Breast Cancer Detection Using Federal Learning

Aishik Biswas

Abstract

Machine learning has now been the technology that is changing the world around in a good way. Humans makes mistakes to learn from them but those mistakes can be life threatening and can cause potentially heavy damage. Especially in the field of Medical Science. Due to wrong diagnosis a patient can die or can lead to heavy damage which can be beyond repair for the lifetime and be incurable. As we all know that machine learning requires huge datasets for the accuracy of the model which is very important and cannot be neglected when it comes to proper diagnosis. For an example a normal tumor can be miss diagnosed as a malignant tumor and treated according leading to huge damage to the patient as well as the hospital and the doctors who has diagnosed it. However due to the concern of privacy its not possible to access all the data present to train the model. Without the access of sufficient data ML models cannot be accurate enough to be used in the industry. Thats where Federated Learning comes in play removing all the concerns of privacy and consent. The project shows how federated learning can be the future of medical science as well as digital health.

1 Introduction

Predictive deep learning algorithms show promise in improving medical diagnosis and treatment, but they need a lot of data to work well. Deep learning models fared badly on data from institutions whose data were not viewed during training, according to a recent study1. Deep learning medical imaging models were explicitly mentioned as relying on confounding variables related with institutional biases rather than predicting apparent disease. When evaluated against held-out data from the same institution, such models may produce high accuracy, but they may not generalise well to other institutions or even between departments within the same institution. Collaborative learning, in which multi-institutional data is utilised to train a single model, is a logical method to expand both data amount and variety.

In the medical arena, the current paradigm for multi-institutional partnerships necessitates the sharing of patient data to a centralised site for model training. Various medical areas, such as radiography, pathology, and genetics, have their own repository. For example, training an AI-based tumour detector necessitates a vast database that covers the complete range of anatomies, pathologies, and input data formats. Because health data is very sensitive and its use is closely regulated6, data like this is difficult to get. Even if data anonymisation could circumvent these restrictions, it is now widely accepted that deleting information such as a patient's name or date of birth is frequently insufficient to protect privacy. For example, data from computed tomography (CT) or magnetic resonance imaging (MRI) can be used to rebuild a patient's face8. Another reason for the lack of systematic data sharing in healthcare is that gathering, curating, and maintaining a high-quality data set requires a lot of time, effort, and money. As a result, such data sets may have great commercial value, making them less likely to be freely shared. Data collectors, on the other hand, frequently maintain fine-grained control over the information they have obtained.

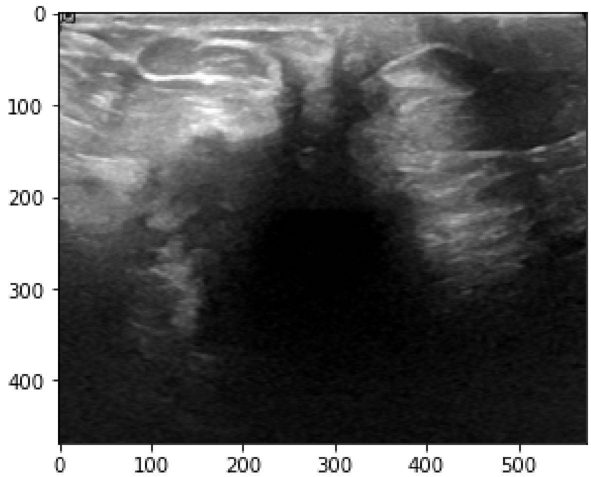
Federated learning (FL) is a learning paradigm that aims to solve the challenges of data governance and privacy by collectively training algorithms without transferring data. It was originally designed for a variety of domains, including mobile and edge device use cases12, but it has lately acquired popularity in healthcare applications. FL allows for collaborative insights, such as in the form of a consensus model, without transferring patient data outside of the institutions' firewalls. A successful application of FL might thus have a huge impact on large-scale precision medicine, resulting in models that make impartial judgments, best reflect an individual's physiology, and are sensitive to uncommon illnesses while still respecting governance and privacy issues. However, FL still needs careful technological thought to guarantee that the algorithm runs well without jeopardising patient safety or privacy. Nonetheless, it has the ability to overcome the drawbacks of methods that rely on a single centralised data pool.

Therefore in this project we have tried to implement Federated Learning to Detect Breast Cancer from several scan datasets of breast cancers and normal healthy breast. In the dataset of Breast Tumor Scans, we have taken data from two classes namely "malignant" and "normal" for cancer detection. Therefore we have applied a simple Neural Network with Multi level perceptron model (MLP) and applied our federal learning in which there is a global model and multiple local models from which the weight get updated and sent to the global model to get trained for the next iteration.

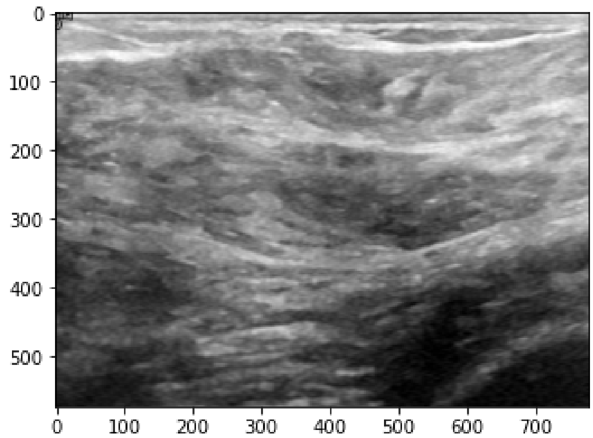
2 Dataset

The Dataset that we have used is one of the datasets that we have found in kaggle "https://www.kaggle.com/datasets/anaselmasry/-datasetbusiwithgt". The datasets in already divided into train and test split. It has test data total of 150 cases from which 84 are labelled as "benign", 42 as "malignant", 24 as "normal". In the training set we have total of 630 casesfrom which 353 labelled as "benign", 168 labelled as "malignant", 109 as "normal". We have excluded out the "benign" labelled caes and went for purely breast cancer detection rather than breast tumor detection.

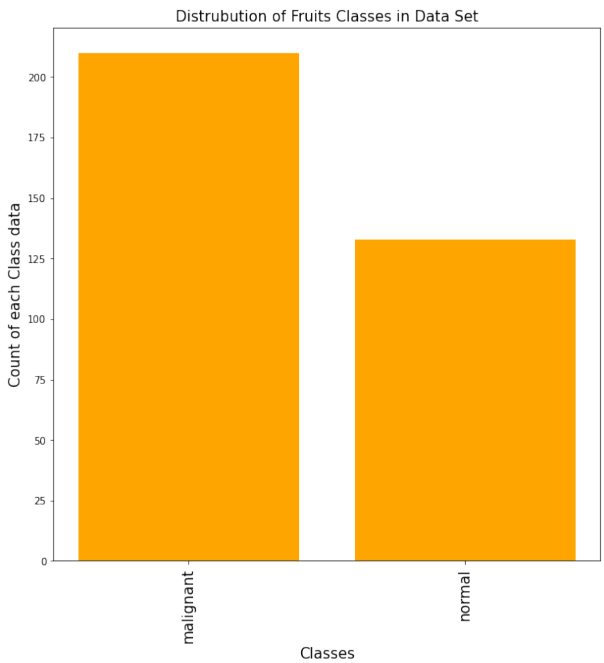
The dataset have 5 types of hearbeats (classes): Normal (N), R-on-T Premature Ventricular Contraction (R-on-T PVC), Premature Ventricular Contraction (PVC), Supra-ventricular Premature or Ectopic Beat (SP or EB), Unclassified Beat (UB).



Malignant Tumor



Malignant Tumor



Distribution graph

3 Appraoch

3.1 Preparing the data

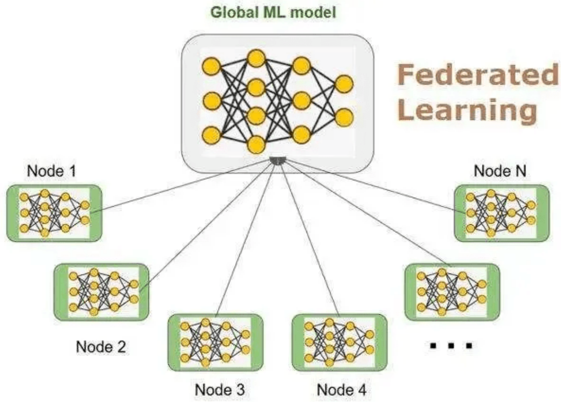
The we dnt have big data therefore we have thick data. Total cases that we have including the training and validation is 343 cases. We save the data class names and count how many case that we have. Therefore we plot a histogram to get visual representation of how many cases are we dealing with and the classes corresponding to the cases. Then we concat and store the data in train and test split with a validation split of 0.3 where 70 percent is the training data and 30 percent is the rest of the testing data. We resize the images into 256 x 256 with bilinear inter polation.

3.2 Creating clients

Creating a list of clients where the xtrain and ytrain datasets are passed with number of number to be created is 100. We are creating 100 local clients from which the weight of the model will be passed on to the global model. The clients will have randomised data and splits. Each client will be shared data accordingly and train. The data shards are then seperated out into data and label lists.

3.3 Starting federated learning

A simple ANN model is created with muli level percepton and have 400 hidden layers. We have Stochastic Gradient Descent (SGD) as out optimiser ata learning rate of 0.01. This is the global model. An instance of the global model is created. So to show the federated learning we have created a simple ANN model. Then we commence a global training loop of 300 iterations. At the first iteration the weight of the global model is served as the weight for all the local models. During each iteration we loop through each client and create a local model and train the data on that local model. Once the model is trained we send the weight of the local model to the global model from there the next local model gets the weight. Before every operation we scale the weights in between 0-1 to reduce the computational expense. Doing this we get and collect all the weights of the local model then average out the weights of all the local models and then fed to the global model. This is done 300 times. Then we calculate the accuracy of the global model at each iteration and Categorical Cross Entropy loss.



Federated learning architecture

4 Results

After training the model over 300 iterations and federated learning process of locally and globally training the model but only taking the weights of of the trained model, we got a very good accuracy of global accuracy being 91.119 and global loss being 1.59965.

```
comm_round: 261 | global_acc: 88.214% | global_loss: 1.6277244091033936
comm_round: 262 | global_acc: 85.619% | global_loss: 1.6424388885498047
comm_round: 263 | global_acc: 86.881% | global_loss: 1.6318743228912354
comm_round: 264 | global_acc: 88.548% | global_loss: 1.625700831413269
comm_round: 265 | global_acc: 91.571% | global_loss: 1.596787452697754
comm_round: 266 | global_acc: 89.952% | global_loss: 1.608684778213501
comm_round: 267 | global_acc: 88.214% | global_loss: 1.6285958290100098
comm_round: 268 | global_acc: 90.238% | global_loss: 1.609269618988037
comm_round: 269 | global_acc: 86.500% | global_loss: 1.6376686096191406
comm_round: 270 | global_acc: 90.452% | global_loss: 1.6039811372756958
comm_round: 271 | global_acc: 90.286% | global_loss: 1.6033092737197876
comm_round: 272 | global_acc: 90.714% | global_loss: 1.6026108264923096
comm_round: 273 | global_acc: 89.643% | global_loss: 1.6113004684448242
comm_round: 274 | global_acc: 90.929% | global_loss: 1.6083978414535522
comm_round: 275 | global_acc: 89.524% | global_loss: 1.6150033473968506
comm_round: 276 | global_acc: 86.286% | global_loss: 1.641101360321045
comm_round: 277 | global_acc: 88.714% | global_loss: 1.6173404455184937
comm_round: 278 | global_acc: 89.000% | global_loss: 1.6218324899673462
comm_round: 279 | global_acc: 90.548% | global_loss: 1.6015233993530273
comm_round: 280 | global_acc: 91.381% | global_loss: 1.5945498943328857
comm_round: 281 | global_acc: 90.952% | global_loss: 1.6001704931259155
comm_round: 282 | global_acc: 89.738% | global_loss: 1.6162526607513428
comm_round: 283 | global_acc: 87.500% | global_loss: 1.6324450969696045
comm_round: 284 | global_acc: 91.929% | global_loss: 1.5894263982772827
comm_round: 285 | global_acc: 90.738% | global_loss: 1.5999544858932495
comm_round: 286 | global_acc: 87.857% | global_loss: 1.625375747680664
comm_round: 287 | global_acc: 87.429% | global_loss: 1.6339106559753418
comm_round: 288 | global_acc: 91.571% | global_loss: 1.5947693586349487
comm_round: 289 | global_acc: 91.095% | global_loss: 1.596139907836914
comm_round: 290 | global_acc: 89.905% | global_loss: 1.6045095920562744
comm_round: 291 | global_acc: 90.881% | global_loss: 1.6024580001831055
comm_round: 292 | global_acc: 91.524% | global_loss: 1.605469822883606
comm_round: 293 | global_acc: 90.262% | global_loss: 1.6065346002578735
comm_round: 294 | global_acc: 89.095% | global_loss: 1.625815749168396
comm_round: 295 | global_acc: 91.429% | global_loss: 1.5952544212341309
comm_round: 296 | global_acc: 87.690% | global_loss: 1.6267505884170532
comm_round: 297 | global_acc: 90.024% | global_loss: 1.604494571685791
comm_round: 298 | global_acc: 89.833% | global_loss: 1.607370138168335
comm_round: 299 | global_acc: 91.119% | global_loss: 1.5996589660644531
```

after 300 iterations

5 Discussion and conclusion

In this paper, I used different approaches to classify Normal and Abnormal heartbeats. LSTM autoencoder successfully detected all the numbers of Normal or Anomaly heart beats. I also showed how through calculating reconstruction loss the problem can be solved by Binary classification using

Threshold value. Also I showed how KNN performed with feature scaling and without SMOTE for the dataset. And the performance of SVM before and after performing Oversampling.

References

[1] Sheller, M.J., Edwards, B., Reina, G.A. et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Sci Rep 10, 12598 (2020). https://doi.org/10.1038/s41598-020-69250-1

[2] Rieke, N., Hancox, J., Li, W. et al. The future of digital health with federated learning. npj Digit. Med. 3, 119 (2020). https://doi.org/10.1038/s41746-020-00323-1

[3] Asynchronous Federated Learning-based ECG Analysis for Arrhythmia Detection Sadman Sakib, Mostafa M. Fouda†, Zubair Md Fadlullah, Khalid Abualsaud, Elias Yaacoub , and Mohsen Guizani

[4] Learnings from Federated Learning in the Real world Christophe Dupuy, Tanya G. Roosta, Leo Long, Clement Chung, Rahul Gupta, Salman Avestimehr

[5] J. Cao, M. L. Brown, E. J. Higgins, and P. J. Degroot, "Automatic thresholds for atrial tachyarrhythmia detection in an implantable medical device," ed: Google Patents, 2019.