Predicting Child Development Delays Using Federated Learning

Geetika Gopi
Carnegie Mellon University

David Mberingabo
Carnegie Mellon University

Yunyi Wang
Carnegie Mellon University

Kay Chen
Carnegie Mellon University

Abstract

The early detection of child development delays is critical to ensure early support and appropriate treatment. In this study, we investigate the use of federated learning to predict child development delays in a privacy preserving way. We obtain a dataset with carefully selected features from a previous study by Borovsky et al [13]. To perform model weight aggregation across multiple rounds of federated computations, we had to use a sequential neural network instead of the Random Forest model used in the study. In conclusion, we compare the performance of our federated learning model against a centralized sequential model. The results of this study indicate that federated learning achieves a performance accuracy similar to that of a centralized learning model. The slightly higher loss for federated learning may be attributed to communication overheads and differences in the weight updates provided by the local client models. Overall, our findings suggest that TensorFlow Federated can be a useful tool for theoretically training a child development delays prediction model in a privacy preserving manner.

1 Introduction

The early detection of child development delays is critical to ensure early support and appropriate treatment. To narrow down our study scope, we particularly look at a common condition called Developmental Language Disorder (DLD). DLD is associated with significant language difficulties in children, such as language learning and use, with extremely limited supporting data to identify the root causes. DLD is a condition that affects around seven percent of children and may have a large impact in their academics, social development, mental health and physical well-being [12], [22]. Despite its prevalence and impact, DLD remains under-diagnosed and under-treated, with many children going unrecognized and unsupported [30].

Past research has indicated that early identification and treatment can significantly help children to improve their outcomes [25]. This suggests the need for prediction tools that can accurately predict potential DLD symptoms. In the recent years, there has been multiple attempts to detect child development delays using machine learning algorithms [12].

It is important to realize that data collected related to DLD diagnostics is highly sensitive in nature. Hence, it is crucial to ensure that any data that is collected for detection purposes should be protected and the child's privacy is protected. Furthermore, even when sensitive health data is provided with consent, there are several ways in which it can potentially compromise children's' privacy. Some of them include, data breaches, re-identification attacks, discrimination and the use of collected data for secondary purposes without explicit consent.

The purpose of this study is to develop a privacy-preserving machine learning model that can predict early indications of developmental language disorder. A full scale implementation would involve creating an application for parents to record their child's development and deviation from the average. A smart listening device such as Alexa or Google Home could be used to collect language data as well as parents own reporting. Instead of a full implementation, our study focuses on developing a machine learning model in a simulated federated learning environment. [26] The federated learning approach will allows multiple devices or parties to collaborate in model training iterations, while keeping the data local, avoiding the need for centralizing sensitive information [26]. The goal of our study is to evaluate the performance of a model trained in a federated learning environment to that of a model trained in a centralized environment. Federated learning is a relatively new solution with limited real-world implementation. Our study illuminates some of the possibilities of federated learning and demonstrates how to simulate, tune and evaluate a federated learning model, starting with a centralized model [26]. This is the first effort made to apply federated learning to predict DLD conditions.

1.1 Problem Statement

Based on description in section 1, we formulate our research question as follows:

RQ: How effective is federated learning as compared to centralized learning when predicting child development delays based on real world data?

1.2 Privacy Risks

Solove and Citron's privacy harms taxonomy [15] is a framework for categorizing the various harms that can result from the collection, use, and dissemination of personal information. The framework provides a useful tool for analyzing and understanding the many different types of privacy harm that can result from the use of personal information, and it helps to inform the development of privacy policies and laws.

Connecting the proposed study to Prof. Solove and Citron's work [15], our research aims to address the following privacy harms:

- **Surveillance** - The use of federated learning means that data doesn't have to be sent to a centralized server, reducing the need for continuous monitoring of a child's activity to compute a development score. Instead, only the model weights would be monitored, which greatly reduces the amount of raw personal data collection.
- **Interrogation** - Our proposed solution would ensure that medical professionals/centralized systems do not have to request sensitive information from parents. It also avoids potential parental discomfort and concerns about sharing information about their child.
- **Identification** - Since the data about children stays locally on the parents' device, parents do not have to worry about their children being linked to sensitive information, such as their developmental statistics. We would like to note that although federated learning provides far better protection against identification than common anonymization, it has been shown to be vulnerable to differential attacks. [10]
- **Secondary use** - Since the data about children stays locally on the device, parents do not have to worry about their child's sensitive information being used for purposes other than intended.
- **Intrusion** - Data about children's development is highly sensitive and can cause discomfort for parents if they believe it is being monitored. Our solution protects both parents and children from this harm by keeping the data locally on the device, reducing the need for continuous monitoring and ensuring a higher level of privacy protection than common anonymization.

2 Related Work

2.1 Understanding and Diagnosis of Developmental Language Disorder

Developmental Language Disorder (DLD) is a communication disorder that interferes with learning, understanding, and using language. Children with DLD typically have normal intelligence and hearing, but experience significant challenges with language development [9]. They may struggle to keep up with their friends and experience difficulty forming relationships with others. DLD is one of the most common developmental disorders, approximately 1 in 14 children in kindergarten are affected by this disorder [9]. This is a neurodevelopmental disorder that will first appear in childhood and continues to as the child get older. Most cases of DLD can be diagnosed after a child enters school but the roots of this disorder are present in their infancy. Treatment for this disorder is best done at an early age and early detection is therefor crucial [9].

Several factors are associated with DLD, including psychosocial factors. The article "Understanding developmental language disorder - the Helsinki longitudinal SLI study (HelSLI): a study protocol" written by Laasonen et al. [21], highlights the role of mother-child interaction quality in moderating the effects of biological disadvantage on cognitive functioning, as well as the impact of temperament on language development, both of which have been underexplored in the context of DLD [21]. Another challenge in diagnosing DLD comes from typically developing children in bilingual households, the lack of knowledge, normative data, and tools can lead to over or under-diagnosis of the condition [21]. Hence, researchers suggest that healthcare professionals combine DLD and bilinguals as they resemble each other in multiple aspects [21]. The mix of these 2 disorders can also lead to inefficiency in identifying DLD in children at an early age. Another method used to detect DLD is through Continuous electroencephalogram (EEG) recording [21]. However, a longitudinal study failed to identify any significant associations between abnormal EEG and later language development. It remains unclear if EEG abnormalities are present in all children with DLD or only in specific subgroups, and if EEG has predictive value for DLD in a longitudinal setting [21]. Researchers also find that DLD is often influenced by genetic factors, with at least three genetic loci and two genes expressed in the brain suggested to contribute to DLD with low confidence. The article [21] well discusses the gaps of understanding and the need for further research and development in the field of Developmental Language Disorders.

2.2 Use of Machine Learning in Healthcare

Over the last few years, machine learning has been increasingly in use for healthcare applications [11]. This is majorly

due to the ability to quickly extract useful insights from large datasets at scale. The recent surge in the use of machine learning for healthcare includes applications like disease diagnosis, health monitoring, drug discovery and personalized treatments [11]. One of the most important application of ML to healthcare is disease diagnosis [18]. Medical researches have long been developing machine learning models to effectively detect diseases such cancer, Alzheimer's disease and heart related conditions.

Other than disease diagnosis, machine learning has been increasingly used for health monitoring, continuous monitoring and analysis of health-related data from individuals [16]. Machine learning is hence an effective way to detect early stage symptoms associated with various healthcare conditions [16].

2.3 Use of Federated Learning in Healthcare

Federated learning is a shared global model that is trained under the coordination of a central server, from a federation of participating edge devices [20]. The model under consideration leverages a decentralized approach, whereby the training data is securely stored on local devices owned by end-users. These devices are employed as computation nodes to perform data operations locally and in parallel. By doing so, a global model is updated without requiring the data to be centralized. Such a decentralized framework offers enhanced privacy and security guarantees, while facilitating scalability and efficiency of model training [20].

The privacy characteristics of federated learning make it suitable for some sensitive data, such as healthcare data. Hospitals could use the FL model to predict the likelihood of patients developing an illness or getting infected [17]. It can also help research diseases because it allows people to create a large dataset and allow researchers who might not have access to certain sensitive data to still access models trained using that sensitive data, while respecting the privacy of the research participants [17]. This can make participants more willing to participate in studies and allow ease of collaboration between researchers, as there is fewer risks if the model was build in a federated learning environment.

The methods used to diagnose DLD are very limited. Due to the complex nature of DLD, there is very little attention from the researchers. DLD is an unknown disorder compared to others such as autism, and ADHD. Most people do not know about this disorder, so parents often do not realize something wrong until a later stage or may never realize it [24]. The first method would be to educate parents about language development and disorder, and empower them with detection tools. Untrained people may confuse DLD with intellectual disability, so it's important for parents to understand the differences and identify this disorder for their children so it can be improved in the early stage [24]. There are no accurate methods to identify DLD, most of the diagnoses depend on parents' or teachers' reflection but when they realize something went wrong it will be when the children can speak and it may be a little bit late. So if DLD can be predicted much earlier, this disorder can be treated much better.

Our research mainly focused on Borovsky's article "Moving towards accurate and early prediction of language delay with network science and machine learning approaches", they think by using powerful computational tools people can improve early DLD risk assessment via parental report of early communication skills [13]. The author argued that early identification of language delay is crucial for effective intervention and treatment and that current diagnostic methods have limitations in their accuracy and reliability. In this study, Borovsky and her group use Random Forest modeling to do the analysis. RF algorithms develop highly accurate classification solutions while requiring only minimal assumptions about the structure of the data, and are robust against over-fitting of selected training subsets [13]. By using this method, Borovsky, and her group will be the first ones to predict future language outcomes. The article's conclusion suggests that network science and machine learning approaches have the potential to significantly improve early prediction and diagnosis of language delay [13].

3 Dataset Description

Our current implementation had to be modified from the original plan proposed in the project proposal. This was because the dataset that was originally selected for this study (the By-Child Summary data from WordBank [19]) had extremely low feature importance to detect the outcome label (`typically_developing`). This was confirmed by training a basic Gaussian Naive Bayes model and observing the confusion matrix. The results indicated an extremely strong correlation between the input feature `health_conditions_encoded` and the outcome label `typically_developing` (see figure 1). This would mean that the low feature importance would negatively impact our model's predictions. Hence, we had to identify a high quality dataset with satisfactory feature importance for our study experiment.

The dataset for our study was provided by the authors of our reference paper, Borovsky et al. [13] The MBCDI-derived dataset [1] is a combination of the Early Identification of Risk for Language Impairment (EIRLI) [28] and Language Acquisition and Semantic Relations (LASER) datasets [13]. Both datasets measure early vocabulary skills and demographic variables along with outcome measures of later language/reading delays.

The combined dataset consists of 14 features including network measures of semantic structure in each child's early productive lexicon, overall vocabulary skill, grammatical ability, and demographic measures. The outcome column of our dataset will be `AnyLangorReadDxOnly`. This label indicates whether a child has a DLD disorder or not. If the child does

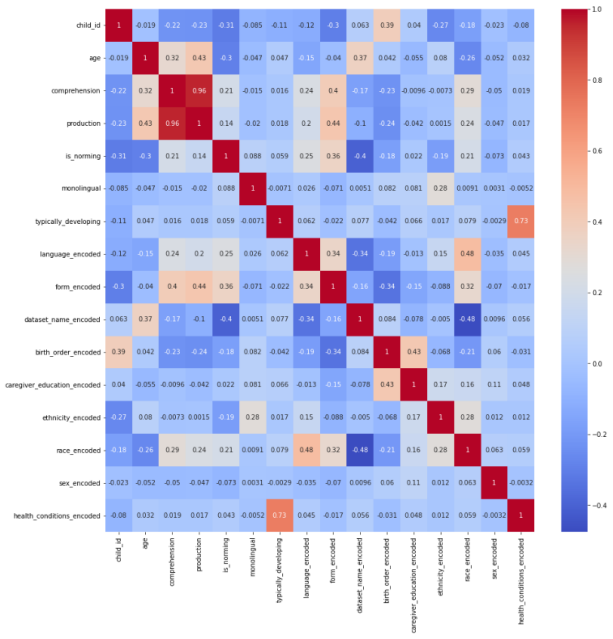

Figure 1: Confusion matrix for the WorkBank dataset model

have a disorder the outcome will be recorded as `Dx`, else will be recorded as `NoDx`. The goal of our machine learning model would be to use the 14 input features to predict the `AnyLangorReadDxOnly` column. Each of these measures and datasets are discussed in detail in the following sections.

3.1 EIRLI Dataset

The Early Identification of Risk for Language Impairment (EIRLI) dataset is a subset of data aimed at exploring the early development of gesture, vocabulary, and language skills in early childhood, and provides a valuable opportunity to investigate the early identification of risk factors for language impairment and reading disorders in children [13]. The dataset includes demographic information and MBCDI data for 391 children who were followed up with their families once a year between 4 and 7 years of age [13].

Parents reported any diagnosis of a language or reading disorder, which was verified via documentation from school records or a clinician [13]. Children were marked as having a positive language/reading disorder outcome status if they were reported and verified to have any language or reading issues at any of these ages. The analysis creating this dataset focused on the subset of children who had item-coded MBCDI data at 16 and/or 28 months, of which 17 (4.3%) were later reported to have a school-age language/reading disorder. Specifically, 4.3% of the children with item-coded MBCDI data at 16 and/or 28 months were later reported to have a school-age language/reading disorder, indicating that these early indicators may be useful in identifying children who may benefit from intervention services [13].

3.2 LASER Datset

The Language Acquisition and Semantic Relations (LASER) dataset includes data from 85 children who were longitudinally followed between 18 and 36 months of age, and evaluated for language delay using the CELF-P2 at 36 months [13]. To facilitate comparability with the EIRLI dataset, similar demographic and vocabulary variables were selected for analysis, including vocabulary data at 18 and 27 months, which aligns with the EIRLI dataset's time points of 16 and 28 months. The LASER dataset's main outcome variable is the CELF-P2, which identifies children with language delay based on standard scores less than 85. Out of the 85 children, 12 (14%) met this criterion for language delay. These children may represent a subgroup at higher risk for persistent language difficulties, and understanding the factors associated with language delay in this group could have important implications for early identification and intervention [13].

3.3 MBCDI Derived Measures

3.3.1 Description of network measures

In order to measure children vocabulary, five network connectivity measures are used.

- **Mean Path Length(MPL)**: MPL is computed as the average distance between all connected pairs of nodes in a network, with the distance representing the shortest number of "hops" between any two-word nodes via semantically-overlapping feature links. Shorter MPLs indicate more efficient networks that can transmit information more easily and quickly between nodes. A shorter MPL would enable faster and more effective retrieval of associated words in a network of related words. [13]

- **Global Clustering Coefficient(GCC)**: GCC is a measure of the overall connectivity of the nouns in a toddler's vocabulary network. The GCC is calculated as the total number of words that are connected in closed triples divided by the total number of connected triples. A connected triple is any set of three words (nodes) that share semantic links, while closed triples include cases where all three nodes are interconnected. GCC values range from 0 to 1, where a value of 0 indicates that the child's lexicon has no connected triples, and a value of 1 indicates that all triplets are closed. Toddlers with a higher GCC have higher semantic connectivity of their vocabulary, with a higher proportion of vocabulary in the semantic word "neighborhoods" compared to children with a lower GCC. [13]

- **The degree(K)**: the degree of the vocabulary item is defined as the number of other nodes in the vocabulary network that share feature overlap with that item. The mean degree (MD) is calculated as the average degree of

all items in each toddler's vocabulary network. A higher MD indicates that a child's vocabulary consists of more words with direct semantic neighbors than a child with a lower MD. Therefore, a higher MD is indicative of a richer vocabulary and a greater semantic network. [13]

- **The Mean Betweeness Centrality(BC)**: BC assesses the degree of centrality of a given vocabulary item within the semantic network by quantifying the number of times that item appears on the shortest path between all pairs of nodes in the network. Specifically, BC reflects the extent to which a word lies on the "between" path connecting other word pairs in the network. The Mean BC, on the other hand, refers to the average betweenness centrality score of all nodes in the network. By examining this metric, we can gain insight into the overall network structure and identify which words serve as central hubs of connectivity. [13]

- **Mean Harmonic Centrality(HC)**: HC is a network measure that captures the degree of connectivity of nodes in a network. Specifically, HC is a variant of closeness centrality that takes into account unconnected nodes in a network, a common feature in early vocabulary development. HC is calculated as the inverse of the sum of all distances of a node between all other nodes, divided by the number of words minus one. This measure ranges between 0 (indicating a lack of connection to other words) and 1 (a complete connection to all other words). Mean harmonic centrality is calculated as the average harmonic centrality of all words in a child's network, providing a comprehensive estimate of how closely each word is connected to other words in the child's vocabulary. [13]

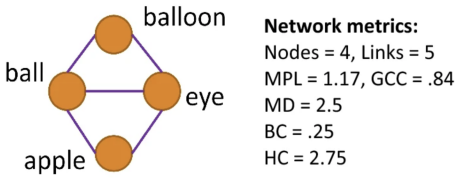

Figure 2: An example of noun-feature network with four nodes (words) and five links

There are other measures in the dataset such as gender, ethnicity, median income of the parents, and family history of DLD, that also significantly contributes to DLD diagnosis in children.

3.4 Feature Importance

To validate the quality of the dataset, mean feature importance and ranking was performed. This was done to ensure that input features of the dataset have a satisfactory correlation with the outcome label, `AnyLangorReadDxOnly`. As seen in Figure 3, the MCBDI-derived features (including betweenness, degree, harmonic centrality, path length and GCC) and vocabulary percentile are the top seven features highly correlated to the outcome label, `AnyLangorReadDxOnly`.

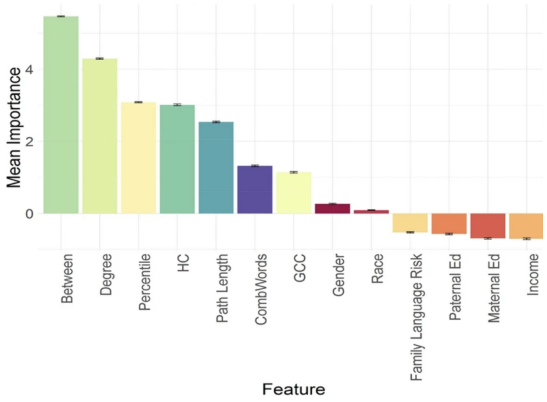

Figure 3: Feature importance in the dataset

4 Methodology

To compare the performance of federated learning against centralized learning, we created a simple sequential neural network model. A sequential neural network is a type of neural network architecture that is designed to process sequential data, such as time series or natural language [29]. It is composed of a sequence of layers, where each layer receives input from the previous layer and produces output for the next layer. The layers are usually arranged in a feedforward manner, with the input layer at the beginning and the output layer at the end [29]. The most commonly used type of sequential neural network is the recurrent neural network (RNN), which is designed to handle sequences of arbitrary length by using a feedback loop that allows the network to maintain an internal state that depends on the previous inputs. Sequential neural networks offer a promising performance for healthcare for disease diagnosis, because of it's ability to process sequential data [23]. By analysing data patterns, sequential neural networks can aid with disease diagnosis and prediction of disease progression [23].

To ensure a stable baseline for performance comparison, the sequential neural network model was created using TensorFlow's Keras API [2] and federated learning was implemented using TensorFlow Federated [4]. The experimental setup is described in detail in the following sections.

5 Implementation

5.1 Code and Dataset Availablity

Due to non-disclosure agreement with the authors of the paper [13], we are unable to publish the datasets for public use. However, the code to our implementation is available at https://github.com/drakstik/Federated_Learning_CDI.

5.2 Centralized Learning Model

The centralized learning model was created using TensorFlow's Keras API [2]. TensorFlow's Keras is a popular API used for building high-level neural networks. Keras is a popular choice for developing machine learning models because of it's developer friendly nature and customization options. The API offers variety of built in layers, activation functions and optimization algorithms.

First, the datasets (EIRLI and LASER) were merged, cleaned and pre-processed. To identify each record a unique client ID was generated using the `uuid` library [8] in python. During the pre-processing stage, all the `NaN` values in the dataset (`GramComplex` column), was imputed with mean averages. This was done because most of the machine learning algorithms cannot handle missing values. Imputing missing data with mean averages is a simple and effective way to impute missing values, without significantly altering the original data distribution. Finally, the binary categorical outcomes `NoDx` and `Dx` was replaced with numerical values (0 and 1), to make it more suitable for binary classification algorithms.

The dataset was then divided into training and validation sets using a 80/20 split ratio. These sets were identified using `X_train`, `X_test`, `y_train`, `y_test`, where `X` represents the input features and `y` represents the outcome label. The neural network was created using TensorFlow Keras Sequential() function [5].

The neural network (see 4) is made up of three fully connected layers. The first two layers are made up of 64 neurons and uses the ReLU activation function. The ReLU activation function is a common choice for neural networks because of its computational efficiency. When the input to a ReLU activation function is greater than zero, the output is equal to the input [3]. When the input is less than or equal to zero, the output is zero. This results in a threshold effect where the neuron only "fires" if the input is above a certain threshold, which can help the network learn more sparse representations of the data [3]. The last layer uses the sigmoid activation function [27], where the output is a probability between 0 and 1 (binary classification).

Model compilation was performed using `adam optimizer` and `binary_crossentropy` loss function. Adam optimizer, which is an extension of stochastic gradient descent (SGD), is an optimization algorithm that computes individual adaptive learning rates for different parameters from estimates of

```
Model: "sequential_1"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense_3 (Dense)             (None, 64)                960

 dense_4 (Dense)             (None, 32)                2080

 dense_5 (Dense)             (None, 1)                 33

=================================================================
Total params: 3,073
Trainable params: 3,073
Non-trainable params: 0
_________________________________________________________________
```

Figure 4: Model summary of the Sequential Neural Network

first and second moments of the gradients. Because of it's ability to converge faster and handle sparse gradients, adam's optimizer is a popular choice for training sequential neural networks [14]. The binary cross entropy loss function measures the difference between the predicted and the true labels [6]. The loss function heavily penalizes when the model when an incorrect prediction is made, thus making it extremely effective in training sequential neural networks. Finally, the model is trained with a batch size of 32 and 10 ephochs.

5.3 Federated Learning Model

We used the TensorFlow Federated (TFF) library to simulate a federated learning environment. TFF is an open-source framework for running federated computations using federated data in a local simulation of a decentralized system. It allows researchers to run federated algorithms on their models and data using a high-level API, while allowing for granular customization with a lower-level interface. TFF does not provide real world scenarios by simulating real clients, instead it spawns local servers which are called clients and a central server which is called the central server.

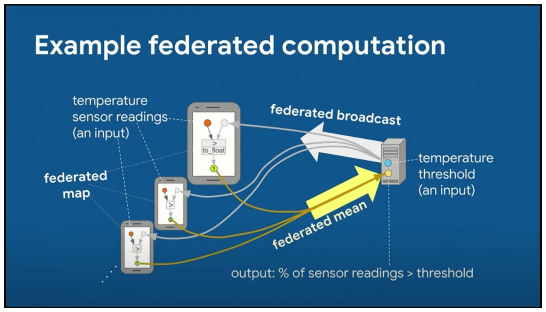

Figure 5: A TFF Computation, as demonstrated by GoogleTalk's tutorial session using TFF. [7]

To simulate clients, TFF requires the data to be inputted as federated data. In TFF, federated data is represented as an array of TensorFlow Datasets, with each dataset representing a single client's data. This data is used to run a federated training process, which is simply our centralized model wrapped as

a TFF model, with some parameters used in simulating client and server optimizers. The model is deployed to clients and weights are transmitted back to the aggregating server in as many rounds of computation as necessary to reach a desirable accuracy. Once the training rounds are done, a final model can then be evaluated against non-private test data stored by the central server, which leads to a final mode accuracy and loss.

This process requires subtle parameterization and we saw little to no adequate online help. We would recommend more community help and a better interface to customize clients' device and network states. Federated learning is best suited for large systems with many users, however, even smaller systems could end up benefiting form it, if the data being handled is too risky. TFF is best suited for simple algorithms, due to the edge device and network overhead. To increase its adoption, TFF should bridge the gap between simulation and deployment in order for researchers to gain more confidence in the utility of federated learning for their system.

5.3.1 Similarities and Differences Between Centralized and FL Model

While implementing a federated version of our centralized model we tried to maintain as much similarity between the two model setups as possible. The models were the same, a Keras sequential neural network, with a similar loss function. The client and server had seperate but similar optimizers in the FL environment with the same learning rate (0.001), which is the same rate used in the Centralized model setup.

Since the federated evaluation process happens solely on the central server, it is similar to a centralized evaluation process. We randomize our data with 42 as the random seed and split the data 80/20 for train/test data respectively, just like the centralized model.

Fundamental differences between our centralized and decentralized models can not be evaded. For example, the format of the input data is different, one is federated data and the other is not. It is also important to note that rounds of federated learning computations do not resemble epochs or iterations of a centralized model, because one is distributed and the other is not. In a federated computation, clients have their own local models running as well, which is not the case in a centralized model running 10 simple epochs. This is a fundamental difference and we are actually trying to measure how this difference presents itself in model accuracy and loss. We nonetheless treat TFF computation rounds similar to a centralized model's *epochs* and run as many rounds as we did epochs.

5.3.2 Drawback of Simulating Decentralized Systems Using TFF

TFF is decentralized in the sense that clients store their data and participate in the training of a global model, by locally training a model. However, the global model is aggregated by a central server, which in turn updates the edge devices to train the model locally and transmit back the weights. This process of receiving an updated global model, training a local model and transmitting weights back for aggregation is a single federated computation round, which should result in an updated loss function and model weights being held by the aggregator server at the end of a round. Obviously, the system is not fully decentralized, however, the theoretical principle remains that clients maintain more control over their data while the model improves in accuracy per round. To achieve consistent results, we ran sample collections of 10 round training computations. We used those samples to tune our model into a consistent accuracy without greatly affecting the differences between our two models or the accuracy and loss of our centralized model.

Unfortunately, TFF can not easily be customized to represent a real world scenario. Real edge devices and real network conditions would have to be simulated as well, which is tasking for any single computer or library. For this reason, our conclusions are solely based on the theoretical efficiency of the two compared models, and do not consider edge device and network performances such as network overhead or timing TFF rounds versus centralized epochs.

6 Ethical Considerations

Since this study is a course requirement without the need for external publication, an IRB approval wasn't required. However, considering the senstive nature of the dataset, multiple measures were taken to ensure security and privacy. The team's data science expert, who is also a student in the Privacy Engineering program, was responsible for maintaining the privacy of the dataset. Unfortunately, TensorFlow Federated (client simulation) cannot be used on cloud based services like Google Collab, hence the dataset had to be downloaded to the team's machines for analysis. To ensure the security of the data, multiple precautions were taken, such as minimizing copies, sharing the raw data on a need-to basis, and password protecting machines. At the end of the study, the dataset was deleted from the team's machines and Google drive.

7 Results and Discussions

The results of the study indicate that the accuracy of the federated learning model was comparable to that of centralized learning. The centralized model exhibited a final accuracy of 0.960 (see figure 6) and the federated model exhibited a

final accuracy of 0.971 (see figure 7) when tested against a test/validation dataset.

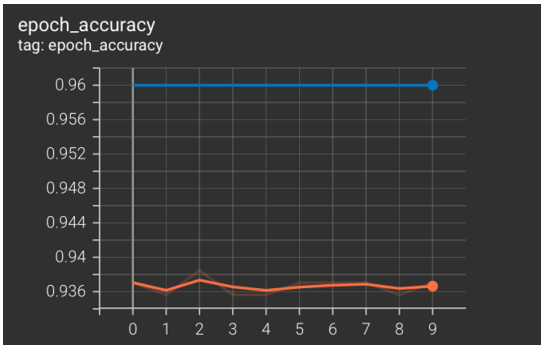

Figure 6: Centralized learning model accuracy

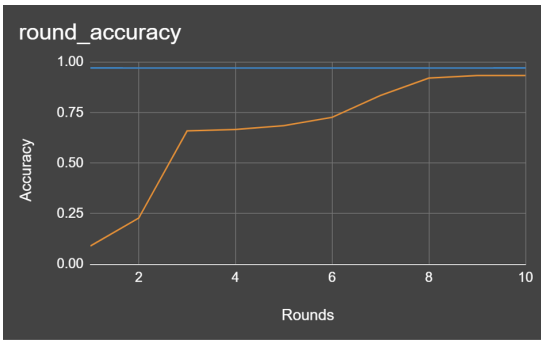

Figure 7: Federated learning model accuracy

However, the federated model had a slightly higher loss rate of 0.40975 (see figure 9) in comparison to the loss rate of the centralized model which was 0.177 (see figure 8).

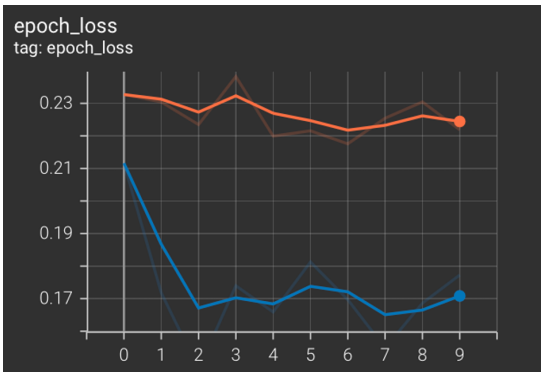

Figure 8: Centralized learning model loss

The differences in loss values between centralized and federated learning maybe due the multiple factors. One main difference is the way in which data is accessed and used. In centralized learning, the model has direct access to raw data and the training takes place in a central server. In federated learning, multiple local models are run simultaneously

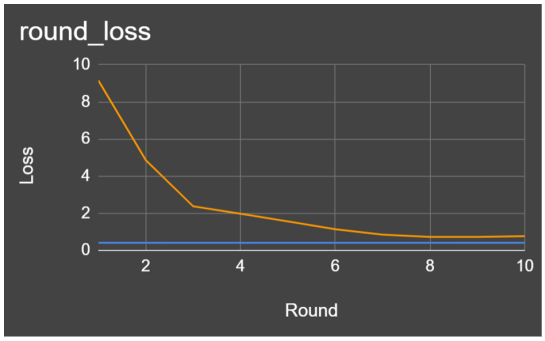

Figure 9: Federated learning model loss

on client devices, and the model updates are based on the average of the local models' parameters. However, due to differences in the local datasets, network conditions, and hardware capabilities, the local models may converge to different optima, which can result in slightly different model updates. These differences can accumulate over time and may result in higher loss values compared to a centralized approach.

8 Conclusion

The results of this study indicate that federated learning achieves a performance accuracy similar to that of a centralized learning model. The slightly higher loss for federated learning maybe attributed to communication overheads and differences in the weight updates provided by the local client models. It is also important to note that the results obtained were for this specific dataset and model architecture. Further research is needed to determine if the results of this study can be generalized to other datasets and models at varying scales. However, our study provides a key finding that it is possible to achieve a similar performance for predicting DLD conditions in a privacy preserving way, without sacrificing on model accuracy.

9 Limitations

We acknowledge the limitations of our study. First, there are geographical differences between both datasets, EIRLI was collected in a large metropolitan region in southern California and LASER was collected in a medium-sized city in northern Florida. This may have impacted the accruacy of our neural network model.

Second, TensorFlow Federated was implemented in virtual environment to simulate multiple clients. However, in a real-world setup, network and hardware differences across clients could affect the performance for federated learning. However, the impact due these factors would be minimal and federated learning can still be an effective way to detect developmental delays in a privacy preserving way.

10 Future Work

There are several opportunities for future work based on the study results. Some of them are discussed below:

- Investigating confusion matrices for Centralized and Federated learning models. This can provide insights into the types of errors made by the models. However, in the case of federated learning, confusion matrices will have to be computer per client. Comparing the number of false positives (FPs) in both learning setups can help identify areas of improvement.
- Examining the impact of different client participation rates. In federated learning, not all clients necessarily participate in every round of training. Future work could investigate how the participation rate of clients affects federated learning performance and privacy.
- This study focused on comparing the accuracy and loss for centralized and federated learning. Future work can explore privacy vs accuracy tradeoff in depth, by investigating cases where privacy preserving methods significantly affect utility.

11 Acknowledgement

We would like to express our gratitude to Arielle Brovosky and the other authors of the paper *"Moving towards accurate and early prediction of language delay with network science and machine learning approaches"* [13], for providing us with the EIRLI and LASER datasets used in this study. Their generous contribution has been instrumental in enabling our research, and we are thankful for their assistance.

References

[1] MBCDI derived measures.

[2] Module: tf.keras | TensorFlow v2.12.0.

[3] ReLU Activation Function Explained | Built In.

[4] TensorFlow Federated.

[5] tf.keras.Sequential | TensorFlow v2.12.0.

[6] Understanding binary cross-entropy / log loss: a visual explanation | by Daniel Godoy | Towards Data Science.

[7] Tensorflow federated tutorial session., july 2020.

[8] uuid, Sept. 2022.

[9] Developmental language disorder. national institute of deafness and other communication disorders, 2023.

[10] ABADI, M., CHU, A., GOODFELLOW, I., MCMAHAN, H. B., MIRONOV, I., TALWAR, K., AND ZHANG, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (oct 2016), ACM.

[11] ALIPANAHI, B., DELONG, A., WEIRAUCH, M. T., AND FREY, B. J. Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. *Nature biotechnology 33*, 8 (2015), 831–838.

[12] BISHOP, D. V., SNOWLING, M. J., THOMPSON, P. A., GREENHALGH, T., CONSORTIUM, C.-., ADAMS, C., ARCHIBALD, L., BAIRD, G., BAUER, A., BELLAIR, J., ET AL. Phase 2 of catalise: A multinational and multidisciplinary delphi consensus study of problems with language development: Terminology. *Journal of Child Psychology and Psychiatry 58*, 10 (2017), 1068–1080.

[13] BOROVSKY, A., T. D. . L. L. Moving towards accurate and early prediction of language delay with network science and machine learning approaches.

[14] BROWNLEE, J. Gentle Introduction to the Adam Optimization Algorithm for Deep Learning, July 2017.

[15] CITRON, D. K., AND SOLOVE, D. J. Privacy harms. *BUL Rev. 102* (2022), 793.

[16] DAGLIATI, A., MARINI, S., SACCHI, L., COGNI, G., TELITI, M., TIBOLLO, V., DE CATA, P., CHIOVATO, L., AND BELLAZZI, R. Machine learning methods to predict diabetes complications. *Journal of diabetes science and technology 12*, 2 (2018), 295–302.

[17] DEFIV. Applications of federated learning in healthcare, 2023.

[18] ESTEVA, A., KUPREL, B., NOVOA, R. A., KO, J., SWETTER, S. M., BLAU, H. M., AND THRUN, S. Dermatologist-level classification of skin cancer with deep neural networks. *nature 542*, 7639 (2017), 115–118.

[19] FRANK, M. C., BRAGINSKY, M., YUROVSKY, D., AND MARCHMAN, V. A. *Variability and consistency in early language learning: The Wordbank project*. MIT Press, 2021.

[20] KONEČNÝ, J.; MCMAHAN, H. Y. F. R. P. S. A. B. D. Federated learning: Strategies for improving communication efficiency.

[21] LAASONEN, M., S. S. L.-N. P. L. M. L. H. R. H. K. P. A. K. B. T. M. P. E. M. K. T. L. P. H. T. B. C. W. G. A. L. L. S. E. K. S. . A. E. Understanding developmental language disorder - the helsinki longitudinal sli study (helsli): a study protocol, 2018.

[22] LAW, J., BOYLE, J., HARRIS, F., HARKNESS, A., NYE, C., ET AL. Prevalence and natural history of primary speech and language delay: Findings from a systematic review of the literature. *International journal of language and communication disorders 35* (2000), 165–188.

[23] LIPTON, Z. C., KALE, D. C., ELKAN, C., AND WETZEL, R. Learning to Diagnose with LSTM Recurrent Neural Networks, Mar. 2017. arXiv:1511.03677 [cs].

[24] MCGREGOR, K. K. How we fail children with developmental language disorder. 982.

[25] NORBURY, C. F., GOOCH, D., WRAY, C., BAIRD, G., CHARMAN, T., SIMONOFF, E., VAMVAKAS, G., AND PICKLES, A. The impact of nonverbal ability on prevalence and clinical presentation of language disorder: Evidence from a population study. *Journal of child psychology and psychiatry 57*, 11 (2016), 1247–1257.

[26] RIEKE, N., HANCOX, J., LI, W., MILLETARI, F., ROTH, H. R., ALBARQOUNI, S., BAKAS, S., GALTIER, M. N., LANDMAN, B. A., MAIER-HEIN, K., ET AL. The future of digital health with federated learning. *NPJ digital medicine 3*, 1 (2020), 119.

[27] SAEED, M. A Gentle Introduction To Sigmoid Function, Aug. 2021.

[28] STANTON-CHAPMAN, T. L., CHAPMAN, D. A., BAINBRIDGE, N. L., AND SCOTT, K. G. Identification of early risk factors for language impairment. *Research in Developmental Disabilities 23*, 6 (2002), 390–405.

[29] TEAM, K. Keras documentation: The Sequential model.

[30] TOMBLIN, J. B., RECORDS, N. L., BUCKWALTER, P., ZHANG, X., SMITH, E., AND O'BRIEN, M. Prevalence of specific language impairment in kindergarten children. *Journal of speech, language, and hearing research 40*, 6 (1997), 1245–1260.