International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 12 | Dec - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Amplifying the Privacy Protection for the Distributed, Structured, and Concerted Data in the Healthcare

K. Ajay[1], K. Sathvika[2], M. Pavan Kumar[3], Arun Singh Kaurav[4]
UG Scholar, Guru Nanak Institutions Technical Campus, Hyderabad, Telangana[1,2]
Assistant Professor, Guru Nanak Institutions Technical Campus, Hyderabad, Telangana [3]

ABSTRACT

Machine learning has gained significant importance in healthcare for enhancing decision-making and improving predictions. However, healthcare data is often distributed across various locations, such as multiple hospitals, which presents a challenge in analyzing this data while protecting sensitive personal information. A new approach ensures privacy during data analysis. We describe the implementation of this method and evaluate its performance using medical datasets. Our method demonstrates how machine learning can be applied to sensitive healthcare data while protecting and safeguarding the privacy.

1. INTRODUCTION

Artificial intelligence and automated decision-making are able to improve the performance parameters like accuracy and efficiency of healthcare applications. Artificial Intelligence has already been shown to outperform medical experts in some areas. For example, deep neural networks are used to classify the rhythms in electrocardiogram (ECG) signals, and the AI systems have been applied to predict breast cancer; other relate studies can be found in. However, it also presents challenges, particularly around security and privacy. A key concern is that sharing a patient's medical data with a third party could unintentionally expose sensitive information, such as the existence of a medical condition.

In the distributed healthcare settings, hospitals have to use some data mining methods to extract valuable insights from patient data. While hospitals can use their own resources and locally stored health information for data mining, sharing data across multiple hospitals can provide more accurate and useful results. However, this comes with challenges due to privacy and legal concerns. Hospitals must often follow strict privacy regulations that limit the sharing of patient data with other parties, such as other hospitals, family doctors, or specialists. A similar issue arises when data is distributed across patients' personal devices, like mobile phones or wearable devices.

1.1. OBJECTIVE

The topic of the collective learning from distributed data has been explored in the literature for many years. Many distributed learning techniques have been proposed, although they do not always focus specifically on privacy concerns. However, these techniques can still help protect privacy by minimizing the amount of data that needs to be shared with other parties or transferred to central servers or the cloud.

1.2 SCOPE OF THE WORK

In this work, we tackle the issue of learning from data distributed across multiple sources without sharing raw data directly. We assume the data is partitioned in the horizontal manner, with different records stored across various locations. Our primary focus is on solving classification tasks using

structured health data, such as data stored in spreadsheets. Building on our prior work, we propose a scalable, privacy-protecting and safeguarding framework for distributed machine learning. This approach provides minimum overhead with respect to the number of participating parties and is capable of how to handle the missing data.

1.3 EXPLANATION:

Designing this system involves creating a decentralized infrastructure where users can securely interact with applications. These interactions allow for managing code storage, tracking balances, and utilizing external data. Below is a structured explanation of how such a system operates.

1.4. PROPOSED SYSTEM:

We develop a minimum word transportation cost metric to assess the similarity between queries and documents. We focus on the challenge of learning from data distributed across multiple sources without directly sharing raw data. The data is assumed to be horizontally partitioned, meaning that different records are stored across various sources. Specifically, we address the classification problem using structured health data, typically stored in spreadsheets. Our method ensures the secure execution of machine learning models over distributed data, protecting sensitive health information throughout the process.

The proposed system enhances security by utilizing encryption techniques to protect sensitive data from unauthorized access, ensuring privacy. It also reduces storage costs by optimizing data storage and employing space-efficient encryption methods. Additionally, the system enables secure semantic matching on encrypted data, allowing accurate data comparison and retrieval while maintaining confidentiality. This approach ensures that data can be processed and searched without compromising security or privacy.

1.5 DESCRIPTION

We assume that owner is trustworthy, and the users have been authorized by the owner. Communication between the owner and users is secured using protocols such as Secure socket layer and Transport layer Security. Our proposed approach addresses a more advanced security model for the cloud server. In our model, a dishonest cloud server may return inaccurate or tampered search results and attempt to infer sensitive information. However, it will not intentionally delete or modify the documents. As a result, our secure semantic search scheme ensures both verifiability and confidentiality within this security framework.

2. METHODOLOGY

2.1.USER INTERFACE DESIGN:

This module provides a secure login interface. Users enter their username and password to access the server. New users must register with their username, password, and email. After registration, the server creates an account, setting the user ID to the user's name and tracking upload/download rates. Once logged in, users are directed to a page for actions like uploading or downloading data.

2.2. USER:

The module allows users to register and login to the system. After logging in, the user can search for files by name and download files, which are initially displayed in an encrypted format. The user can also send a trapdoor request to the server, which, if accepted, allows the user to request permission from the file owner. Once permission is granted, the file is downloaded in plain text.

2.3. OWNER:

The module allows users to register and login. Once logged in, the Data Owner can upload files to the database and send requests to Data Users

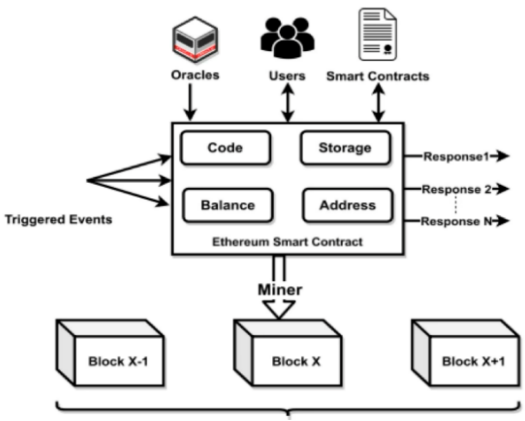
2.4. CLOUD SERVER:

The module allows the server to login and access all Data Owners' and users' information. It can view all stored data files and send key requests to users. Additionally, the Cloud Server can monitor and review any attacker activity related to files.

2.5 EXISTING WORK

The existing algorithm already is straightforward. It is simple and its effectiveness made it popular in diverse fields like image recognition, medical diagnostics, and recommendation systems. But It faces many challenges in the secure environments, mainly when dealing with the encrypted data.

2.6 SYSTEM ARCHITECTURE:



3 RESULTS:

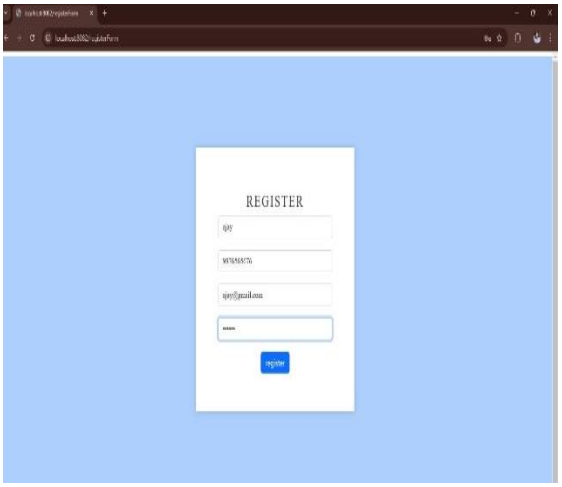


Figure.1

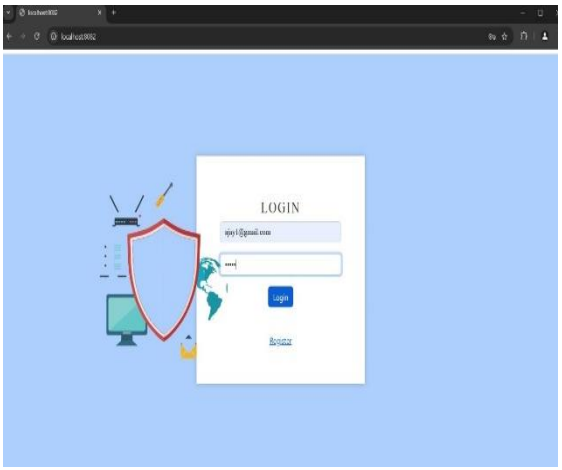


Figure.2

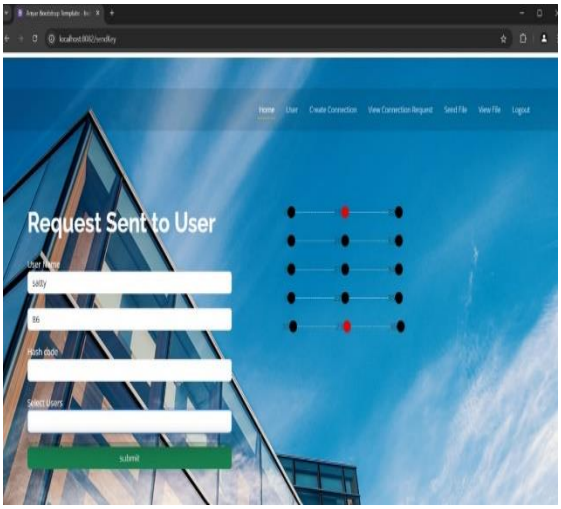


Figure.3

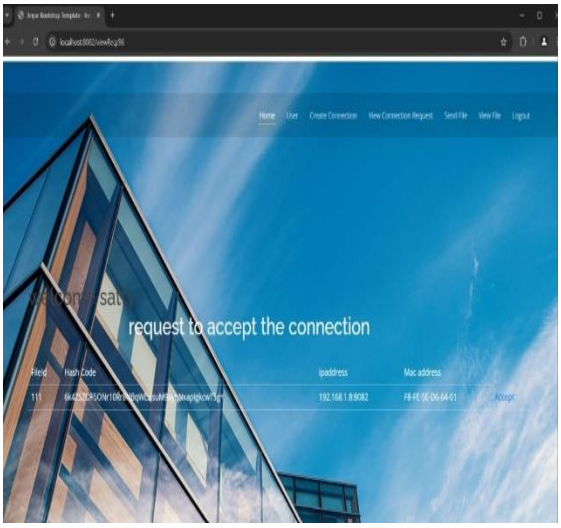


Figure.4

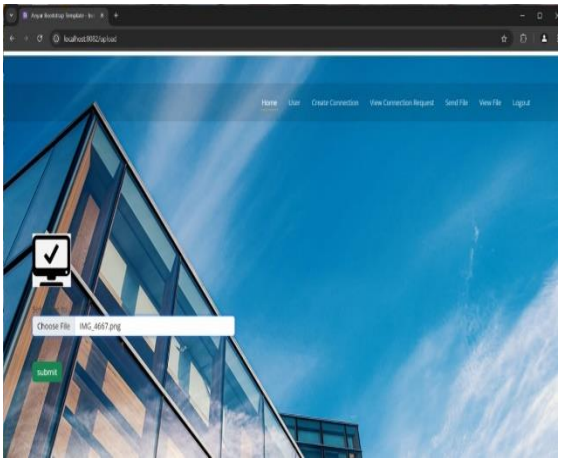


Figure.5

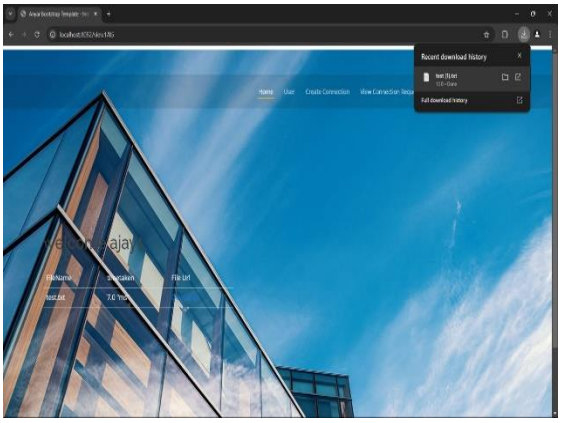


Figure.6

4. FUTURE ENHANCEMENT

In the future, we aim to extend the framework to handle scenarios where parties do not follow the honest-but-curious model, addressing more complex security challenges. Future work will explore extending the framework to handle more complex security models.

5. CONCLUSION

Our proposed work is better than the existing models by around 10 -15% in the overall score, around 13-14% in the accuracy, and around 0.250 in the healthcare datasets. We also demonstrate the implementation on cloud to evaluate the latency and also the scalability. The algorithm also has the linear overhead with respect to the number of parties and also can handle missing data. Our framework provides very efficient, scalable, and secure machine learning models, enhancing healthcare systems without compromising the privacy.

6. REFERENCES

[1] Y. Hannun, P. Rajpurkar, M. Haghpanahi, G. H. Tison, C. Bourn, M. P. Turakhia, and A. Y. Ng, ‘‘Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network,’’ Nature Med., vol. 25, no. 1, pp.65–69,Jan.2019.

[2] S. McKinney et al., ‘‘International evaluation of an AI system for breast cancer screening,’’ Nature, vol.577,pp.89–94,Jan.2020.

[3] X. Liu, L. Faes, A. U. Kale, S. K. Wagner, D. J. Fu, A. Bruynseels, T. Mahendiran, G. Moraes, M. Shamdas, C. Kern, J. R. Ledsam, M. K. Schmid, K. Balaskas, E. J. Topol, L. M. Bachmann, P. A. Keane, and A. K. Denniston, ‘‘A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis,’’ Lancet Digit. Health, vol. 1, no. 6, pp. e271–e297, Oct. 2019.

[4] R. Aggarwal, V. Sounderajah, G. Martin, D. S. W. Ting, A. Karthikesalingam, D. King, H. Ashrafian, and A. Darzi, ‘‘Diagnostic accuracy of deep learning in medical imaging: A systematic review and meta-analysis,’’ npj Digit. Med., vol. 4, no. 1,p.65,Dec.2021.

[5] S. D. Lustgarten, Y. L. Garrison, M. T. Sinnard, and A. W. Flynn, ‘‘Digital privacy in mental healthcare: Current issues and recommendations for technology use,’’ Current Opinion Psychol., vol. 36, pp.25–31,Dec.2020.

[6] D. Pascual, A. Amirshahi, A. Aminifar, D. Atienza, P. Ryvlin, and R. Wattenhofer, ‘‘EpilepsyGAN: Synthetic epileptic brain activities with privacy preservation,’’ IEEE Trans. Biomed. Eng., vol. 68, no. 8, pp. 2435–2446, Aug. 2021

[7] A. Saeed, F. D. Salim, T. Ozcelebi, and J. Lukkien, ‘‘Federated self-supervised learning of multisensory representations for embedded intelligence,’’ IEEE Internet Things J., vol. 8, no. 2, pp. 1030–1040, Jan. 2021.

[8] F. Forooghifar, A. Aminifar, and D. Atienza, ‘‘Resource-aware distributed epilepsy monitoring using self-awareness from edge to cloud,’’ IEEE Trans. Biomed. Circuits Syst., vol. 13, no. 6, pp. 1338–1350,Dec.2019.

[9] D. Sopic, A. Aminifar, A. Aminifar, and D. Atienza, ‘‘Real-time eventdriven classification technique for early detection and prevention of myocardial infarction on wearable systems,’’ IEEE Trans. Biomed. Circuits Syst., vol. 12, no. 5, pp. 982–992,Oct.2018.

[10] D. Sopic, A. Aminifar, A. Aminifar, and D. Atienza, ‘‘Real-time classification technique for early detection and prevention of myocardial infarction on wearable devices,’’ in Proc. IEEE Biomed. Circuits Syst. Conf. (BioCAS), Oct. 2017, pp. 1–4.

[11] R. Zanetti, A. Arza, A. Aminifar and D. Atienza, ‘‘Real-time EEG-based cognitive workload monitoring on wearable devices,’’ IEEE Trans.

Biomed. Eng., vol. 69, no. 1, pp. 265–277, Jan. 2022. [Online].Available:https://ieeexplore.ieee.org/document/9464276, doi: 10.1109/TBME.2021.3092206.

[12] F. Emekci, O. D. Sahin, D. Agrawal, and A. El Abbadi, ‘‘Privacy preserving decision tree learning over multiple parties,’’ Data Knowl. Eng., vol. 63, no. 2,pp.348–361,Nov.2007.

[13] J. S. Davis and O. Osoba, ‘‘Improving privacy preservation policy in the modern information age,’’ Health Technol., vol. 9, no. 1, pp. 65–75, Jan. 2019.

[14] J. Vaidya, B. Shafiq, W. Fan, D. Mehmood, and D. Lorenzi, ‘‘A random decision tree framework for privacy-preserving data mining,’’ IEEE Trans. Dependable Secure Comput., vol. 11, no. 5, pp. 399–411,Sep.2014.

[15] P. Jurczyk and L. Xiong, ‘‘Distributed anonymization: Achieving privacy for both data subjects and data providers,’’ in Proc. IFIP Annu. Conf. Data Appl. Secur. Privacy. Berlin, Germany: Springer, 2009. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-03007-9_13

[16] L. Sweeney, ‘‘K-anonymity: A model for protecting privacy,’’ Int. J. Uncertainty, Fuzziness Knowl.-Based Syst., vol. 10, no. 5, pp. 557–570, Oct. 2002.

[17] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, ‘‘Ldiversity: Privacy beyond K-anonymity,’’ ACM Trans. Knowl. Discovery Data, vol. 1, no. 1, p. 3, 2007. [18] N. Li, T. Li, and S. Venkatasubramanian, ‘‘T-closeness: Privacy beyond K-anonymity and L-diversity,’’ in Proc. IEEE 23rd Int. Conf. Data Eng., Apr. 2007, pp. 106–115.

[19] A. Aminifar, Y. Lamo, K. Pun, and F. Rabbi, ‘‘A practical methodology for anonymization of structured health data,’’ in Proc. 17th Scandin. Conf. Health Informat. 2019, pp. 127–133. [Online]. Available:https://ep.liu.se/en/conference-article.aspx?series= ecp&issue=161&Article_No=22

[20] A. Aminifar, F. Rabbi, V. K. I. Pun, and Y. Lamo, ‘‘Diversity-aware anonymization for structured health data,’’ in Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC), Nov. 2021, pp. 2148–2154.

[21] Health Informatics—Pseudonymization, International Organization for Standardization, Geneva, Switzerland, Standard ISO 25237:2017, Jan. 2017[Online].Available: https://www.iso.org/standard/63553.html

[22] C. Dwork, ‘‘Differential privacy,’’ in Proc. 33rd Int. Colloq. Automata, Lang., Program. (ICALP) (Lecture Notes in Computer Science). Berlin, Germany: Springer-Verlag, 2006. [Online]. Available: https://link.springer.com/chapter/10.1007/11787006_1

[23] M. Kantarcioglu, ‘‘A survey of privacy-preserving methods across horizontally partitioned data,’’ in Privacy-Preserving Data Mining. Boston, MA, USA: Springer, 2008, pp. 313–335. [Online]. Availablehttps://link.springer.com/chapter/10.1007/978-0-387-70992-5_13

[24] J. Vaidya, ‘‘A survey of privacy-preserving methods across vertically partitioned data,’’ in Privacy-Preserving Data Mining. Boston, MA, USA: Springer, 2008, pp. 337–358. [Online]. Available: https://link.springer.com/chapter/10.1007/978-0-387-70992-5_14.

[25] W. Du and Z. Zhan, ‘‘Building decision tree classifier on private data,’’ in Proc. IEEE Int. Conf. Privacy, Secur. Data Mining (CRPIT), vol. 14.

Australia: Austral. Comput. Soc., 2002, pp. 1–8. [Online].Available: https://dl.acm.org/doi/10.5555/850782.850784

[26] J. Konečný, H. Brendan McMahan, D. Ramage, and P. Richtárik, ‘‘Federated optimization: Distributed machine learning for on-device intelligence,’’ 2016, arXiv: 1610.02527.

[27] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Agüera y Arcas, ‘‘Communication-efficient learning of deep networks from decentralized data,’’ 2016, arXiv:1602.05629.