Learning for Healthcare: Multi-Modal COVID-19 Diagnosis at the Edge," in *IEEE Open Journal of the Computer Society*, vol. 3, pp. 172-184, 2022, doi: 10.1109/OJCS.2022.3206407.

[45] K. M. Elshabrawy, M. M. Alfares and M. A. . -M. Salem, "Ensemble Federated Learning for Non-II D COVID-19 Detection," *2022 5th International Conference on Computing and Informatics (ICCI)*, New Cairo, Cairo, Egypt, 2022, pp. 057-063, doi: 10.1109/ICCI54321.2022.9756090.

LIST OF PUBLICATION(S)

1. Abhilasha Sharma, Riti Rathore, "*Secure AI in Healthcare: Advanced Privacy-Preserving Machine Learning Techniques for Medical Data*". The paper has been **Accepted** at the **International Conference on Artificial Intelligence and Sustainable Innovation 2025 (ICAISI-2025)**, May 2025. Indexed by **Scopus.** Paper Id: 1458.



23/DSC/26 RITI RATHORE <ritirathore_23dsc26@dtu.ac.in>

International Conference on Artificial Intelligence and Sustainable Innovation - 2025 : Submission (1458) has been created.

1 message

Microsoft CMT <noreply@msr-cmt.org> Sat, Apr 26, 2025 at 1:49 AM

To: ritirathore_23dsc26@dtu.ac.in

Hello,

The following submission has been created.

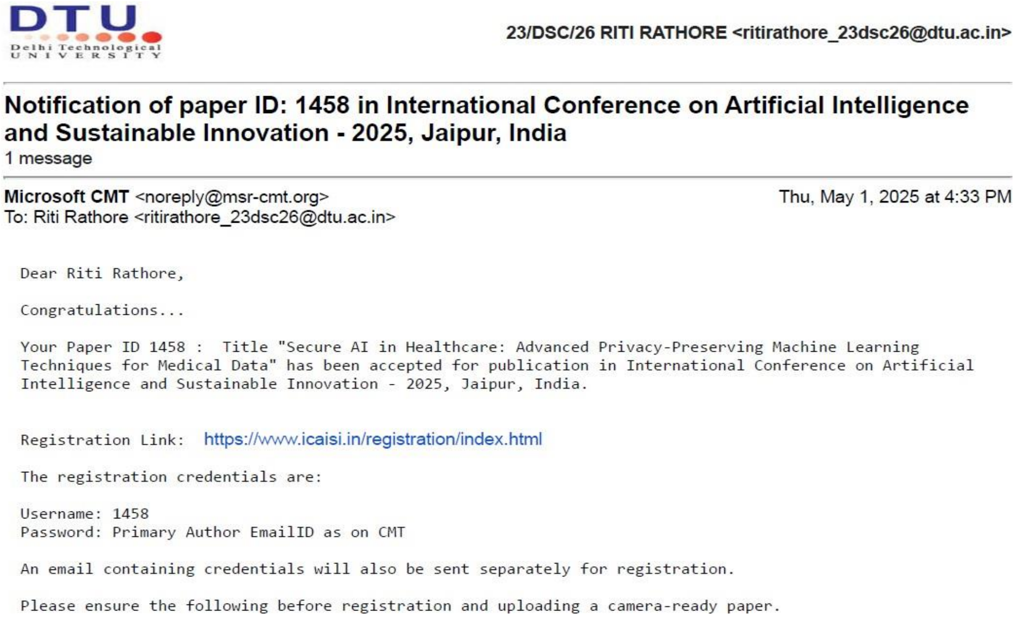
Track Name: Track 4: AI in Healthcare & Medical Sustainability

Paper ID: 1458

Paper Title: Secure AI in Healthcare: Advanced Privacy-Preserving Machine Learning Techniques for Medical Data

Abstract:
The demand for analyzing healthcare data can be attributed to the curiosity for personalized prediction, treatment, and monitoring. The rapid growth of healthcare information by the demand of healthcare requires better strategies for healthcare data analysis. However, while healthcare data analytics has been proposed to combination of information, network expertise, mixed in models that are trained on this private information; the use of late deep systems, improving an architecture time and complexity. Healthcare data is information about a patient's healthcare status. It includes various types of data such as structured and non-structured, or private and national healthcare data. The global movement of having health data for the public is producing many initiatives. While healthcare data analytics have demonstrated some promising results, there are still challenges, particularly in models trained on private data. Privacy Preserving Deep Learning techniques in the healthcare domain addressed the critical challenge of protection of the privacy of the patient and ensuring the judicious usage of data for models in machine learning. In this paper, we have discussed the comparative study of the key techniques which involves Federated Learning, Differential Privacy, Homomorphic Encryption, Secure Multi-party Computation and Synthetic Data Generation. These techniques will provide robust solutions for data-confidentiality and secure model training. This paper discusses about the amalgamation of these advanced technologies with regulatory compliance, which helps in emphasizing the potential of balancing innovation with ethical responsibility to transform healthcare.

Fig. 1. Submission Proof of Paper 1



23/DSC/26 RITI RATHORE <ritirathore_23dsc26@dtu.ac.in>

Notification of paper ID: 1458 in International Conference on Artificial Intelligence and Sustainable Innovation - 2025, Jaipur, India

1 message

Microsoft CMT <noreply@msr-cmt.org> Thu, May 1, 2025 at 4:33 PM

To: Riti Rathore <ritirathore_23dsc26@dtu.ac.in>

Dear Riti Rathore,

Congratulations...

Your Paper ID 1458 : Title "Secure AI in Healthcare: Advanced Privacy-Preserving Machine Learning Techniques for Medical Data" has been accepted for publication in International Conference on Artificial Intelligence and Sustainable Innovation - 2025, Jaipur, India.

Registration Link: https://www.icaisi.in/registration/index.html

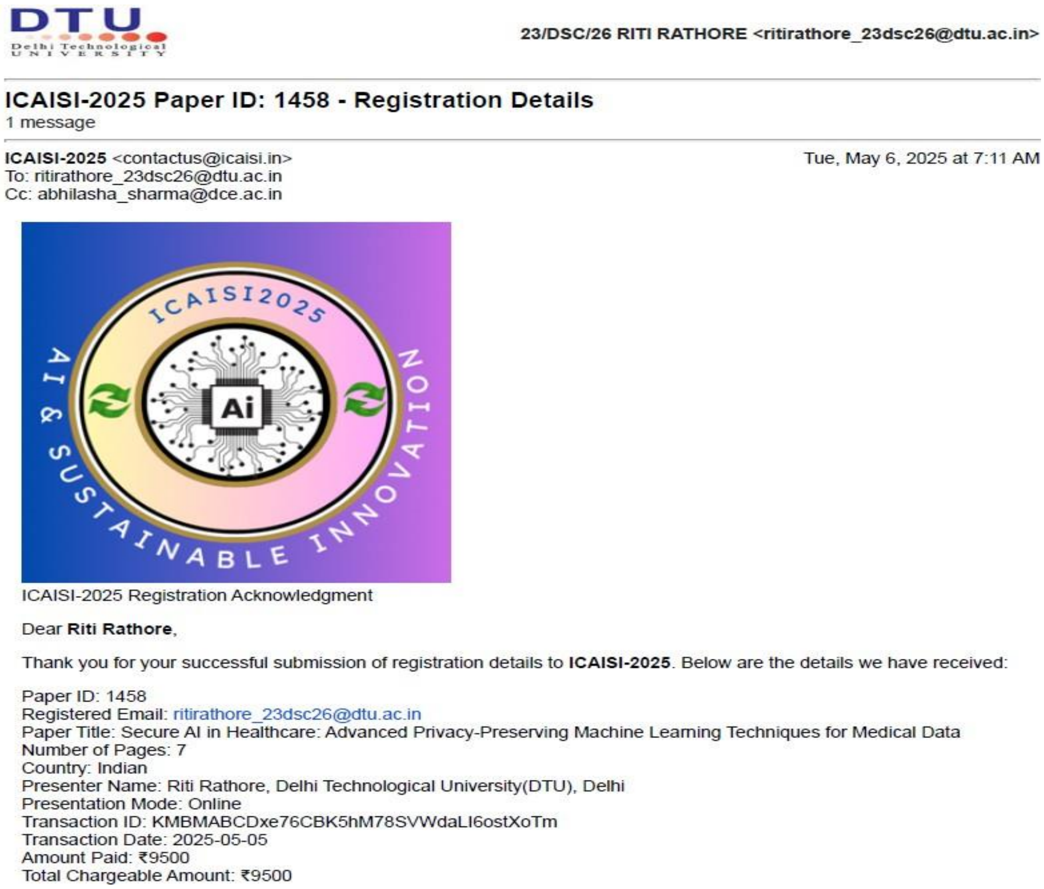
The registration credentials are:

Username: 1458
Password: Primary Author EmailID as on CMT

An email containing credentials will also be sent separately for registration.

Please ensure the following before registration and uploading a camera-ready paper.

Fig. 2. Acceptance Proof of Paper 1



23/DSC/26 RITI RATHORE <ritirathore_23dsc26@dtu.ac.in>

ICAISI-2025 Paper ID: 1458 - Registration Details
1 message

ICAISI-2025 <contactus@icaisi.in> Tue, May 6, 2025 at 7:11 AM
To: ritirathore_23dsc26@dtu.ac.in
Cc: abhilasha_sharma@dce.ac.in

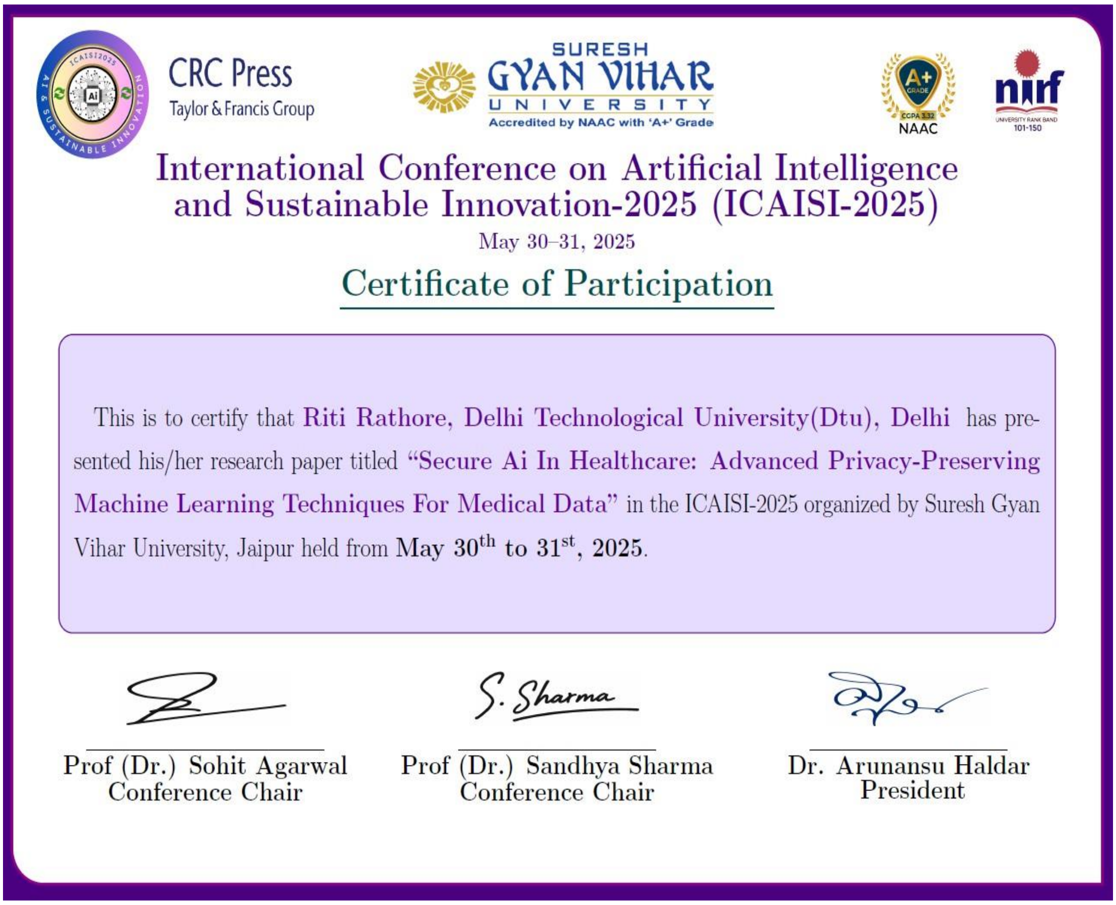
ICAISI-2025 Registration Acknowledgment

Dear **Riti Rathore**,

Thank you for your successful submission of registration details to **ICAISI-2025**. Below are the details we have received:

Paper ID: 1458
Registered Email: ritirathore_23dsc26@dtu.ac.in
Paper Title: Secure AI in Healthcare: Advanced Privacy-Preserving Machine Learning Techniques for Medical Data
Number of Pages: 7
Country: Indian
Presenter Name: Riti Rathore, Delhi Technological University(DTU), Delhi
Presentation Mode: Online
Transaction ID: KMBMABCDxe76CBK5hM78SVWdaLI6ostXoTm
Transaction Date: 2025-05-05
Amount Paid: ₹9500
Total Chargeable Amount: ₹9500

Fig. 3. Payment Proof of Paper 1



ICAISI2025
AI & SUSTAINABLE INNOVATION
CRC Press
Taylor & Francis Group
SURESH GYAN VIHAR UNIVERSITY
Accredited by NAAC with 'A+' Grade
A+
NAAC
nirf
UNIVERSITY RANK BAND 101-150

International Conference on Artificial Intelligence and Sustainable Innovation-2025 (ICAISI-2025)
May 30–31, 2025

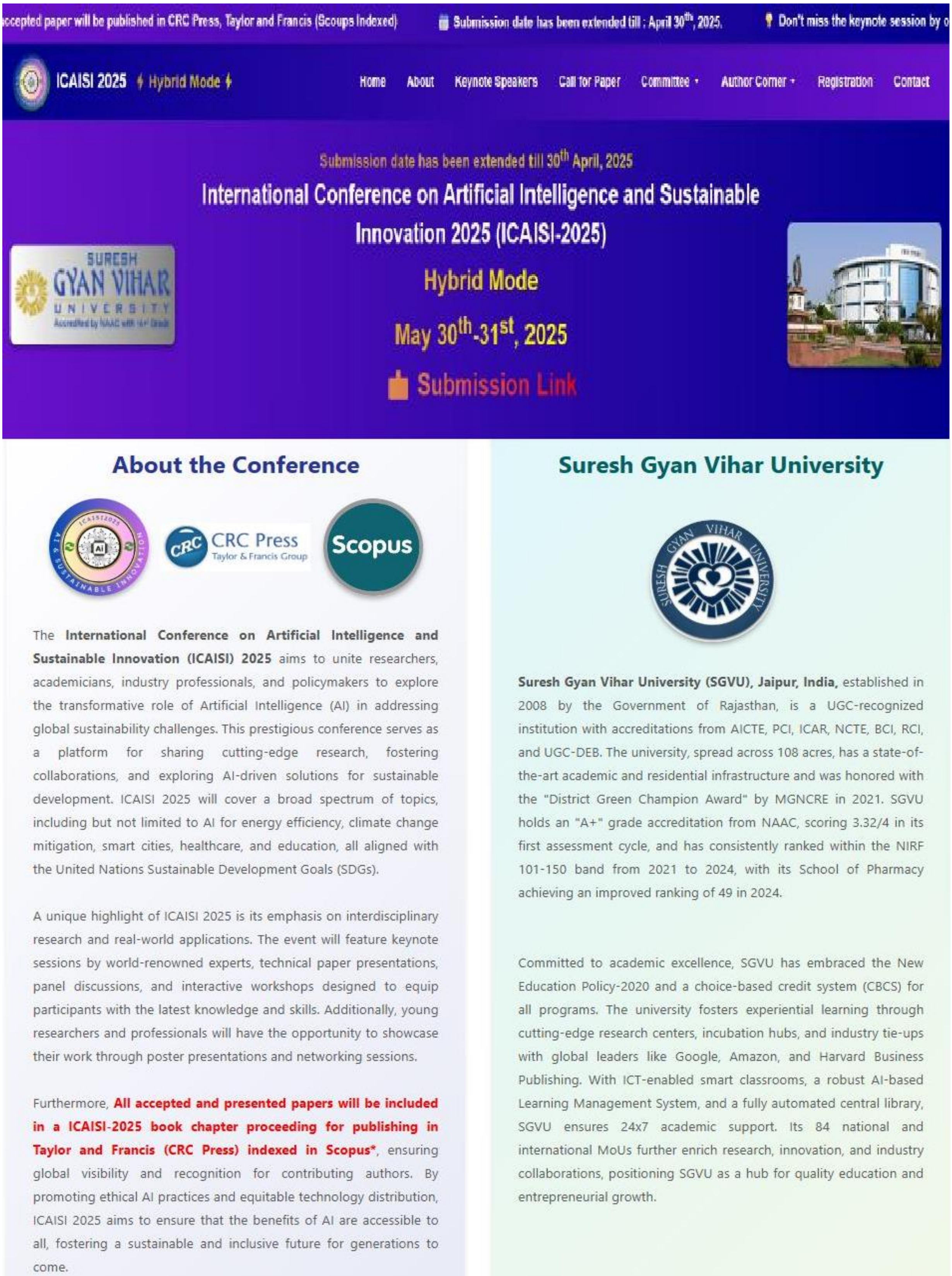
Certificate of Participation

This is to certify that Riti Rathore, Delhi Technological University(Dtu), Delhi has presented his/her research paper titled "Secure Ai In Healthcare: Advanced Privacy-Preserving Machine Learning Techniques For Medical Data" in the ICAISI-2025 organized by Suresh Gyan Vihar University, Jaipur held from May 30th to 31st, 2025.

S. Sharma

Prof (Dr.) Sohit Agarwal Conference Chair	Prof (Dr.) Sandhya Sharma Conference Chair	Dr. Arunansu Haldar President

Fig. 4. Participation Certificate of Paper 1



ccepted paper will be published in CRC Press, Taylor and Francis (Scoups Indexed) Submission date has been extended till : April 30th, 2025. Don't miss the keynote session by o

Submission date has been extended till 30th April, 2025

International Conference on Artificial Intelligence and Sustainable Innovation 2025 (ICAISI-2025)

Hybrid Mode

May 30th-31st, 2025

About the Conference

The **International Conference on Artificial Intelligence and Sustainable Innovation (ICAISI) 2025** aims to unite researchers, academicians, industry professionals, and policymakers to explore the transformative role of Artificial Intelligence (AI) in addressing global sustainability challenges. This prestigious conference serves as a platform for sharing cutting-edge research, fostering collaborations, and exploring AI-driven solutions for sustainable development. ICAISI 2025 will cover a broad spectrum of topics, including but not limited to AI for energy efficiency, climate change mitigation, smart cities, healthcare, and education, all aligned with the United Nations Sustainable Development Goals (SDGs).

A unique highlight of ICAISI 2025 is its emphasis on interdisciplinary research and real-world applications. The event will feature keynote sessions by world-renowned experts, technical paper presentations, panel discussions, and interactive workshops designed to equip participants with the latest knowledge and skills. Additionally, young researchers and professionals will have the opportunity to showcase their work through poster presentations and networking sessions.

Furthermore, **All accepted and presented papers will be included in a ICAISI-2025 book chapter proceeding for publishing in Taylor and Francis (CRC Press) indexed in Scopus***, ensuring global visibility and recognition for contributing authors. By promoting ethical AI practices and equitable technology distribution, ICAISI 2025 aims to ensure that the benefits of AI are accessible to all, fostering a sustainable and inclusive future for generations to come.

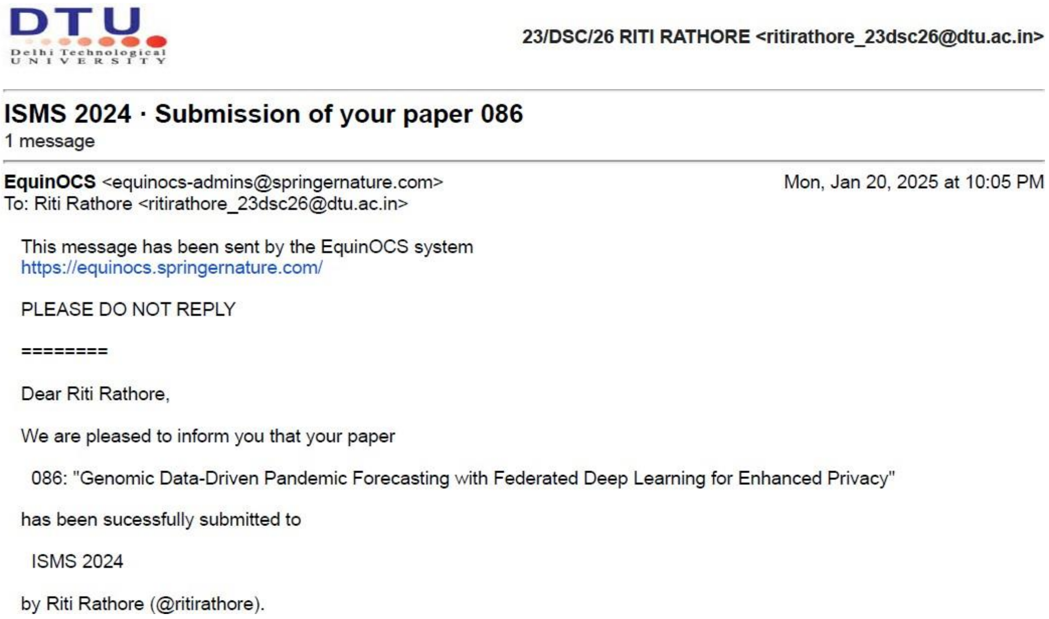
Suresh Gyan Vihar University

Suresh Gyan Vihar University (SGVU), Jaipur, India, established in 2008 by the Government of Rajasthan, is a UGC-recognized institution with accreditations from AICTE, PCI, ICAR, NCTE, BCI, RCI, and UGC-DEB. The university, spread across 108 acres, has a state-of-the-art academic and residential infrastructure and was honored with the "District Green Champion Award" by MGNCRE in 2021. SGVU holds an "A+" grade accreditation from NAAC, scoring 3.32/4 in its first assessment cycle, and has consistently ranked within the NIRF 101-150 band from 2021 to 2024, with its School of Pharmacy achieving an improved ranking of 49 in 2024.

Committed to academic excellence, SGVU has embraced the New Education Policy-2020 and a choice-based credit system (CBCS) for all programs. The university fosters experiential learning through cutting-edge research centers, incubation hubs, and industry tie-ups with global leaders like Google, Amazon, and Harvard Business Publishing. With ICT-enabled smart classrooms, a robust AI-based Learning Management System, and a fully automated central library, SGVU ensures 24x7 academic support. Its 84 national and international MoUs further enrich research, innovation, and industry collaborations, positioning SGVU as a hub for quality education and entrepreneurial growth.

Fig. 5 Indexing Proof of Paper 1

2. Abhilasha Sharma, Riti Rathore, "*Genomic Data-Driven Pandemic Forecasting with Federated Deep Learning for Enhanced Privacy*". The paper has been **Accepted** and **Presented** at the 7th International Conference on Information Systems and Management Science (ISMS 2024), February 2025. Indexed by **Scopus.** Paper Id: 086.



DTU Delhi Technological University

23/DSC/26 RITI RATHORE <ritirathore_23dsc26@dtu.ac.in>

ISMS 2024 · Submission of your paper 086

1 message

EquinOCS <equinocs-admins@springernature.com> Mon, Jan 20, 2025 at 10:05 PM

To: Riti Rathore <ritirathore_23dsc26@dtu.ac.in>

This message has been sent by the EquinOCS system
https://equinocs.springernature.com/

PLEASE DO NOT REPLY

========

Dear Riti Rathore,

We are pleased to inform you that your paper

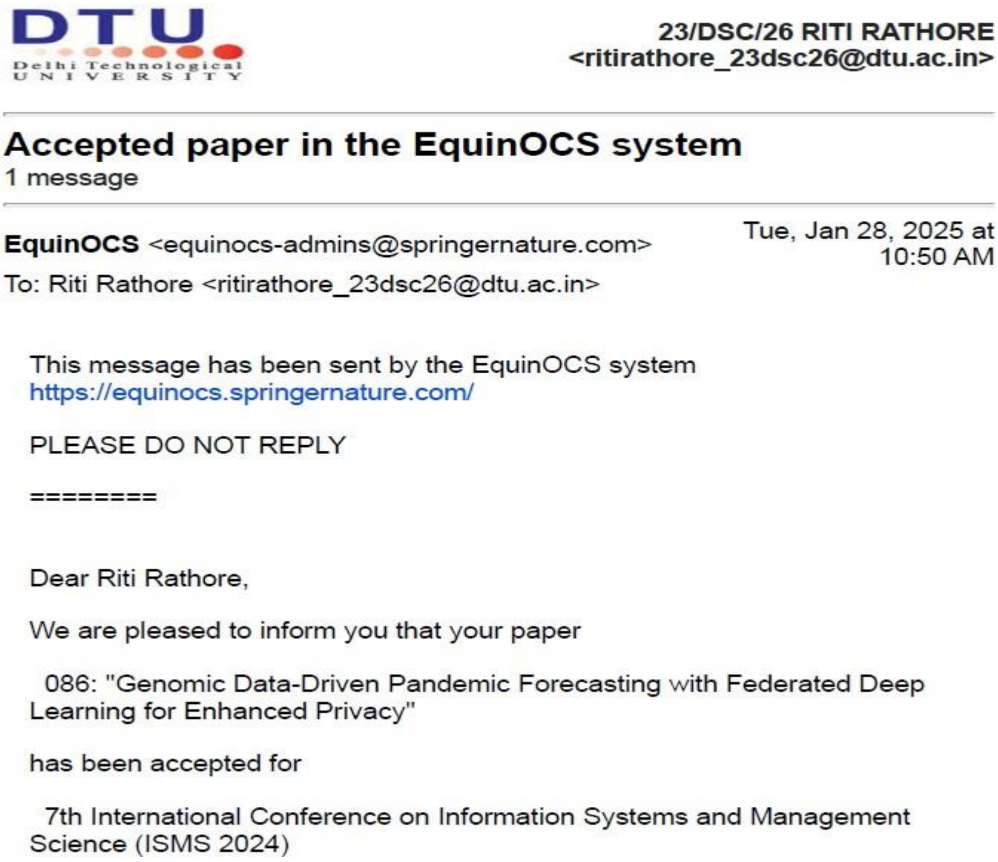
086: "Genomic Data-Driven Pandemic Forecasting with Federated Deep Learning for Enhanced Privacy"

has been sucessfully submitted to

ISMS 2024

by Riti Rathore (@ritirathore).

Fig. 6. Submission Proof of Paper 2



DTU Delhi Technological University

23/DSC/26 RITI RATHORE <ritirathore_23dsc26@dtu.ac.in>

Accepted paper in the EquinOCS system

1 message

EquinOCS <equinocs-admins@springernature.com> Tue, Jan 28, 2025 at 10:50 AM

To: Riti Rathore <ritirathore_23dsc26@dtu.ac.in>

This message has been sent by the EquinOCS system
https://equinocs.springernature.com/

PLEASE DO NOT REPLY

========

Dear Riti Rathore,

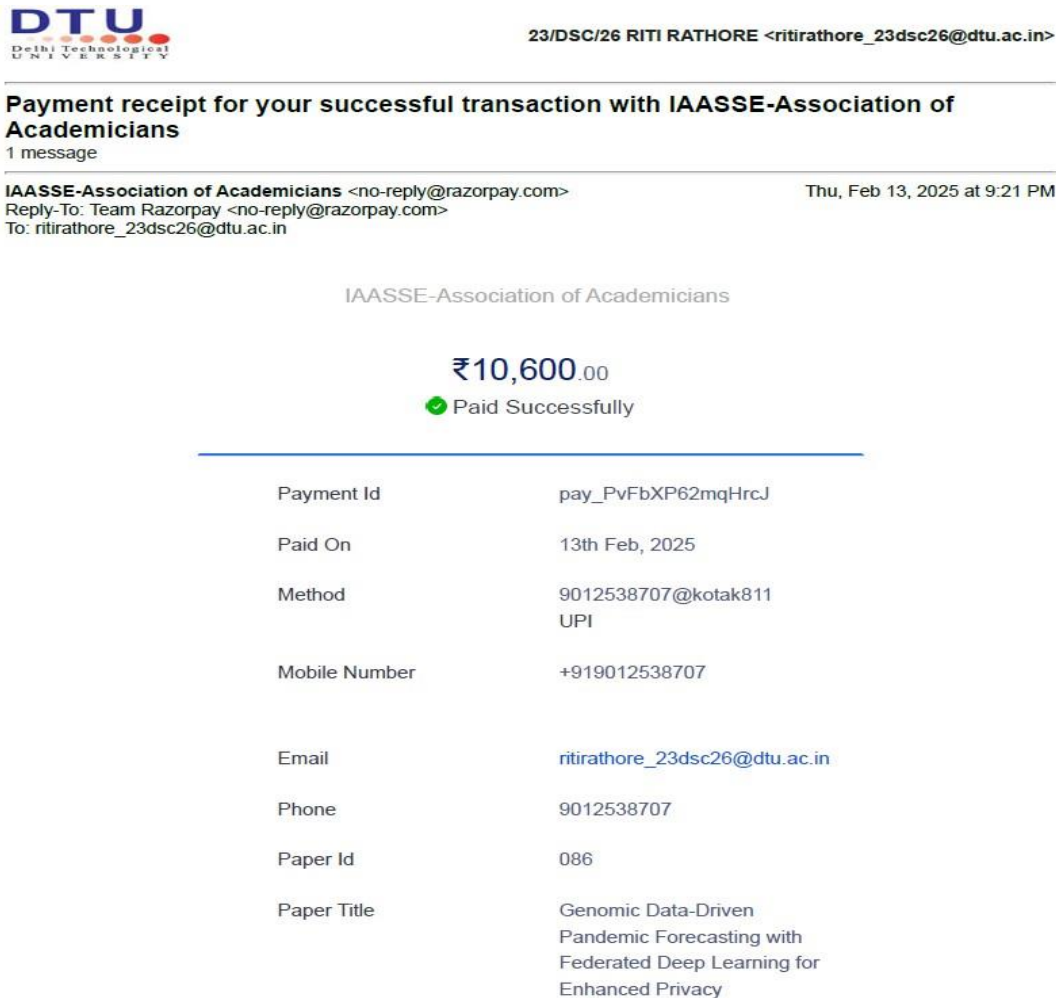
We are pleased to inform you that your paper

086: "Genomic Data-Driven Pandemic Forecasting with Federated Deep Learning for Enhanced Privacy"

has been accepted for

7th International Conference on Information Systems and Management Science (ISMS 2024)

Fig. 7. Acceptance Proof of Paper 2



DTU Delhi Technological University

23/DSC/26 RITI RATHORE <ritirathore_23dsc26@dtu.ac.in>

Payment receipt for your successful transaction with IAASSE-Association of Academicians
1 message

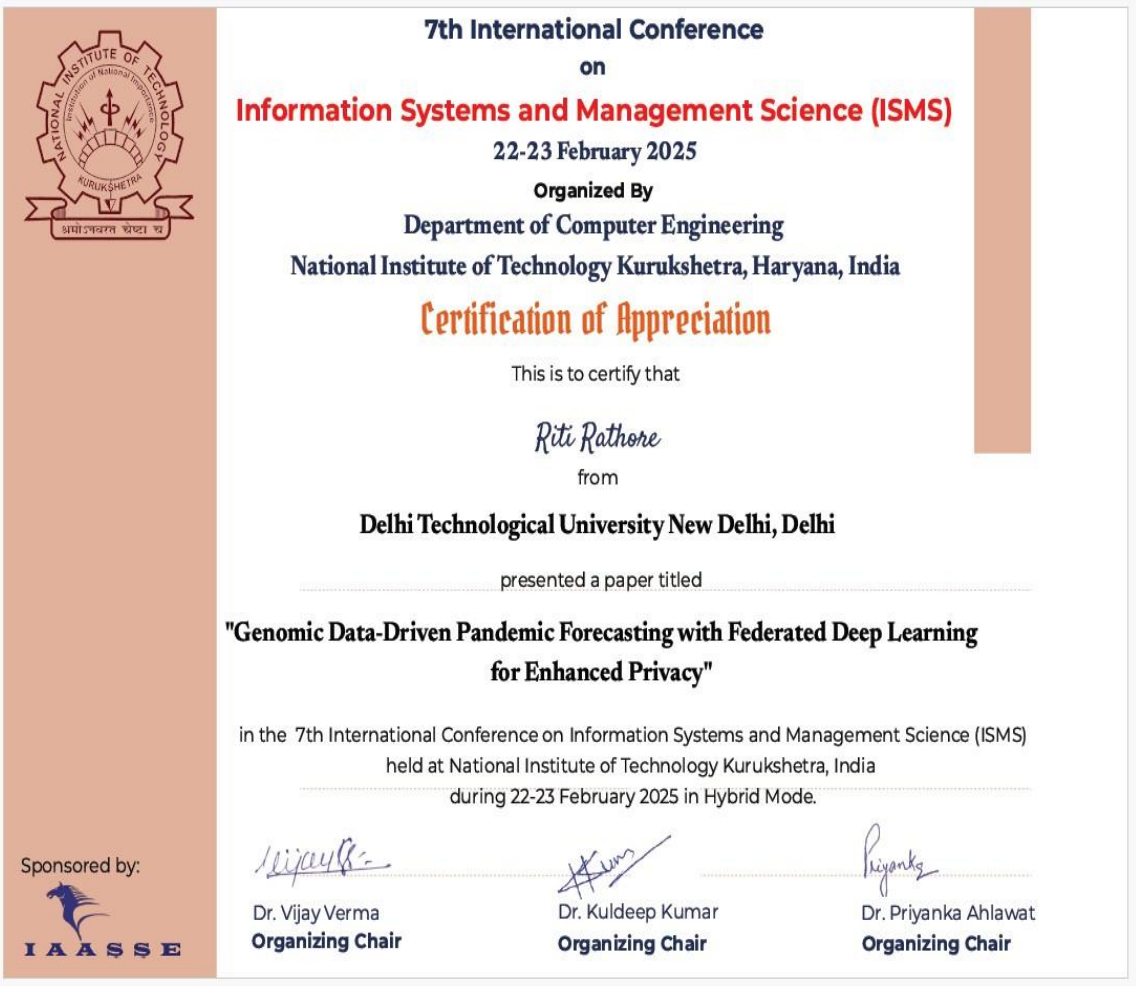
IAASSE-Association of Academicians <no-reply@razorpay.com> Thu, Feb 13, 2025 at 9:21 PM
Reply-To: Team Razorpay <no-reply@razorpay.com>
To: ritirathore_23dsc26@dtu.ac.in

IAASSE-Association of Academicians

₹10,600.00
✓ Paid Successfully

Payment Id	pay_PvFbXP62mqHrcJ
Paid On	13th Feb, 2025
Method	9012538707@kotak811 UPI
Mobile Number	+919012538707
Email	ritirathore_23dsc26@dtu.ac.in
Phone	9012538707
Paper Id	086
Paper Title	Genomic Data-Driven Pandemic Forecasting with Federated Deep Learning for Enhanced Privacy

Fig. 8. Payment Proof of Paper 2



7th International Conference
on
Information Systems and Management Science (ISMS)
22-23 February 2025
Organized By
Department of Computer Engineering
National Institute of Technology Kurukshetra, Haryana, India

Certification of Appreciation

This is to certify that

Riti Rathore

from

Delhi Technological University New Delhi, Delhi

presented a paper titled

"Genomic Data-Driven Pandemic Forecasting with Federated Deep Learning for Enhanced Privacy"

in the 7th International Conference on Information Systems and Management Science (ISMS) held at National Institute of Technology Kurukshetra, India during 22-23 February 2025 in Hybrid Mode.

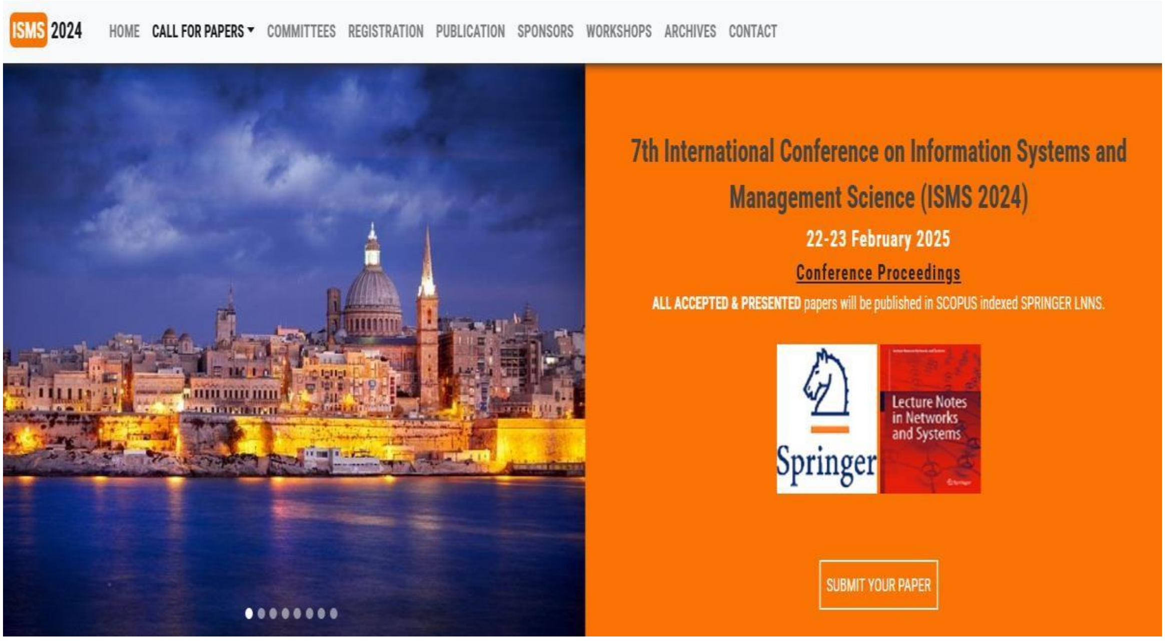
Dr. Vijay Verma
Organizing Chair

Dr. Kuldeep Kumar
Organizing Chair

Dr. Priyanka Ahlawat
Organizing Chair

Sponsored by:
IAASSE

Fig. 9. Participation Certificate of Paper 2



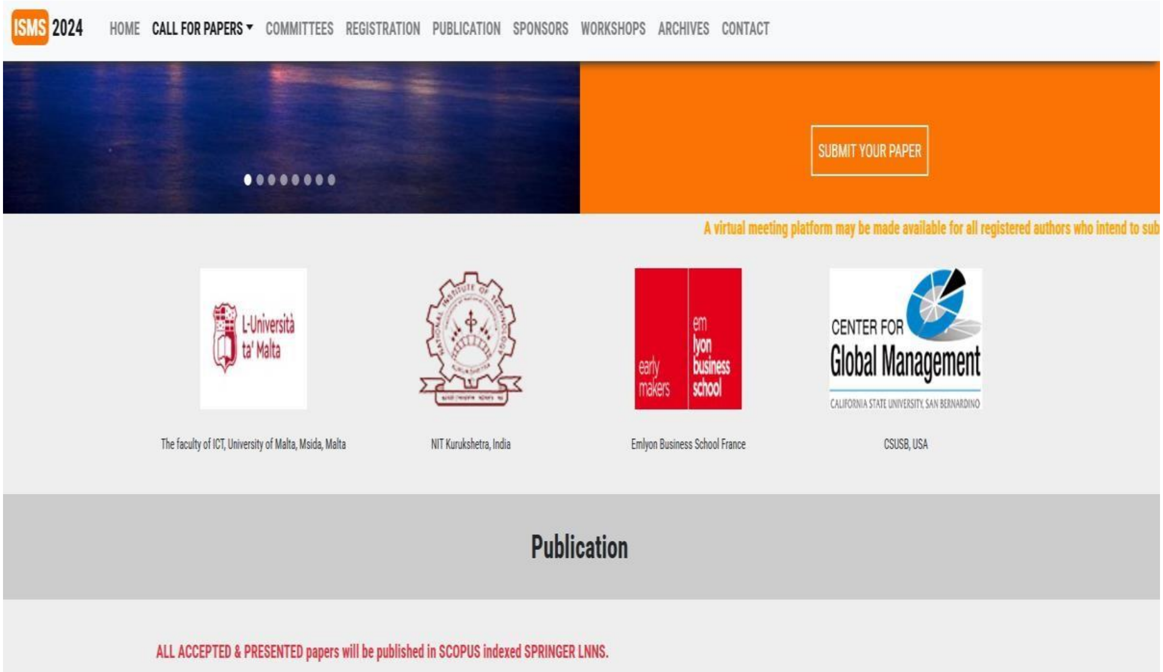
7th International Conference on Information Systems and Management Science (ISMS 2024)

22-23 February 2025

Conference Proceedings

ALL ACCEPTED & PRESENTED papers will be published in SCOPUS indexed SPRINGER LNNS.

Fig. 10. Indexing Proof of Paper 2



A virtual meeting platform may be made available for all registered authors who intend to sub

The faculty of ICT, University of Malta, Msida, Malta

NIT Kurukshetra, India

Emlyon Business School France

CSUSB, USA

Publication

ALL ACCEPTED & PRESENTED papers will be published in SCOPUS indexed SPRINGER LNNS.

Fig. 11. Indexing Proof of Paper 2

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

PLAGIARISM VERIFICATION

Title of the Thesis Privacy Preserving Machine Learning in Healthcare for Pandemic Prediction using Genomic Data

Total Pages 55 Name of the Scholar Riti Rathore

Supervisor (s)

(1) Dr. Abhilasha Sharma

(2)

(3)

Department Department of Software Engineering

This is to report that the above thesis was scanned for similarity detection. Process and outcome is given below:

Software used: Turnitin Similarity Index: 9 %, Total Word Count: 10,031

Date: 20/05/2025

Riti Rat.

Candidate's Signature

Abhilasha Sharma

Signature of Supervisor(s)

thesis- copy Rithi-10-54.pdf

Delhi Technological University

Document Details

Submission ID

trn:oid:::27535:96815553

Submission Date

May 20, 2025, 12:38 PM GMT+5:30

Download Date

May 20, 2025, 12:41 PM GMT+5:30

File Name

thesis- copy Rithi-10-54.pdf

File Size

1.6 MB

45 Pages

10,031 Words

58,721 Characters

9% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Cited Text

Match Groups

74 Not Cited or Quoted 9%
Matches with neither in-text citation nor quotation marks

0 Missing Quotations 0%
Matches that are still very similar to source material

1 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

4%	Internet sources
5%	Publications
6%	Submitted works (Student Papers)

Integrity Flags

1 Integrity Flag for Review

Hidden Text
73 suspect characters on 1 page
Text is altered to blend into the white background of the document.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

thesis- copy Rithi-10-54.pdf

Delhi Technological University

Document Details

Submission ID trn:oid:::27535:96815553	45 Pages
Submission Date May 20, 2025, 12:38 PM GMT+5:30	10,031 Words
Download Date May 20, 2025, 12:41 PM GMT+5:30	58,721 Characters
File Name thesis- copy Rithi-10-54.pdf	
File Size 1.6 MB	

*% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Disclaimer
Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (it may misidentify writing that is likely AI generated as AI generated and AI paraphrased or likely AI generated and AI paraphrased writing as only AI generated) so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?
The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?
Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.

DECLARATION

We/I hereby certify that the work which is presented in the Major Project-II/Research Work entitled Privacy preserving ML In Healthcare for Pandemic Prediction using Genome Data in fulfilment of the requirement for the award of the Degree of Bachelor/Master of Technology in M-tech and submitted to the Department of SE, Delhi Technological University, Delhi is an authentic record of my/our own, carried out during a period from 2023-2025, under the supervision of Dr. Abhilasha.

The matter presented in this report/thesis has not been submitted by us/me for the award of any other degree of this or any other Institute/University. The work has been published/accepted/communicated in SCI/ SCI expanded/SSCI/Scopus indexed journal OR peer reviewed Scopus indexed conference with the following details:

Title of the Paper: Secure AI In Healthcare : Advanced Privacy Preserving ML techniques for medical data
Author names (in sequence as per research paper): Abhilasha Sharma, Riti Rathore
Name of Conference/Journal: ICAISI - 2025
Conference Dates with venue (if applicable): 30-05-2025 (hybrid)
Have you registered for the conference (Yes/No)?: yes
Status of paper (Accepted/Published/Communicated): Accepted
Date of paper communication: 26/4/2025
Date of paper acceptance: 4/05/2025
Date of paper publication: yet to be done

Student(s) Roll No., Name and Signature

Riti Rathore
23/DSC/26

SUPERVISOR CERTIFICATE

To the best of my knowledge, the above work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere. I, further certify that the publication and indexing information given by the students is correct.

Abhilasha Sharma

Place: Delhi

Date: 26/5/2025

Supervisor Name and Signature

NOTE: PLEASE ENCLOSE RESEARCH PAPER ACCEPTANCE/ PUBLICATION/COMMUNICATION PROOF ALONG WITH SCOPUS INDEXING PROOF (Conference Website OR Science Direct in case of Journal Publication).

DECLARATION

We/I hereby certify that the work which is presented in the Major Project-II/Research Work entitled Privacy preserving ML in Healthcare for Pandemic Prediction Using Genomic Data in fulfilment of the requirement for the award of the Degree of Bachelor/Master of Technology in Mtech and submitted to the Department of SC, Delhi Technological University, Delhi is an authentic record of my/our own, carried out during a period from 2023-25, under the supervision of Dr Abhilasha Sharma.

The matter presented in this report/thesis has not been submitted by us/me for the award of any other degree of this or any other Institute/University. The work has been published/accepted/communicated in SCI/ SCI expanded/SSCI/Scopus indexed journal OR peer reviewed Scopus indexed conference with the following details:

Title of the Paper: Genomic Data Driven Pandemic Forecasting with federated learning for Enhanced Privacy
Author names (in sequence as per research paper): Abhilasha Sharma, Riti Rathore
Name of Conference/Journal: ISMS 2024
Conference Dates with venue (if applicable): 22nd Feb 2025
Have you registered for the conference (Yes/No)?: Yes
Status of paper (Accepted/Published/Communicated): Accepted
Date of paper communication: 20-01-2025
Date of paper acceptance: 28-01-2025
Date of paper publication: yet to be done

Student(s) Roll No., Name and Signature

Riti Rathore
23/DSC/26

SUPERVISOR CERTIFICATE

To the best of my knowledge, the above work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere. I, further certify that the publication and indexing information given by the students is correct.

Abhilasha Sharma

Supervisor Name and Signature

Place: Delhi

Date: 26/5/2025

NOTE: PLEASE ENCLOSE RESEARCH PAPER ACCEPTANCE/ PUBLICATION/COMMUNICATION PROOF ALONG WITH SCOPUS INDEXING PROOF (Conference Website OR Science Direct in case of Journal Publication).