Evaluasi Statistik Performa Algoritma Federated Learning pada Klasifikasi Tumor Otak = Statistical Evaluation of Federated Learning Algorithm Performance in Brain Tumor Classification

Raisha Zhafira, author

Abstrak

Berdasarkan Global Cancer Observatory (GLOBOCAN), tumor otak diestimasikan berada di urutan ke-19 sebagai tumor yang paling umum terjadi dan ke-12 sebagai penyebab utama kematian akibat kanker di dunia pada tahun 2020. Walaupun begitu, informasi terkait epidemiologi tumor otak di Indonesia masih sangat terbatas. Belum diwajibkannya pendataan kasus tumor di Indonesia merupakan salah satu alasannya. Tumor otak dapat dideteksi menggunakan pencitraan medis, seperti computed tomography (CT) scan dan magnetic resonance imaging (MRI). Deteksi dini tumor otak merupakan hal yang penting karena dapat meningkatkan tingkat keberlangsungan hidup dari pasien. Saat ini, banyak perkembangan teknologi yang dapat dimanfaatkan untuk membantu kehidupan manusia, salah satunya adalah deep learning (DL). Akan tetapi, data medis merupakan data yang sensitif, sehingga menjadi salah satu tantangan dalam penerapan DL di bidang kesehatan. Untuk mengatasi privasi dan keterbatasan data, terdapat metode federated learning (FL) yang memungkinkan untuk dilakukannya pelatihan data lokal pada klien tanpa menyebarkan data klien tersebut. Pada penelitian ini, akan dibentuk simulasi klasifikasi tumor otak menggunakan DL berbasis FL. Tujuan utama dari penelitian ini adalah untuk menganalisis performa model yang dihasilkan dari federated learning dan membandingkannya dengan metode training konvensional. Terdapat empat cycle data dengan tiga cycle berasal dari dataset pertama (M. Nickparvar) dan satu cycle dari dataset kedua (J. Cheng). Hasil akurasi dan F1-score tertinggi dari simulasi federated didapatkan pada epoch (jumlah putaran pelatihan data pada tiap klien) 15 dan round (jumlah putaran mulai dari tahap pembagian parameter model global kepada klien sampai dengan agregasi model) 15, yaitu 0.8375 dan 0.8384 (cycle 1), 0.7625 dan 0.7567 (cycle 2), 0.8375 dan 0.8308 (cycle 3), serta 0.7333 dan 0.7255 (dataset 2). Hasil akurasi dan F1-score tertinggi dari simulasi standard pelatihan lokal pada tiap cycle adalah 0.75 dan 0.7568 (cycle 1), 0.6875 dan 0.6677 (cycle 2), 0.675 dan 0.6744 (cycle 3), serta 0.7222 dan 0.7085 (dataset 2). Hasil akurasi dan F1-score tertinggi dari simulasi standard pelatihan all data pada tiap cycle adalah 0.7625 dan 0.7644 (cycle 1), 0.6875 dan 0.6723 (cycle 2), 0.775 dan 0.7766 (cycle 3), serta 0.6 dan 0.5355 (dataset 2). Berdasarkan pengujian hasil simulasi, korelasi epoch dan round terhadap performa model signifikan pada dataset kedua (Pacc-epoch = 0.019; PF1-epoch = 0.006; Pacc-round = 0.008; PF1-round = 0.025), tetapi hanya korelasi round yang signifikan pada dataset pertama (cycle 1 Pacc-round < 0.001 dan PF1-round < 0.001; cycle 2 Pacc-round $= 0.004$ dan PF1-round $= 0.003$; cycle 3 Pacc-round < 0.001 dan PF1-round < 0.001). Selain itu, performa model global hasil federated learning lebih baik daripada performa model lokal dan model pelatihan standard. Tidak ditemukan perbedaan signifikan antara performa model dataset pertama dengan cycle yang berbeda (Pacc between cycles $= 0.679$; PF1 between cycles $= 0.770$) serta tidak ditemukan juga perbedaan signifikan antara performa model dataset pertama dan kedua (Pacc cycle 1-dataset 2 = 0.103; Pacc cycle 2-dataset 2 = 0.334; Pacc cycle 3-dataset 2 $= 0.103$; PF1 cycle 1-dataset 2 $= 0.140$; PF1 cycle 2-dataset 2 = 0.120; Pacc cycle 3-dataset 2 $= 0.140$).

.....Based on the Global Cancer Observatory (GLOBOCAN), brain tumors are estimated to rank 19th among

the most common tumors and 12th as the leading cause of cancer-related deaths worldwide in 2020. However, information regarding the epidemiology of brain tumors in Indonesia remains very limited. One reason is that case registration for tumors is not yet mandatory in Indonesia. Brain tumors can be detected using medical imaging, such as computed tomography (CT) scan and magnetic resonance imaging (MRI). Early detection of brain tumors is crucial as it can improve the survival rates of patients. Currently, many technological advancements can be utilized to aid human life, one of which is deep learning (DL). However, medical data is sensitive, presenting a challenge in applying DL in healthcare. To address privacy and data limitation problems, there is a method called federated learning (FL) that enables local data training on clients without sharing the clients' data. This study aims to simulate brain tumor classification using DL based on FL. The main objective of this research is to analyze the performance of the model generated from federated learning and compare it with conventional training methods. There are four data cycles, with three cycles from the first dataset (M. Nickparvar) and one cycle from the second dataset (J. Cheng). The highest accuracy and F1-score from the federated simulation were achieved at epoch (number of training rounds on each client) 15 and round (number of rounds starting from the global model parameter distribution to the clients until model aggregation) 15, which are 0.8375 and 0.8384 (cycle 1), 0.7625 and 0.7567 (cycle 2), 0.8375 and 0.8308 (cycle 3), and 0.7333 and 0.7255 (dataset 2). The highest accuracy and F1-score from the standard local training simulation in each cycle are 0.75 and 0.7568 (cycle 1), 0.6875 and 0.6677 (cycle 2), 0.675 and 0.6744 (cycle 3), and 0.7222 and 0.7085 (dataset 2). The highest accuracy and F1-score from the standard all data training simulation in each cycle are 0.7625 and 0.7644 (cycle 1), 0.6875 and 0.6723 (cycle 2), 0.775 and 0.7766 (cycle 3), and 0.6 and 0.5355 (dataset 2). Based on the simulation, the correlation between epoch and round on model performance is significant in the second dataset (Pacc-epoch = 0.019; PF1-epoch = 0.006; Pacc-round = 0.008; PF1-round = 0.025), but only the round correlation is significant in the first dataset (cycle 1 Pacc-round < 0.001 and PF1-round < 0.001; cycle 2 Pacc-round $= 0.004$ and PF1-round $= 0.003$; cycle 3 Pacc-round < 0.001 and PF1-round < 0.001). Moreover, the performance of the global model resulting from federated learning is better than the local model performance and standard all data training model performance. No significant difference was found between the performance of the first dataset with different cycles (Pacc between cycles $= 0.679$; PF1 between cycles = 0.770) nor between the performance of the first and second datasets (Pacc cycle 1-dataset 2 = 0.103; Pacc cycle 2-dataset 2 = 0.334; Pacc cycle 3-dataset 2 = 0.103; PF1 cycle 1-dataset 2 $= 0.140$; PF1 cycle 2-dataset 2 = 0.120; Pacc cycle 3-dataset 2 = 0.140).