Indonesian Journal of Data and Science

Volume 5 Issue 1
ISSN 2715-9936
https://doi.org/10.56705/ijodas.v5i1.116

Research Article

Federated Learning for Bronchus Cancer Detection Using Tiny Machine Learning Edge Devices

Musa Dima Genemo [1,*]

[1] University of Stirling, Stirling FK9 4LA, United Kingdom, musa.ju2002@gmail.com

Correspondence should be addressed to Musa Dima Genemo; musa.ju2002@gmail.com

Received 26 February 2024; Accepted 21 March 2024; Published 31 March 2024

Abstract:

In deep learning, acquiring sufficient data is crucial for making informed decisions. However, due to concerns regarding security and privacy, obtaining enough data for training models in the era of deep learning is challenging. There is a growing need for machine learning (ML) solutions that can derive accurate conclusions from small data while preserving privacy. Smartphones, which are widely used and generate large amounts of data, can serve as an excellent source for data generation. One suitable approach for regularly evaluating real-world data from edge devices is Tiny Machine Learning (TinyML). With the increasing number of edge devices involved in transmitting private data, it's vital to have a method that allows computations to be performed on edge devices and pushed to the edge rather than over the network. Considering these obstacles, the combination of TinyML edge devices and Federated Learning can be applied in the early treatment of Bronchus Cancer. Under the framework of federated learning, local edge devices are trained independently and then integrated into the server without exchanging edge device data. This approach enables the creation of secure models without sharing information, resulting in a highly efficient solution with enhanced data security and accessibility. This article provides a comprehensive discussion of the key challenges addressed in recent literature, accompanied by an extensive examination of relevant studies. Additionally, a novel model based on edge devices and federated learning is proposed.

Keywords: Artificial Intelligence, Federated Learning, TinyML, Edge Device, Internet of Things.
Dataset link: The datasets from the current study are not publicly available due to data privacy. However, they can be obtained from the corresponding author upon reasonable request once the ownership of the data is approved.

1. Introduction

The healthcare sector has felt the impact of artificial intelligence, particularly with the advancements in deep learning. These advancements have brought about revolutionary technologies in radiology, including X-rays, ultrasounds, MRIs, and others. Deep learning is commonly used in e-healthcare applications with the goal of improving efficiency, virtual care management, and cancer diagnosis by learning from extensive personalized datasets. To ensure the security and privacy of data, we can utilize tiny machine learning and federated learning to develop a secure model without the need for sharing information.

According to the World Health Organization (WHO), cancer is the leading cause of death in developing countries [1]. Research shows that bronchus tumors are the fifth most common tumor and can be successfully treated if detected early [2]. However, detecting diseases in bronchus X-rays is a challenging task that requires the expertise of radiologists. The focus on electronic healthcare provides an opportunity to overcome this challenge by using TinyML and federated learning through edge devices. This approach will enable the training of models for researchers and medical practitioners while protecting individual privacy. Moreover, the distributed nature of the local dataset makes it suitable for federated learning using TinyML edge devices. Please see **Figure 1** for a visual representation of how data can be shared using TinyML and federated learning.

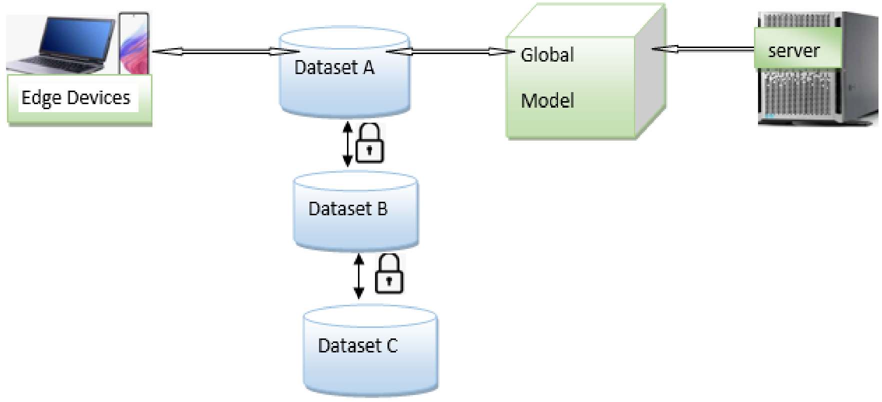

Figure 1. Federated learning using TinyML edge device

Even though many imaging centers have extensive imaging databases, many of the images are unlabeled, making it impossible for the model to learn. Every hospital needs a skilled radiologist to manually label the radiographs. As a result, creating a training dataset can take a long time, which is a key stumbling block in the development of AI for medical imaging. Federat ed learning is used to overcome this challenge by training algorithms across multiple healthcare institutions to generate superior AI models through collaboration. A given algorithm collects labeled patient data from a variety of sources to supplement its learning base and, as a result, improve its ability to detect patients across a large population. In this context, federated learning (FL) has emerged as a promising alternative for developing cost-effective smart healthcare applications that are also more private [3]–[8]. To offer the primary data and model, we employed the World Health Organization (WHO) as a central server. **Figure 2** shows how this model is constructed and how data is transferred.

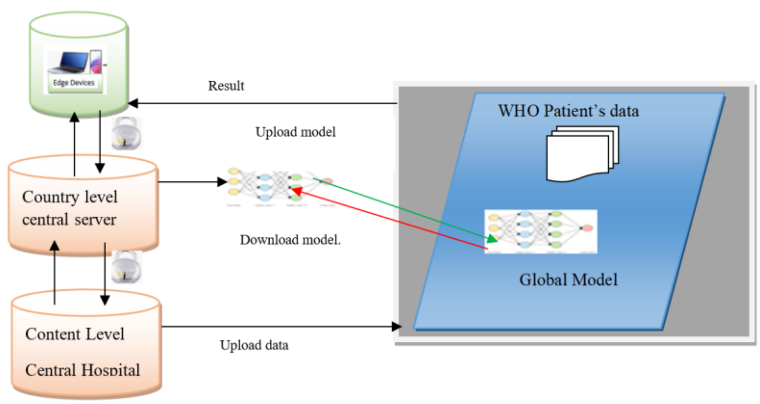

Figure 2. Proposed model for data sharing using TinyMl and Federated Learning

FL and TinyML models can update data from different health data clients, including edge devices [9]–[12], without needing direct access to the local data. This approach reduces the risk of privacy breaches by limiting the exposure of sensitive user information and preferences. Moreover, FL leverages extensive computation and dataset resources from various health data clients to train AI models, resulting in enhanced quality of health data training, including accuracy. This improvement is not achievable with centralized AI approaches that rely on limited data and computational capabilities.

2. Method:

Federated learning can be used to collect massive amounts of data from various locations around the world, remotely train models, and then import the generated results of these models for further usage through IoT and cloud computing technology. However, due to the confidentiality of hospital data, sharing data is a challenge. The proposed model aims to help hospitals adopt a common model to solve problems without sharing local data or having a common model. Another example of federated learning, which is applied in various medical situations, is the segmentation of brain tumors. This task is challenging because the desired training data may not be readily available. To classify the disease, we followed three basic procedures:

a. Infection segmentation based on local contrast of symptoms using a local contrast haze reduction (LCHR).
b. Retrieval and merging of geometric features and LBP using a canonical correlation analysis (CCA) approach. At this stage, we introduced noise in the form of irrelevant and redundant information, which is subsequently removed by the Neighborhood Component Analysis (NCA).
c. Subsequently, a multi-class SVM is used to classify the final reduced features.

The presence of noise in the image pixels can lead to incorrect labeling. To address this issue, we utilized LCHR to extract significant features for future segmentation and classification methods. This approach enables us to overcome the challenges of misclassification, improve the model's accuracy, and reduce execution time. To enhance the local contrast of the affected region in the input image, we applied equation 1 to a p × p frame size.

$$\theta = \alpha * \frac{\mu}{\sigma} \tag{1}$$

α represents constant parameter with values [0, 1] and μ is global mean, σ is the sandardeviation. The Local mean can be obtained by applying equation 2.

$$\lambda(k) = \frac{1}{m * m} \sum_{x=0}^{m-1} * \sum_{y=0}^{m-1} k(x, y) \tag{2}$$

Where k (x, y) is the original input image, The following formula can be used to measure the standard deviation:

$$\sigma = \sqrt{\frac{1}{m * m} \sum_{x=0}^{m-1} * \sum_{y=0}^{m-1} [k(x, y) - n(k)2]} \tag{3}$$

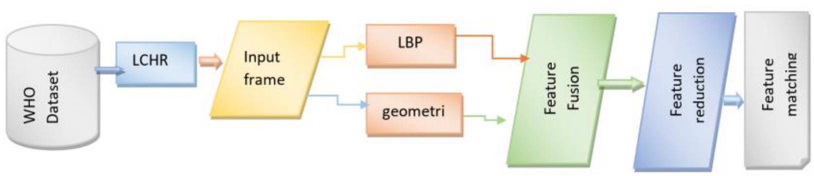

Figure 3. The proposed model for bronchus tumor classification

The complement of the input image I (p)'s contrast is generated by

$$I(p) = \theta * (p - n) + \frac{n}{\sigma}, where\ p\ \varepsilon(x, y) \tag{4}$$

Early detection of tumors is beneficial for early treatment. The current study focuses on local data, which makes it very difficult to develop a global model. As a result, an FL-based model is proposed for segmenting images into meaningful portions. For this we followed some major steps: (a), weighted function-based channel selection, (b) morphological operations refining, and (c) finally mapping and drawing ROI. The main benefit of grouping Lab features is that it allows you to check the information for each individual channel. The extracted features are mapped using cross-entropy activation function.

$$C(fi, \omega) = -\sum_{c}^{C} \beta(on, L\omega) LOG(\beta(n, \omega)) \tag{5}$$

Where probability is represented by β feature represented by f, wj represent class, represent number of observations. We reduce features into short vectors after extracting them to eliminate redundancy and provide efficient and valuable information. For lung tumor categorization, we used geometric and LBP features. Local Binary Pattern (LBP) is used to examine local texture formation to determine local patterns rather than on pixel level. LBP uses distinct simple primitives to explain complicated formations in an image [9].

$$\psi p,r = \sum_{p=0}^{p-1} s(gp - gc)2p * s(x) = \begin{cases} 1\ if\ x \geq 0 \\ 0, otherwise \end{cases} \tag{6}$$

And we used geometiric featuature to check If there is swelling, we used Geometric characteristics. This will allow us to see if the shape around the problem area has changed.

$$\rho = \sum_{i=0}^{n} * \sum\nolimits_{j=0}^{m} A[i,j], \gamma = \varphi(2l + 2w) \tag{7}$$

Where $\gamma\ is\ parameter$

To combine various retrieved features, we used feature fusion. We can cut computing costs by merging features. For feature fusion, we used canonical correlation analysis (CCA).

$$\delta = \ max \frac{\max * \ x_a^T\ \ x_{\ b}\ \ \ Cabcx_c}{\sqrt{x_a^T Caax_a\ \ \ * x_b^T Cbbx_b\ _* \ x_c^T Cccx_c}} \tag{8}$$

C_{abc} covariance matrix between feature set $C_{aa} = AA^T$, $C_{bb} = BB^T$, $C_{cc} = CC^T$ represents covariance within three feature sets. When the matrices within feature sets are nonsingular then CCA can be obtained by computing generalized Eigenproblem

3. Results and Discussion

In this section, we discuss the steps and parameters that were employed during the computation of the results. In the bronchus tumor classification, we tested the proposed model. During testing, the accuracy and error rate of the proposed model were both taken into account. Many public datasets have been used to compare the performance and effectiveness of the model. We employed various classifiers (quadratic SVM, cubic SVM, cosine KNN, and M-SVM) to validate the performance of our model. Throughout the validation phase, we found M-SVM to be effective. The proposed segmentation result has an overall accuracy of 95.95 percent. The results obtained are listed in **Table 2** to **4**.

Table 1. Dataset

Disease type	Total images	Training images	Testing images	Cross-fold-validation
Adenocarcinomas	1383	692	691	10
Squamous cell carcinomas	1076	538	538	10
Large cell carcinomas	1180	590	590	10
Bronchial carcinoids	423	212	211	10
Total	4062	2032	2030	40

The data is collected from CT scans and X-ray images. A publicly available dataset is also used for benchmarking purposes. The images are converted into a format suitable for machine learning, and features are extracted through pre-processing techniques. The diverse data collected is annotated to identify tumors, lesions, or abnormalities in the images. The dataset is then divided into training (70%) and testing sets (30%), which are crucial for training and evaluating the performance of the federated learning model.

Table 2. Classification results using CCA based features fusion.

Method	Sensitivity (%)	Specificity (%)	Precision (%)	AUC	FPR	Accuracy (%)
ESD	92.35	93.46	95.00	0.98	0.0312	92.03
Quadratic SVM	93.25	95.86	95.10	0.99	0.0312	94.22
Cubic SVM	94.55	96.14	95.12	0.99	0.0312	93.45
Cosine KNN	90.50	89.99	90.99	0.97	0.0350	92.02

Method	Sensitivity (%)	Specificity (%)	Precision (%)	AUC	FPR	Accuracy (%)
M-SVM	94.75	95.00	95.11	0.99	0.0300	95.38

Table 2 shows the overall performance of the model in detecting bronchus cancer using CCA-based feature fusion. The model achieved an accuracy level of 95.38%.

Table 3. Proposed classification after NCA based feature reduction.

Method	Sensitivity (%)	Specificity (%)	Precision (%)	AUC	FPR	Accuracy (%)
ESD	94.00	94.19	94.98	0.98	0.0312	94.99
Quadratic SVM	94.52	94.49	94.95	0.98	0.0312	93.90
Cubic SVM	93.65	93.89	94.00	0.98	0.0200	92.95
Cosine KNN	89.00	89.00	89.99	0.96	0.0399	89.89
M-SVM	95.99	96.00	96.00	0.99	0.0300	95.95

Table 3 presents the overall performance of the model utilized for bronchus cancer detection through labeling. The model achieved an accuracy rate of 95.95% by implementing NCA-based feature fusion, surpassing the accuracy achieved with CCA.

Table 4. Confusion matrix for NCA after based reduction approach

Adenocarcinomas	91	9		
Squamous cell carcinomas	8	92	1	1
Large cell carcinomas	0	1	99	1
Bronchial carcinoids	1	1	0	98
	Adenocarcinomas	Squamous cell carcinomas	Large cell carcinomas	Bronchial carcinoids

Table 4 presents the results obtained from NCA. The rows in this table correspond to the actual class labels, while the columns represent the predicted class labels. The matrix displays the counts of instances for each combination of true and predicted classes. The diagonal line running from the top left to the bottom right represents the correct predictions (true positives), while the elements outside the diagonal represent misclassifications. This confusion matrix provides a comprehensive understanding of the classification model's performance across different classes.

4. Conclusion

The lack of diagnostic tools in developing countries has a significant impact on people's lives. Therefore, it is critical to diagnose illnesses early and provide accessible and affordable treatments. We have developed an automated approach for detecting lung tumors using saliency assessment and prominent feature selection. Various features are retrieved, and a CCA-based fusion strategy is used to select the best features using the NCA method. Compared to existing approaches, we used an M-SVM classifier for classification and achieved an accuracy of 95.95 percent. We also calculated the computing cost of the proposed system, and the cost was minimized after the selection procedure. However, the suggested approach has several drawbacks, such as reduced accuracy for complex images. We plan to implement a more robust system in the future.

References

[1] L. Liu, Y. Y. Tsui, and M. Mandal, “Skin lesion segmentation using deep learning with auxiliary task,” *Journal of Imaging*. mdpi.com, 2021, doi: 10.3390/jimaging7040067.

[2] P. Rajpurkar *et al.*, “CheXNet: radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv archivePrefix.” 2017, doi: 10.48550/arXiv.1711.05225.

[3] S. Warnat-Herresthal, H. Schultze, K. L. Shastry, and ..., “Swarm learning for decentralized and confidential clinical machine learning,” *Nature*. nature.com, 2021, doi: 10.1038/s41586-021-03583-3.

[4] M. J. Sheller, B. Edwards, G. A. Reina, J. Martin, S. Pati, and ..., “Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data,” *Scientific reports*. nature.com, 2020, doi: 10.1038/s41598-020-69250-1.

[5] G. Kaissis, A. Ziller, J. Passerat-Palmbach, and ..., “End-to-end privacy preserving deep learning on multi-institutional medical imaging,” *Nature Machine …*. nature.com, 2021, doi: 10.1038/s42256-021-00337-8.

[6] G. A. Kaissis, M. R. Makowski, D. Rückert, and ..., “Secure, privacy-preserving and federated machine learning in medical imaging,” *Nature Machine …*. nature.com, 2020, doi: 10.1038/s42256-020-0186-1.

[7] D. C. Nguyen *et al.*, “Federated Learning for Industrial Internet of Things in Future Industries,” *IEEE Wirel. Commun.*, vol. 28, no. 6, pp. 192–199, Dec. 2021, doi: 10.1109/MWC.001.2100102.

[8] M. D. Genemo, “Suspicious activity recognition for monitoring cheating in exams,” *Proc. Indian Natl. Sci. Acad.*, vol. 88, no. 1, pp. 1–10, Mar. 2022, doi: 10.1007/s43538-022-00069-2.

[9] Q. Yang, Y. Liu, T. Chen, and Y. Tong, “Federated machine learning: Concept and applications,” *ACM Trans. Intell. …*, 2019, doi: 10.1145/3298981.

[10] D. C. Nguyen, M. Ding, P. N. Pathirana, A. Seneviratne, J. Li, and H. Vincent Poor, “Federated Learning for Internet of Things: A Comprehensive Survey,” *IEEE Commun. Surv. Tutorials*, vol. 23, no. 3, pp. 1622–1658, 2021, doi: 10.1109/COMST.2021.3075439.

[11] V. Mothukuri, R. M. Parizi, S. Pouriyeh, Y. Huang, A. Dehghantanha, and G. Srivastava, “A survey on security and privacy of federated learning,” *Futur. Gener. Comput. Syst.*, vol. 115, pp. 619–640, Feb. 2021, doi: 10.1016/j.future.2020.10.007.

[12] G. Shrestha, Deepsikha, M. Das, and N. Dey, “Plant Disease Detection Using CNN,” in *2020 IEEE Applied Signal Processing Conference (ASPCON)*, Oct. 2020, pp. 109–113, doi: 10.1109/ASPCON49795.2020.9276722.