Federated Ensemble Learning for Healthcare Applications

Enmanuel Aquino, ThanhVi Dang, Jeremy Martínez, Ujjwal Samanta

I. Background Overview

It is well known that in the latest decades, technology has been pushing the boundaries of how we approach and look for solutions to the day-to-day issues we face as part of our lives, Data Science and Machine Learning models have been spearheading these improvements. Every day, significant progress has been made regarding how well these models can perform; this idea holds for the backbone of human development: health.

Improvement in the Healthcare Applications framework has been vital to pivot to any other area of knowledge. Naturally, ensuring the well-being and longevity of our people is directly proportional to the well-being, safekeeping, and good use of the data available to us in the form of medical histories and unmined knowledge.

Just a couple of decades before [1], medical reports worldwide continued to be taken physically without any form of digitalization or constructive data form. Even these days, many countries, which are mainly part of societies in ways of development, still lack the benefit of Electronic Medical Records or EMR as we know them. Naturally, we want to argue about the importance of Machine Learning methods in the domain of Healthcare. In that case, we first need to understand the importance of the data being available for processing, hence, why the year 1960 was pivotal in this aspect. That is when Larry Weed conceived the first record idea regarding electronic patient information recording methods.

Jump a couple of decades forward, and we find ourselves in a highly developed society with more information than it can process. The complexity and rise of needs in the healthcare industry grow exponentially, and so does the usage of Artificial Intelligence (AI) and Data Mining to achieve clinical decision support, solid electronic health record systems, improve the hardware involved in the diagnosis and corrective processes of health, and much more.

The problem lies in a straightforward concept engraved in the usage of Machine Learning models, even if there have been outlooks to improve through the years constantly, these models still struggle to offer sound and precise results when the data used to train them is very sparse, noisy, contains many outliers or in general, is hard to analyze in a way that makes patterns somewhat simple to detect. When we encounter high-dimensional or very skewed data, the traditional already known Machine Learning models need help finding the underlying patterns that result in the usable and essential insights that we look for.

Naturally, something as complex as human nature is, by extension, studying for patterns and insight discovering in the health domain turns out to be very complicated as well. Some of the topics from the early days of science until today still capture the interest of scientists around the world and are related to healthcare, for instance, [2] Genomic Sequences, Predictive Analytics in Healthcare for diseases as relevant as Cancer, Sequence Tracking for Viral diseases like Zika or Dengue and much more.

Without a doubt, there is a need for good data discovery, classification, mining, and prediction. We are already at a point where much data is available for use, but how can we solve the need for more precise and reliable Machine Learning Models to pair them with something as important as health confidently? That is where “Ensemble Learning” comes into play. We will be looking into this topic further in the research process.

Lastly, a vital aspect of this entire problem that needs to be addressed is related to the ethics behind healthcare. Much of the information that is involved in the diagnosis and corrective processes of the healthcare domain is obscured or protected due to medical patient safety. Naturally, psychological and integrity safety is one of the big challenges that Data Mining faces in the context of health, luckily, a solution for that and another significant component of this research project is in ways of development, and that is the “Federated” aspect of learning, which eliminated the concept of centralized data [3] and creates opportunities to collaboratively build the data sources for model training without jeopardizing the safety of all the individual pieces of information involved.

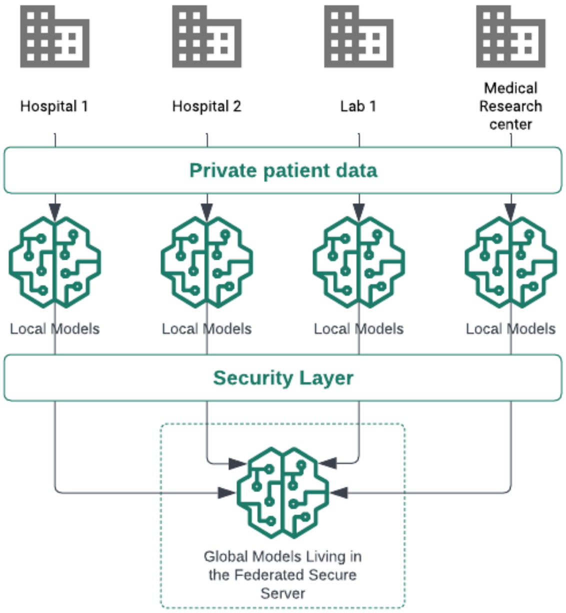

Figure 1 (Authentic): A Federated Ensemble Learning High-Level Diagram

To briefly provide some context regarding Figure 1, we can see how in that framework, we have several sources of data related to the healthcare domain, such as hospitals, labs, and medical research centers, all of these could have their local model running of the information they have at hand, in the context of ensemble learning, they would have several models all of which ultimately feed into a single global model with improved accuracy.

II. Topic Overview

Before getting into details regarding the comparative study, and based on the fact that Federated Learning is a relatively new field of study, we set to

A. *Ensemble Learning*

It consists of utilizing several models running in parallel or sequentially over the same dataset with the primary objective of taking advantage of each of their virtues against the data being used. Covering more ground in this aspect allows for much more flexibility when working with complex data, as we mentioned before.

A local model for Ensemble Learning for Healthcare applications could look like this:

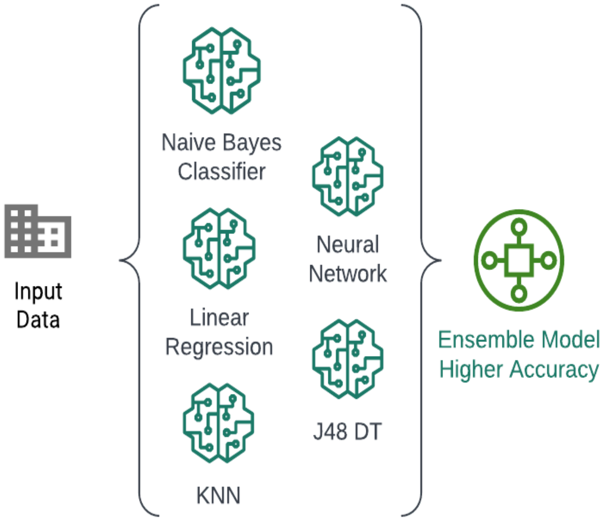

Figure 2 (Authentic): Ensemble Model Visualization

We can think of this figure as how each of the "Local Models" in **Figure 1** would look internally based on the premise of our Federated Framework having each one of its units working based on an Ensemble Learning model like this one.

B. *Federated Learning*

Comes into play when we need to train a model based on information that originates from scattered sources, and we want to achieve that without exposing the information to any sort of breach or threat, once again, due to the nature of medical patient information, this is a great alternative to consume different data sources while not breaching any security norm. Refer to **Figure 1** for a visual representation.

[4] Federated Learning can be executed based on different paradigms, but using a Central Orchestrator set-up is one of the most effective and used alternatives, this is how the framework behaves algorithmically:

- **Central server** chooses a model to be trained.
- **Central server** sends the initial model to all the **scattered nodes**.
- All the **particular nodes** go over their own training in a **local** environment.
- Results from local training are uploaded to the **Central server,** and a response is generated without accessing the local private data.

In the same way a Neuronal Network does, a model based on Federated Ensemble learning also has many hyper-parameters for tweaking and optimization. Some of these can be the number of federated rounds or the local learning rate for each particular model. We can think of Federated Learning as a network, reason why they have these key similarities with a Neural Network.

C. *Federated Learning and Healthcare*

Decentralized learning is the dominant idea behind FL, where the users' sensitive data is never sent outside of their local institutions. The standard machine learning (ML) approaches require centralizing the training data into a common store, meaning that all data is collected and modeled in one centralized location, creating an extremely easy step to manipulate data analysis. However, there are many downsides to this centralized setup. Due to network latency or connection loss, battery duration, and other unforeseen problems, this back-and-forth communication might negatively affect the user experience. [10, 18, 19]. The solution to this is federated learning, where we utilize the goodness of decentralized data on individual devices by training data locally, only sending to the centralized server the model parameters instead of the training data.

As various research studies have emphasized, federated learning is a notable paradigm due to its capability to protect and preserve data privacy [10, 11, 12, 15, 17, 19, 21]. However, note that FL itself does not guarantee data privacy. It is a model that *reduces* privacy concerns by collaboratively extracting knowledge and learning complex model while storing personal data in their own perspective sources. Another caveat to FL is its vulnerability in producing overfitting results and its high variance performance [10]. Although the model is not perfect, it provides great compromises in other aspects.

FL was initially developed by Google to aggregate dispersed intelligence without jeopardizing the data privacy and security of its users. It is gaining popularity as a result of the convergence of emerging new technology and applications. They proposed this algorithm to be a viable method for the natural language processing task. This task allows Google to implement support for their virtual keyboard that supports more than 600 language varieties [13, 17]. Functions such as next-word prediction, auto-correction, suggested words, etc. often contain sensitive information that FL has promising applications for [13, 17].

Federated learning has been proven to be successful in many fields due to its many benefits in comparison to traditional ML approaches. FL embraces data security, data diversity, real-time continual learning, as well as hardware efficiency [15, 19, 23]. The applications that are most relevant in FL are smart devices [13, 17], organizations, IoT [23], healthcare [14, 21, 22], and many more. Researchers are continuing to see success in this mechanism because it sets forth many intrinsic advantages in real-world applications.

Most importantly, FL holds significant potential for connecting dispersed healthcare data sources while protecting patients' privacy. As the field of biomedical data analytics surges, FL becomes a successful tool in this field to preserve confidential health records while catering to the patients' personalized care. It also promotes scientific research on large datasets that aims to improve standardized medical records as it is notoriously known for being fragmented and inaccessible due to the nature of the electronic health records (EHR) system [15, 22] and rigorous regulations of HIPAA [24]. This collaborative modeling mechanism FL provides enables multiple parties involves in the healthcare industry (insurance companies, providers, patients, hospital institutions, pharmaceutical companies, etc.) to deal with large-scale data representation to hopefully one day build joint data access globally for better patient care.

Although the traditional ML methodology (especially deep learning) is becoming more and more promising in building clinical-grade accuracy results and robust statistical models for medical data, the concerns of the data repository and privacy remain a restriction, therefore, a limitation to fully encompass its values. The assurance of FL is straightforward: to enable ML from non-co-located data, by addressing issues with privacy and data governance. Each data controller in an FL context establishes its own governance procedures and privacy guidelines, as well as managing data access, and has the authority to withdraw it. This creates many more new opportunities in healthcare informatics: novel research on rare diseases, avoiding duplications of records from multiple institutions, growing a global dataset to capture larger data variability, new business avenues, etc [19]. In the context of EHR, FL helps represent patient similarity learning [16], leveraging patient data communications between different providers [22], including predictive modelings such as cardiac events [11], mortality [14], ICU stay-length [14], readmission [18], and more. For clinician usage, FL aids in ensuring consistency of diagnosis, diagnosis segmentation, and leveraging patient data communications between different providers [19]. It also has been proven the advantages in the field of medical imaging [15] and phenotyping [25, 26]. For the patients, FL could ensure high-quality clinical decisions and personalization of care [15, 19]. FL also recognizes and eliminates biases or discrimination in medical practices [27]. For medical manufacturers and researchers, FL can help develop software and hardware that data collect and combine without revealing patient-specific information

FL has been demonstrated to be multifaceted in the field of healthcare Itis a very promising paradigm for creating models that are strong, unbiased, accurate, safe, resilient, and secure. FL elegantly addresses issues related to the sensitivity of medical data by allowing the various parties to collaboratively train together without having to exchange or centralize data sets. Seeing from previous research, FL has had an impact on nearly all the stakeholders in the medical process. Although there are many more technical questions related to FL that are unanswered, we can still see the tremendous prospective of improving precision in medicine and patient medical care.

III. Comparative Overview

As stated in section II, although both methodologies aim to provide a predictive model to respond to a question, key differences must be analyzed. Indeed, the algorithms may have some similarities, but there is a clear difference in terms of the initial approach between ensemble and federated learning. Federated learning runs multiple algorithms with different databases dispersed from local servers to a central server, while ensemble learning uses one dataset throughout its different techniques.

Analyzing the effectiveness of each methodology, even taking out how FL tremendously impacts data privacy in the health sector, there is no convenience in comparing algorithms of federated learning versus ensemble learning. Therefore, the next two subsections will compare algorithms within the same methodology.

IV. Comparative Random Forest vs AdaBoost

Recalling ensemble learning, it is when there is more than one classifier or regressor to predict an outcome. Before comparing the algorithms, there is an important concept in ensemble learning that seeks to evaluate the effectiveness of the models, which is a bias-variance tradeoff. This concept helps draw light on the prediction error of a predictive model. **Figure 3** illustrates that there is a tradeoff between variance and bias. Depending on the technique used, it will help minimize variance or bias. The challenge is to look for the method that throws the best results toward the optimal point.

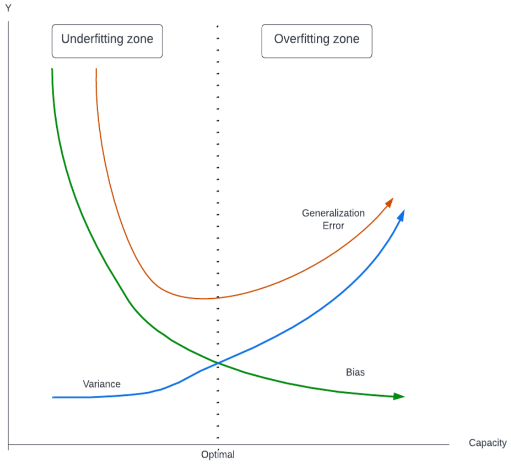

Figure 3 (Non-authentic) [8]: Bias-Variance Tradeoff

Random Forest is one of the algorithms that helps minimize variance. It consists of two main concepts, such as decision trees and parallel running. **Figure 4** shows more than one decision tree, which depends on numerous independent variables trying to predict the desired output. As a result, it uses the same algorithm with different variations running in parallel over the initial dataset. The final step will combine the result of all the decision trees and use a criterion of voting or averaging the most frequent prediction.

Beyond the core concept of random forest, other characteristics are equally important in using a technique that minimizes variance. First, the way the sampling of a data set is used via sampling with replacement. This way of sampling will grant the opportunity for the same observations to be introduced more than once in the same decision tree. Therefore, the atypical observations of the dataset will have less probability of being selected, resulting in a reduction in variance and generalization error, **Figure 3**. Second, optimization of parameters. In practice, three parameters affect the accuracy of the model [7]. Depending on the usage of these parameters, it will improve or not the final output; here are the parameters for tuning:

- Number of random independent variables
- Amount of decision trees
- Tree size, in terms of the splitting decisions.

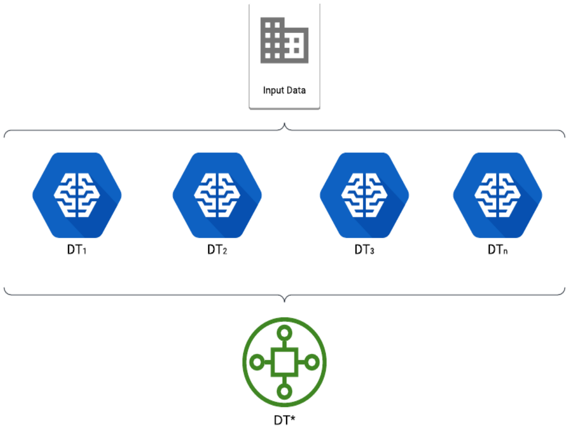

Figure 4 (Authentic): Random Forest Visualization

In contrast, **AdaBoost** represents one of the algorithms inside the boosting methods family that minimizes bias. AdaBoost runs sequentially over the initial dataset in comparison to Random Forest. The core concept of this algorithm, it tries to reduce the errors made by the previous predictors, passing from weak to strong predictors, **Figure 5**. In the first step, it trains a default classifier to return a prediction, for example, decision trees. After calculating the training errors, it concentrates on the misclassified observations by assigning them a higher weight. When it is updated, it trains the model to deliver better predictions.

Therefore, this method minimizes the bias because it tries sequentially adjusting the errors made by the previous predictors, thus reducing the difference between the real value and the predicted value. As well this algorithm has several parameters to optimize for better prediction, such as the learning rate and the number of predictors. The second parameter is more sensitive due to the fact that there is a tradeoff in limiting the number of rounds. A higher number of predictors will cause overfitting because the algorithm trains the model so well in the training phase that having a different set of test data might cause an error to increase.

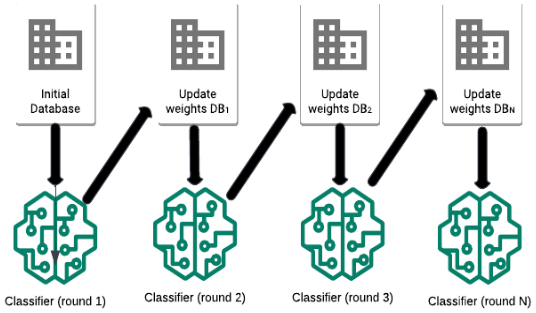

Figure 5 (Authentic): AdaBoost Visualization

V. Comparative FederatedAveraging vs Fed-Ensemble

In principle, the Federated Learning concept is to run models in a decentralized manner via individual clients with their local datasets. As a result, they update their results to the global model in the central server without ever uploading the local datasets. **Federated Averaging** is an algorithm that uses the principles of stochastic gradient descent for each client, and its final step makes an average of the predictions. Each individual database is trained using gradient descent, which tries to minimize the error of the predictions. Once the individual results are available, the central server takes a weighted average.

Any Federated Learning technique must take precautions for the computational cost that will be required in the run time. In the case of Federated Averaging, it takes into consideration three parameters [9]:

- Percentage of the individual dataset is going to be used per round.
- The number of training passes each client does per round.
- Local minibatch size used to update.

Despite the usefulness of Federate Learning, there is an improved technique inside the FL methods called **Fed-Ensemble**. This method uses the core concept of Federated Averaging. The difference relies on how iteratively the ensemble models used behind the scenes through random permutations are updated to improve the generalization error and reduce variance. The user can establish the number of ensemble models, a predetermined parameter, to be used in the selected rounds. In addition, throughout the iterations, the models selected can learn from the dataset, hence improving the overall accuracy.

References

[1] EMR: 100% electronic medical records: Scranton online. The University of Scranton. (2022, May 12). Retrieved December 4, 2022, from https://elearning.scranton.edu/resources/article/emr-the-progress-to-100-percent-electronic-medical-records/#:~:text=The%20History%20of%20EHR's&amp;text=Also%20in%20the%201960's%2C%20the,developed%20by%20the%20Regenstrief%20Institute.

[2] Team, D. F. (2021, April 2). Data Science in healthcare - 7 applications no one will tell you. DataFlair. Retrieved December 4, 2022, from https://data-flair.training/blogs/data-science-in-healthcare/

[3] Thursday, A. 06, &amp; Learningoptimization, C. M. L. O.-device. (n.d.). Federated learning: Collaborative machine learning without centralized training data. – Google AI Blog. Retrieved December 4, 2022, from https://ai.googleblog.com/2017/04/federated-learning-collaborative.html

[4] Wikimedia Foundation. (2022, December 2). Federated learning. Wikipedia. Retrieved December 4, 2022, from https://en.wikipedia.org/wiki/Federated_learning

[5] Tan, Pang-Ning. (2006). Introduction to Data Mining. Pearson Education

[6] Geron, Aurelien. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. 2nd Edition.

[7] Cutler, Adele. (2011). Random Forests. https://www.researchgate.net/publication/236952762_Random_Forests

[8] Fortmann-Roe, Scott. (2012). Understanding the Bias-Variance Tradeoff. https://courses.washington.edu/me333afe/Bias_Variance_Tradeoff.pdf

[9] McMahan, Brendan. (2017). Communication-Efficient Learning of Deep Networks from Decentralized Data. https://www.semanticscholar.org/paper/Communication-Efficient-Learning-of-Deep-Networks-McMahan-Moore/d1dbf643447405984eeef098b1b320dee0b3b8a7

[10] Shi, Naichen. (2021). Fed-ensemble: Improving Generalization through Model Ensembling in Federated Learning. https://arxiv.org/pdf/2107.10663.pd

[11] Brisimi TS, Chen R, Mela T, Olshevsky A, Paschalidis IC, Shi W (2018) Federated learning of predictive models from federated electronic health records. Int J Med Inform 112:59–67

[12] Gong, X., Sharma, A., Karanam, S., Wu, Z., Chen, T., Doermann, D., & Innanje, A. (2022). Preserving Privacy in Federated Learning with Ensemble Cross-Domain Knowledge Distillation.

[13] Hard, A., Rao, K., Mathews, R., Ramaswamy, S., Beaufays, F., Augenstein, S., ... & Ramage, D. (2018). Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*.

[14] Huang, L. et al. Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records. *J. Biomed. Inform.* 99, 103291 (2019).

[15] Kaissis, G.A., Makowski, M.R., Rückert, D. *et al.* Secure, privacy-preserving and federated machine learning in medical imaging. *Nat Mach Intell* 2, 305–311 (2020). https://doi.org/10.1038/s42256-020-0186-1

[16] Lee, J., Sun, J., Wang, F., Wang, S., Jun, C. H., & Jiang, X. (2018). Privacy-preserving patient similarity learning in a federated environment: development and analysis. *JMIR medical informatics*, *6*(2), e7744

[17] McMahan, HB., Ramage, D., Talwar, K., & Zhang, L. (2017). Learning differentially private recurrent language models. *ArXiv e-prints*, arXiv-1710.

[18] Min X, Yu B, Wang F (2019) Predictive modeling of the hospital readmission risk from patients' claims data using machine learning: A case study on copd. Sci Rep 9(1):2362

[19] 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

[20] Roy, A. G., Siddiqui, S., Pölsterl, S., Navab, N. & Wachinger, C. Braintorrent: a peer-to-peer environment for decentralized federated learning. *arXiv preprint arXiv:1905.06731* (2019).

[21] Sharma, P., Shamout, F. E., & Clifton, D. A. (2019). Preserving patient privacy while training a predictive model of in-hospital mortality. *arXiv preprint arXiv:1912.00354*.

[22] Xu, J., Glicksberg, B.S., Su, C. et al. Federated Learning for Healthcare Informatics. J Healthc Inform Res 5, 1–19 (2021). https://doi.org/10.1007/s41666-020-00082-4

[23] Zhou, J., Zhang, S., Lu, Q., Dai, W., Chen, M., Liu, X., ... & Herrera-Viedma, E. (2021). A survey on federated learning and its applications for accelerating industrial internet of things. *arXiv preprint arXiv:2104.10501*.

[24] Gostin, L. O. (2001). National health information privacy: regulations under the Health Insurance Portability and Accountability Act. *Jama*, *285*(23), 3015-3021.

[25] Kim, Y., Sun, J., Yu, H., & Jiang, X. (2017, August). Federated tensor factorization for computational phenotyping. In *Proceedings of the 23rd ACM SIGKDD International conference on knowledge discovery and data mining* (pp. 887-895).

[26] Liu, D., Dligach, D., & Miller, T. (2019, August). Two-stage federated phenotyping and patient representation learning. In *Proceedings of the conference. Association for Computational Linguistics. Meeting* (Vol. 2019, p. 283). NIH Public Access.

[27] Obermeyer, Z., & Mullainathan, S. (2019, January). Dissecting racial bias in an algorithm that guides health decisions for 70 million people. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 89-89).