Proceedings of the 2nd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/96/20241280

Implementation of Distributed Machine Learning in Medicine Industry

Yaoyang Xia

School of Computer Science and Technology, Anhui University, Hefei, China

E12214093@stu.ahu.edu.cn

Abstract. As a matter of fact, machine learning (ML) has become a cornerstone in modern science and engineering, evolving from early algorithms like the perceptron to sophisticated deep learning models. The rise of distributed machine learning (DML) has enabled scalable, privacy-preserving analysis of large healthcare datasets in recent years. With this in mind, this study summarizes the principles of DML, contrasting it with traditional as well as parallel computing approaches. At the same time, this research discusses its applications in medical imaging, predictive analytics, and real-time monitoring. According to the analysis, despite significant advancements, challenges related to data privacy, heterogeneity, and scalability persist. Based on the evaluations, future research is needed to overcome these limitations and further integrate DML into healthcare. Overall, these results underscore the transformative potential of DML in enhancing healthcare delivery and outcomes and pave a path as well as offer a guideline for implementations of machine leaning in medic industry.

Keywords: Distributed machine learning, federated learning, medical imaging, predictive analytics, privacy preservation.

1. Introduction

Machine learning (ML) is integrated into our daily lives and is a standard tool in many fields of science and engineering. The concept of machine learning originated in the mid-20th century, when early computational models and algorithms laid the foundation for artificial intelligence (AI). The perceptron algorithm developed by Rosenblatt in 1958 was a leap forward as a simple linear classifier that mimics biological neurons [1]. This innovation is one of the first attempts to mimic human cognition and sets the stage for future neural network research. In the 1980s and 1990s, there were significant advances in machine learning. In particular, the backpropagation algorithm was proposed, which made the training of multi-layer neural networks more efficient [2]. This period also saw the emergence of SVMS and other methods that were critical to improving the task of pattern recognition and classification in high-dimensional Spaces [3].

Deep learning models are used well in image classification tasks and are widely adopted in various fields such as healthcare. Nowadays, considering the huge data sets and the scarcity of computing resources, the ability of deep learning to automatically learn from raw data has a wide range of applications and plays an important role in various fields. In recent years, distributed machine learning (DML) has gradually replaced traditional centralized machine learning, especially when it comes to handling large-scale data and computationally intensive tasks. Distributed machine learning involves training models across multiple machines or nodes, it can realize parallel processing and greatly reduce model convergence time. The method is widely used in medical image analysis, where data sets are often large and require a lot of computing power to process.

An important milestone in distributed machine learning is the development of federated learning, which enables training models across multiple decentralized devices or servers without sharing raw data. This has advantages in the medical field where the privacy of patient data cannot be compromised. Recent studies have confirmed that federated learning excels in training robust models for medical imaging tasks such as tumor detection and segmentation [4, 5]. These studies highlight the potential of joint learning to enhance inter-agency collaboration while maintaining data privacy and security. As the volume and complexity of healthcare data increases, we are increasingly using distributed machine learning in healthcare. Examples include electronic health records (EHRs), medical images, and real-time monitoring data from wearable devices. In addition, with the rise of edge computing and the Internet of Medical Things (IoMT), there is a growing need for real-time analysis of medical devices. DML frameworks improve the efficiency and responsiveness of healthcare delivery by enabling models to be trained and deployed at the edge [5].

2. Descriptions of distributed machine learning algorithms

Traditional machine learning is trained on a single model with all the data in one place. The efficiency of this calculation is very low, and the performance is not very good when the data set is large and the model complexity is high. In addition, distributed machine learning sends the task of raw data and computation to multiple machines, which can train the model in parallel, which has two main modes, data parallelism and model parallelism. Data parallelism involves placing complete models on different devices and then dividing the data into parallel computations on each device. Data parallelism divides the data set into multiple subsets, each assigned to a different compute node (GPU) with a full copy of the model on each node. Each node processes a different subset of data, calculates the gradient, synchronizes the gradient through set communication, and updates the model parameters. Model parallelism is a technique in which a single model is divided into multiple machines, each is responsible for a part of the computational model. This approach is especially useful when the model itself is too large to work with a single machine, such as a deep neural network with millions of parameters. In this setup, each node computes the forward and backward transmission of the neural network for itself and communicates with other nodes to share the intermediate output. The primary difference between DML and traditional machine learning lies in the data and computation management. While traditional machine learning relies on a centralized approach where all data and computations occur on a single node, distributed machine learning utilizes a decentralized framework to divide tasks and resources across multiple nodes, thereby enhancing scalability and efficiency [6].

Distributed machine learning also differs significantly from conventional parallel computing. In parallel computing, the focus is on subdividing a single task into smaller subtasks that can be executed simultaneously on multiple processors on the same machine. This process, however, assumes a shared memory architecture and low-latency communication between processors. In contrast, distributed machine learning operates on a network of independent machines with their local memory and storage. This necessitates more sophisticated algorithms to handle the high-latency and potentially unreliable communication between nodes.

3. Data pre-processing

Data preprocessing is of great significance in distributed machine learning, and addressing data privacy and heterogeneity is a challenge. This study will typically collect medical data from multiple institutions, so the data is fragmented and varies in format and quality. So standardizing data across institutions is the first step. In Federated Learning (FL), a widely used DML framework, data is kept local at each site and only model updates are shared [7]. This study typically uses techniques such as data normalization (standardizing pixel intensity), data enhancement (rotating, flipping), and processing missing data to address the heterogeneity of medical images [8]. Privacy-preserving methods like differential privacy and secure multi-party computation are applied to ensure that sensitive patient information is not compromised during the model training process [7].

For distributed medical image analysis, CNNs are often employed due to their ability to capture spatial hierarchies in images. CNNs are particularly effective in tasks such as detecting anomalies in CT and MRI scans [9]. This approach maintains data privacy while benefiting from the collaborative learning of a more generalized model. For time-series medical data, such as patient histories, RNNs and LSTM networks are more appropriate due to their capability to capture temporal dependencies. Federal learning architectures such as FedAvg (Federal Average) algorithms are often used in healthcare Settings. The architecture enables each institution (the client) to perform local updates to the model before sending updated model parameters to a central server for aggregation. This process iterates until convergence, enabling the model to learn across a wide range of medical data sources without centralizing sensitive data [10] In conclusion, the application of distributed machine learning in the medical field must first preprocess the data, select a suitable model, and utilize a privacy-protecting distributed architecture such as federated learning.

4. Federated learning for medical imaging:

Federated learning has been widely used in medical imaging, and it supports collaborative model training across institutions while maintaining data privacy. For instance, Rajendran et al. demonstrated the use of federated learning for cancer detection through a cloud-based system that allowed multiple hospitals to train shared models without transferring raw imaging data [11]. This approach not only enhances the accuracy of cancer detection models but also addresses the challenge of data ownership and legal restrictions, such as those imposed by HIPAA and GDPR. By allowing models to be trained on diverse datasets from multiple institutions, federated learning improves model robustness and generalizability, a key advantage for diagnostic applications where diverse patient populations is crucial for effective outcomes [11]. A comparisons for ROC are depicted in Fig. 1

In another study, Huang et al. applied federated learning to predict mortality and length of stay using distributed electronic medical records. Their approach involved clustering patients based on similar characteristics, which improved the efficiency of model training and led to more accurate predictions. This work exemplifies how federated learning can optimize healthcare operations by predicting outcomes based on distributed data sources without centralizing patient information, thus preserving data privacy and reducing the computational load on individual institutions [12].

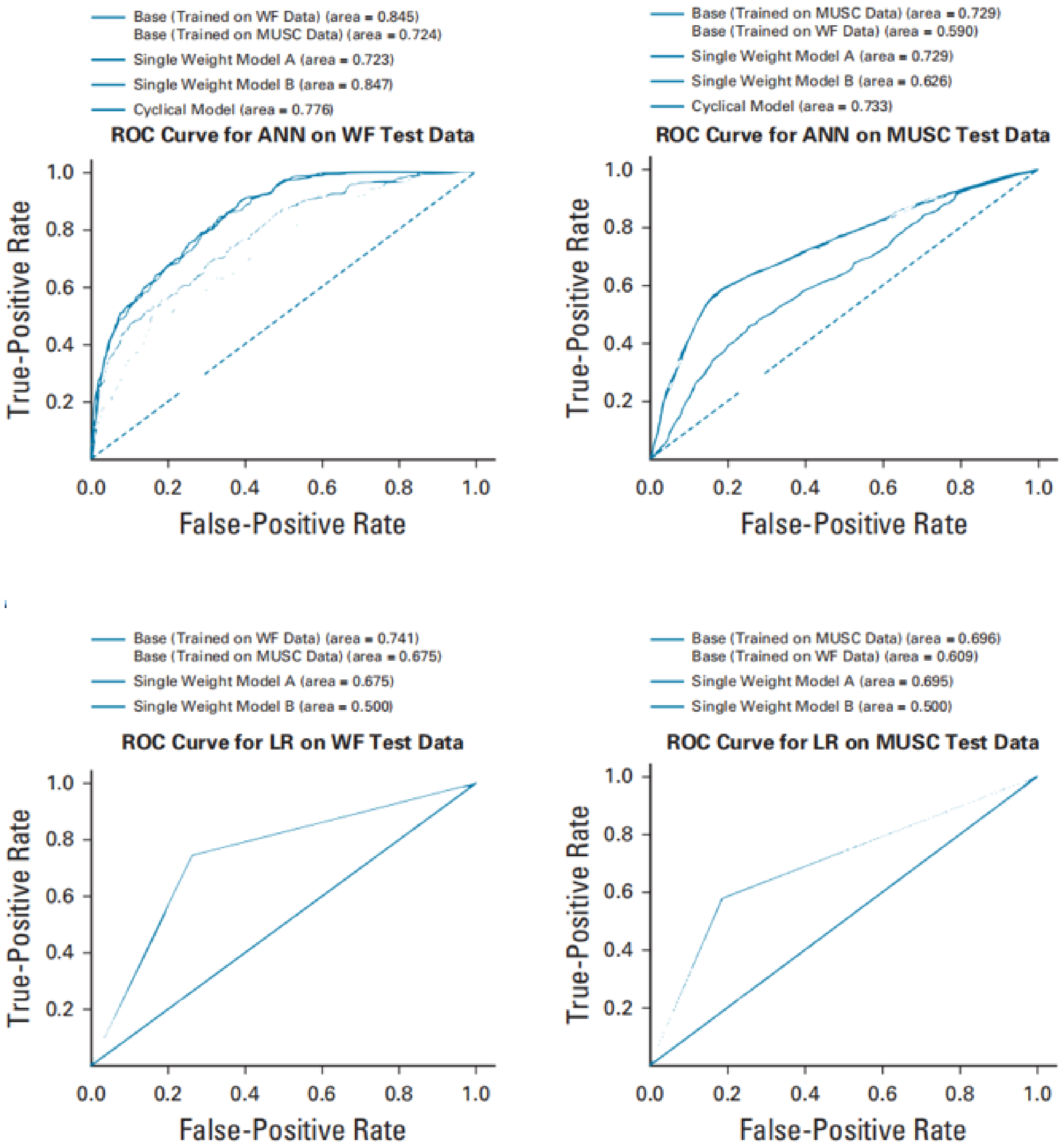

Figure 1. Typical ROC curves for comparisons [11].

5. Predictive analytics and real-time monitoring

Beyond medical imaging, distributed machine learning (DML) has played a pivotal role in predictive analytics and real-time monitoring, particularly within the Internet of Medical Things (IoMT). Deploying machine learning models across IoMT devices enables real-time health monitoring and early diagnosis of chronic diseases. Federated learning, for instance, has been applied to predict diseases like diabetes through wearable sensors and home-based devices, ensuring data privacy while providing real-time insights. This approach minimizes latency and enhances patient monitoring efficiency (MDPI) [12, 13]. A typical error analysis is given in Fig. 2.

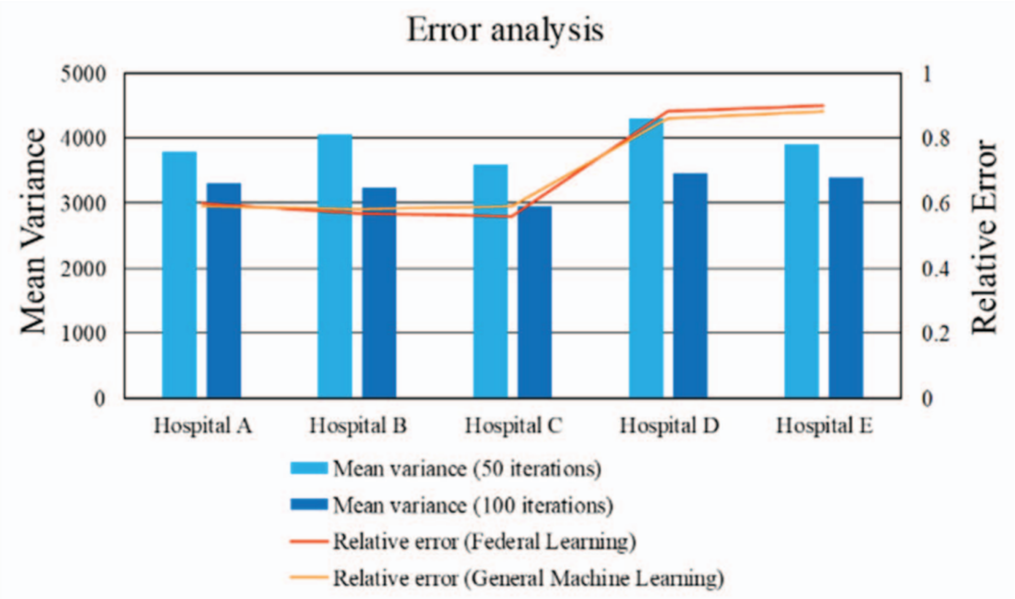

Figure 2. Typical error analysis [12].

6. Pandemic response and public health

DML has also played a crucial role in addressing global health crises such as the COVID-19 pandemic. Distributed learning techniques have been used to analyze data across regions to model the spread of the virus and predict the outcomes of interventions. Guhathakurata et al. employed machine learning models to predict COVID-19 cases and outcomes using distributed data from different countries. By leveraging a distributed framework, their models were able to provide real-time predictions without centralizing sensitive health data, a critical factor given the international nature of the pandemic and the diverse sources of data required for accurate modeling [14]. The use of distributed machine learning in public health extends beyond pandemic response. It is also used to analyze data related to vaccination campaigns, healthcare resource allocation, and patient outcomes across diverse geographic locations. These applications highlight the versatility of DML in managing large-scale, multi-institutional data for critical healthcare operations.

7. Limitations and prospects

Despite advances in distributed machine learning (DML) applications, there are still some limitations and challenges. One of the main limitations is data privacy and security concerns. Although the federated learning model addresses privacy concerns by keeping raw data locally, the exchange of model updates still poses potential risks. Adversarial attacks targeting aggregated model updates or intermediate communication phases could compromise the security of patient data. Ensuring robust security measures and secure aggregation protocols remains a key challenge [7]. Another challenge is the heterogeneity of data between different institutions. In healthcare, data collected from various sources can vary in format, quality, and presentation. This variability affects the performance of DML models because they must be generalized across different data sources. Techniques such as data normalization and standardization are employed to mitigate these issues, but achieving consistent performance across different data sets remains challenging [8]. Scalability is also a concern for DML systems, especially when dealing with large data sets or large numbers of distributed nodes. The communication overhead between nodes will increase, which will affect the efficiency and speed of model training. Developing more efficient algorithms and communication protocols to handle large-scale distributed systems is an ongoing area of research [6].

Looking ahead, the future of DML in healthcare is bright. Advances in edge computing and the Internet of Medical Things (IoMT) are expected to enhance real-time monitoring and predictive analytics. Future research could focus on improving algorithms for better scalability and efficiency, integrating advanced security measures, and developing ways to process heterogeneous data more efficiently. The continued development of DML frameworks could lead to more powerful, privacy-protecting solutions, further driving the integration of AI in healthcare.

8. Conclusion

To sum up, this study discusses the development of DML and its applications in medical imaging, predictive analysis and real-time monitoring, and introduces the basic concepts and advantages of federated learning. However, some issues remain, such as data privacy, heterogeneity, and scalability challenges. The future research will move in this direction to further develop the application of distributed machine learning in medicine. Overall, these results provide an insight for implementation of AI techniques in various industry.

References

[1] Rosenblatt F 1958 The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. Psychological Review vol 65(6) pp 386-408.

[2] Rumelhart D E, Hinton G E, Williams R J 1986 Learning Representations by Back-Propagating Errors Nature vol 323(6088) pp 533-536

[3] Cortes C and Vapnik V 1995 Support-Vector Networks Machine Learning vol 20(3) pp 273-297

[4] Sheller M J, Edwards B, Reina G A, Martin J, Pati S, Kotrotsou A and Milchenko M 2020 Federated Learning in Medicine: Facilitating Multi-Institutional Collaborations Without Sharing Patient Data Scientific Reports vol 10(1) p 12598

[5] Chen X, Ran X, Chen X and Ran X 2019 Deep learning with edge computing: A review Proceedings of the IEEE vol 107(8) p 1655-1674

[6] Dean J and Ghemawat S 2008 MapReduce: Simplified Data Processing on Large Clusters. Communications of the ACM Communications of the ACM vol 51(1) pp 107-113.

[7] Kairouz P, McMahan H B, Avent B, Bellet A, Bennis M, Bhagoji A N and Zhao S 2021 Advances and open problems in federated learning Foundations and Trends® in Machine Learning vol 14(1–2) pp 1-210

[8] Sheller M J, Reina G A, Edwards B, Martin J and Bakas S 2020Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data Scientific Reports vol 10(1) pp 1-12

[9] Singh P, Gupta R and Agarwal N 2020 Anomaly detection in medical images using convolutional neural networks Proceedings of the 2020 IEEE International Conference on Image Processing (ICIP) p 8943456

[10] Zhang H and Qie Y 2023 Applying deep learning to medical imaging: A review Applied Sciences vol 13(18) p 10521

[11] Rajendran S, Obeid J S, Binol H, Foley K, Zhang W, Austin P, Brakefield J, Gurcan M N and Topaloglu U 2021 Cloud-based federated learning implementation across medical centers JCO Clinical Cancer Informatics vol 5 pp 1-11

[12] Huang L, Shea A L, Qian H, Masurkar A, Deng H and Liu D 2019 Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records Journal of Biomedical Informatics vol 99 pp 103291

[13] Xu J, Glicksberg B S, Su C, Walker P, Bian J and Wang F 2021 Federated learning for healthcare informatics. Journal of healthcare informatics research vol 5 pp 1-19.

[14] Guhathakurata S, Saha S, Kundu S, Chakraborty A and Banerjee J S 2021 South Asian countries are less fatal concerning COVID-19: a hybrid approach using machine learning and M-AHP Computational Intelligence Techniques for combating COVID-19 pp 1–26