Protection of Patient Privacy on Mobile Device Machine Learning

Matthew Nguyen
Old Dominion University

Protection of Patient Privacy on Mobile Device Machine Learning

Matthew Nguyen

Jiangwen Sun (Mentor)

Old Dominion University

November 22, 2021

Abstract- An existing StudentLife Study mobile dataset was evaluated and organized to be applied to different machine learning methods. Different variables like user activity, exercise, sleep, study space, social, and stress levels are optimized to train a model that could predict user stress level. The different machine learning methods would test if both patient data privacy and training efficiency can be ensured.

Keywords- centralized machine learning, decentralized machine learning, federated learning, federated multi-task learning

I. INTRODUCTION

The concerns for patient data privacy on mobile devices have become increasingly discussed as healthcare transitions towards a technologically based system. Data is constantly being collected and stored on our mobile devices by applications and websites. However, many patients who have their private data stored on their mobile devices wish to keep their information secure without being shared or accessed by others. As the growth of mobile technology progresses towards the field of artificial intelligence, machine learning, and deep learning, the use of patient data is crucial in the development of such technologies. More importantly, mobile patient data needs to be kept private and secure without compromising the advancement of new technology.

In this work, we explored different methods of machine learning to apply to the StudentLife Study mobile dataset from Dartmouth College to predict user stress levels. The dataset was first sorted and analyzed to see the desired data points and dataset variables we wanted to work with. We intended to use the variables of activity, exercise, sleep, study space, and social level, to predict and compare it to user stress levels. The different methods of machine learning include potentially looking at centralized machine learning, decentralized federated learning, and decentralized multi-federated learning. Our task was to analyze the results presented from each learning method and understand which machine learning method gave us the most accurate stress level while maintaining mobile patient data privacy.

II. StudentLife Study Mobile Dataset

The dataset used from a Dartmouth study provided us with both passive and automatic sensing information from the phones of 48 undergraduate and graduate students over the span of 10 weeks that comprises 53 GB continuously collected data, 32,000 self-report data points, and pre/post surveys. The StudentLife app used on user mobile devices during this study has collected the following data points on the users involved:

- objective sensing data including sleep (bedtime, duration, wake up), conservation duration, conversation frequency, and physical activity (stationary, walk, run)
- location-based data of location, co-location, indoor and outdoor mobility
- Self-reports with affect (PAM), stress, behavior, Boston bombing reaction, cancelled classes, class opinion, comment, Dartmouth now, Dimension incident, Dimension protest, dining halls, events, exercise, Green Key, lab, mood, loneliness, social and study spaces.
- academic performance data: class information, deadlines, grades (grades, term GPA, cumulative GPA), piazza data
- dining data with meals data, location and time
- seating data with seating position of students in Android programming

The large dataset was downloaded and evaluated using HPC (High Performance Computing). The dataset was then transferred from a json text file into a readable table using the Matlab and Microsoft Excel application. We consolidated the dataset into the following variables intended to be used for the machine learning application:

Dataset variables to measure stress levels	Description
Activity	Location, alone working, alone relaxing, with other people working, and with other people relaxing.
Exercise	Location, if they have/not exercised, if they

	could/not exercise because of schedule, the length of time for exercise, and length of time for walking.
Sleep	Location, hours slept the previous night, rating sleep quality, and trouble staying awake.
Study Space	Location, place of study, rate of productivity, and ambient noise level
Social	Location, the number of people they interacted with
Stress	Outcome (Binary variable based on response) (0 = no stress, 1 = stress)

A total of 15 users were found to have data available for activity, exercise, sleep, study space, social, and stress, which had sufficient data available for training machine learning models. For users to qualify to be used in training a model, they must include entries for all necessary variables that exceed a total of 10 data entries. All variable data points had a response time along with each entry. The data entries were organized in chronological order based on response time associated with the entry.

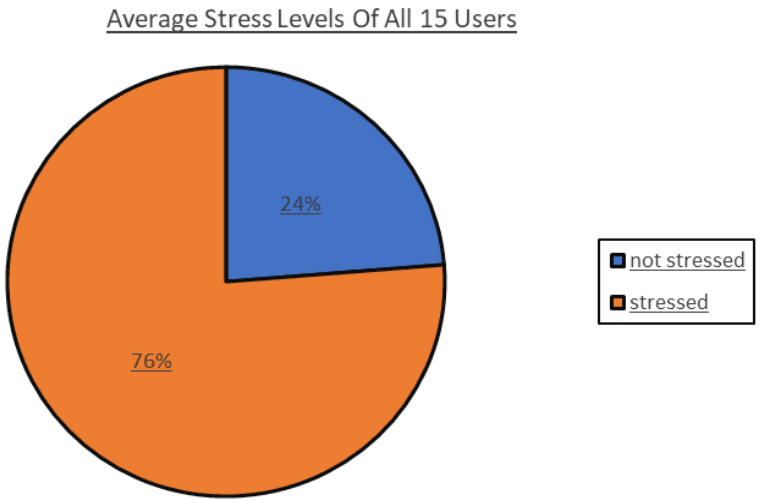

Fig. 1. The pie chart shows the percentage of stress levels of all 15 users throughout the 10 week study period. Users were stressed for around 76% of the time and not stressed for about 24% of the time during the study.

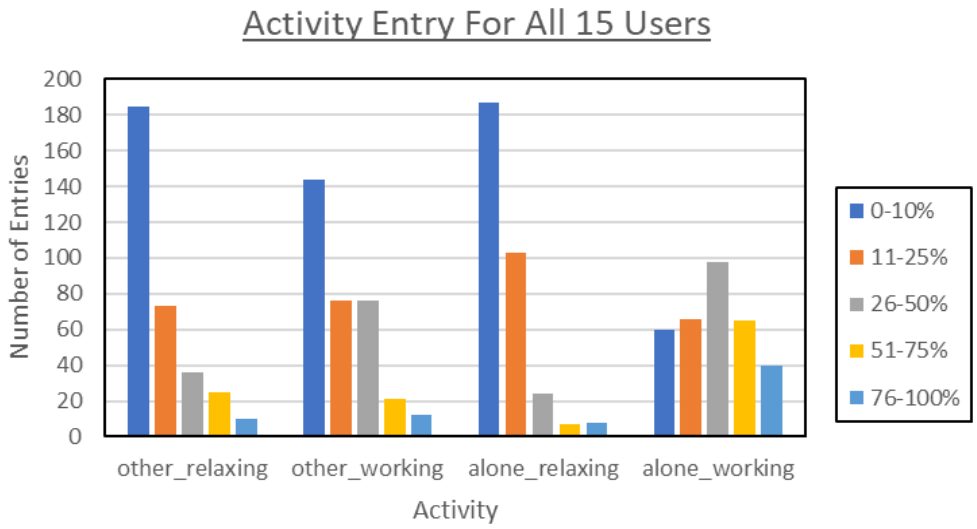

Fig. 2. The bar graph shows the percent of the activity being accomplished by all 15 users throughout the 10 week study period. The majority of time during this study was spent working alone with a normal distribution representation. Relaxing with others, working with others, and relaxing alone skews to the right towards the 0-10% category, which indicates that those activities were done the least.

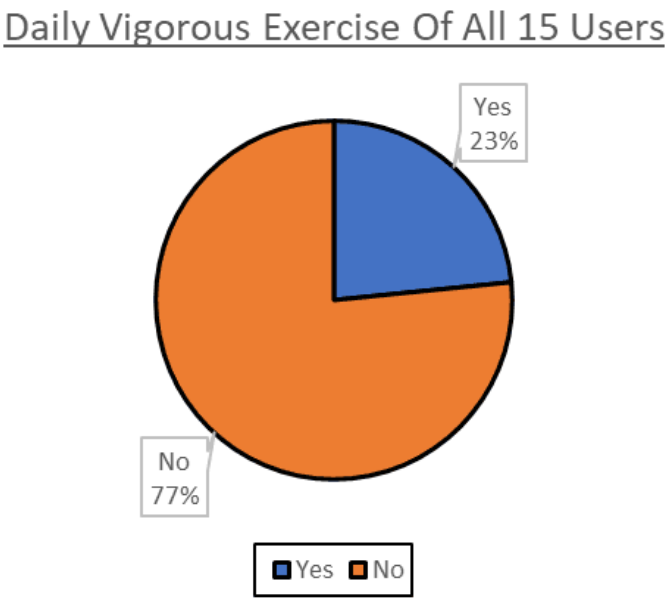

Fig. 3. The pie chart shows the entry responses asking if daily vigorous exercise was done by all 15 users throughout the 10 week study period. The majority of entries submitted by users stated that they did not participate in vigorous exercise that day.

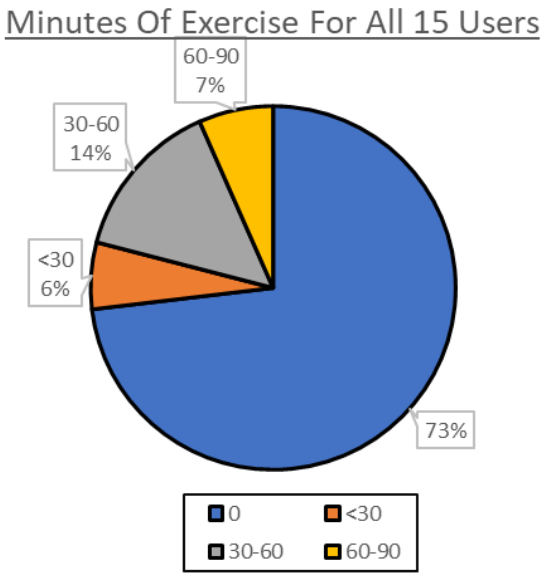

Fig. 4. The pie chart shows the percentage of minutes of exercise done by all 15 users throughout the 10 week study period. The majority of entries submitted by users stated that they participated in 0 minutes of exercise. For the users who did exercise, 14% of them spent roughly 30 minutes to 1 hour exercising.

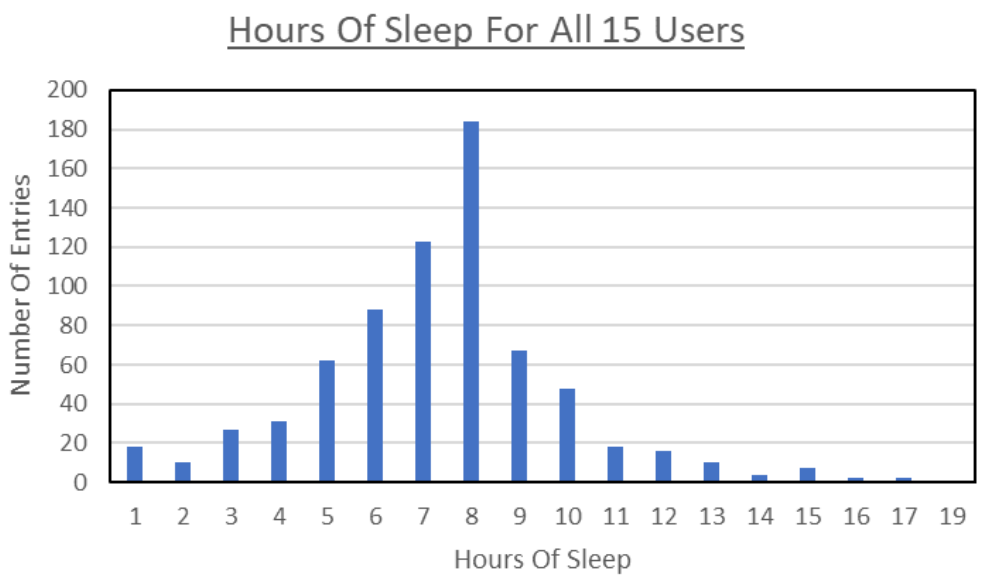

Fig. 5. The bar graph shows the number of entries of hours of sleep recorded by all 15 users throughout the 10 week study period. This shows a majority of users had around 7 to 8 hours of sleep.

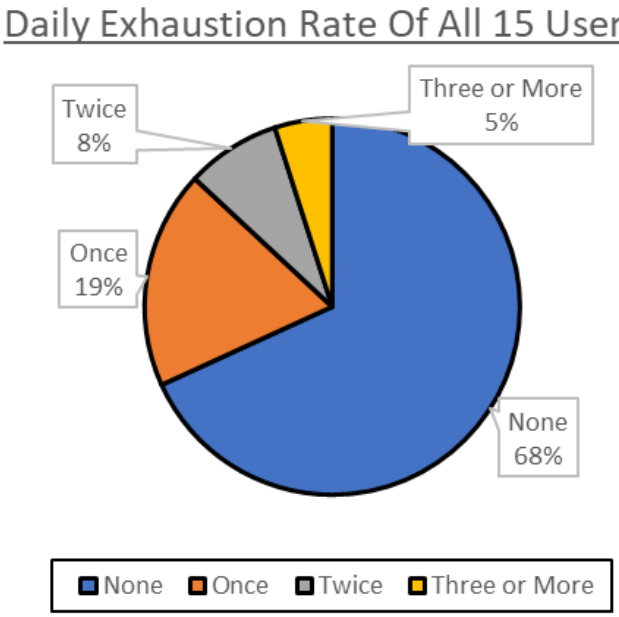

Fig. 6. The pie chart shows the percent of daily exhaustion experienced by all 15 users throughout the 10 week study period in correlation with their sleep. This shows that users were primarily not exhausted and about 20% of the time felt exhausted once.

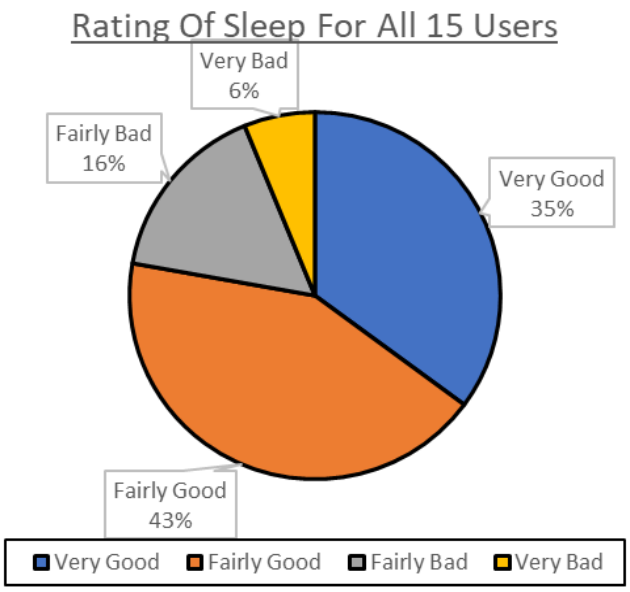

Fig. 7. The pie chart shows the percentage of sleep rating given by all 15 users throughout the 10 week study period in correlation to their sleep. This shows a majority felt that they had very good or fairly good sleep based on the hours of sleep they received.

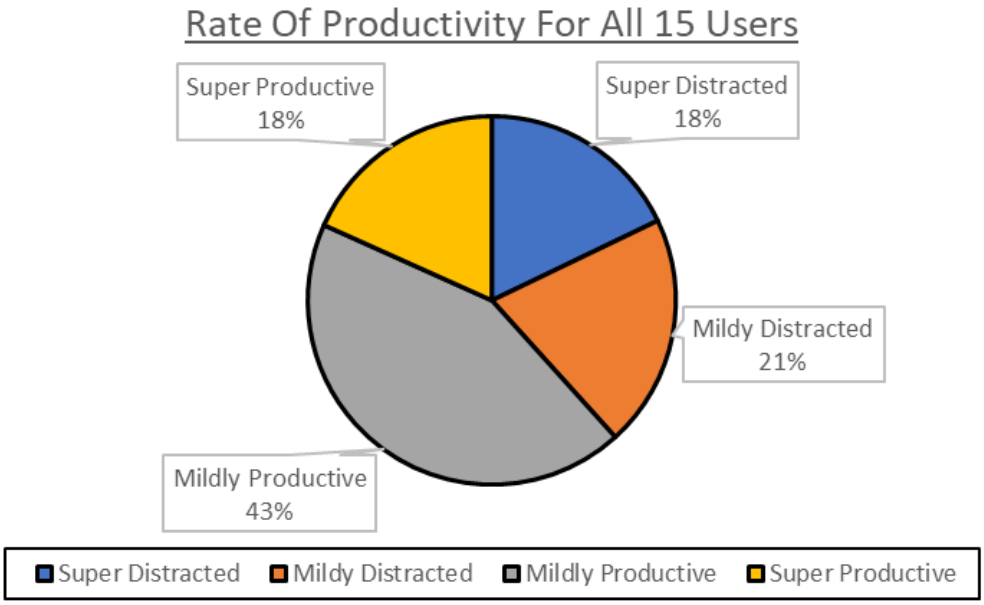

Fig. 8. The pie chart shows the percentage of productivity accomplished by all 15 users throughout the 10 week study period in their designated study space. From the entries submitted by users, 43% of the time spent studying was mildly productive while roughly 20% of the time studying was super productive, super distracte, or mildly distracted.

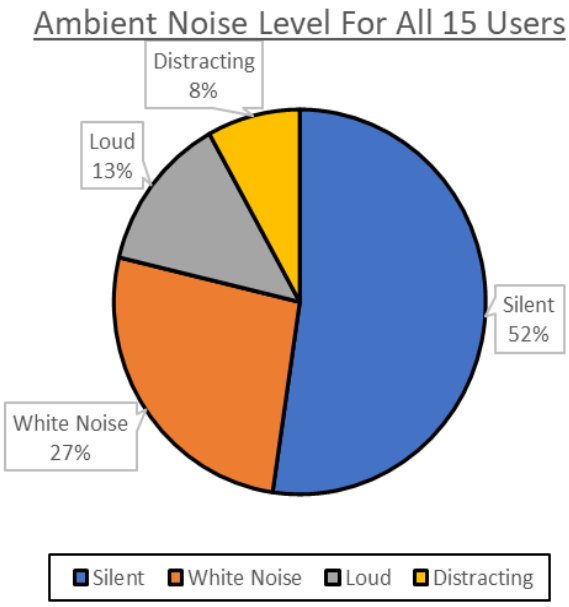

Fig. 9. The pie chart shows the percentage of ambient noise level recorded by all 15 users throughout the 10 week study period in their designated study space. From the entries submitted by users, 52% of the time spent studying was in a silent space while 27% of the time studying had white noise.

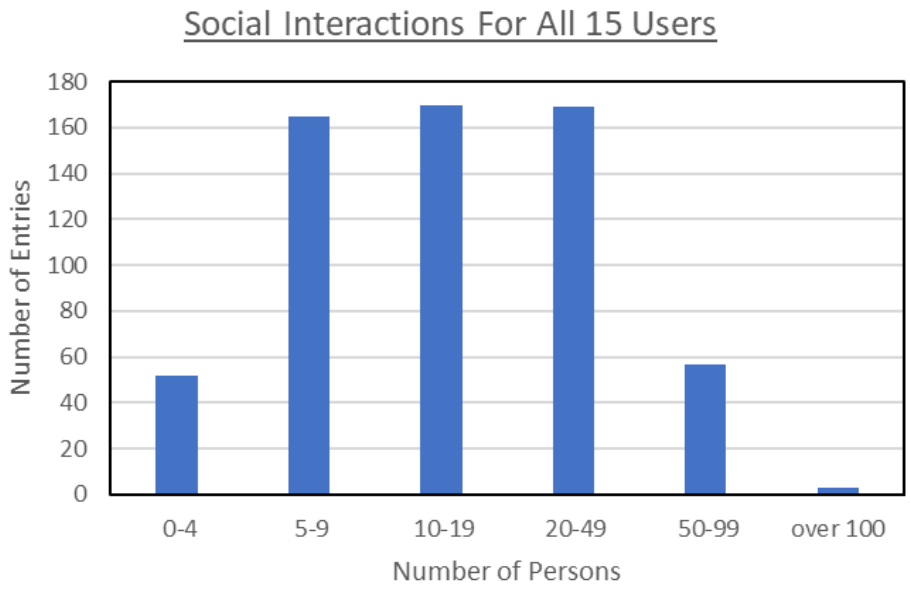

Fig. 10. The bar graph shows the daily number of people interacted with by all 15 users throughout the 10 week study. The majority of days where users had social interaction involves primarily encountering a number of 5 to 49 people.

III. Machine Learning Methods

Machine learning traditionally involves training a model in a centralized method where individual device information is stored over a shared cloud database for Internet of Things (IoT). The individualized data is then trained to produce a broad model which gets redistributed and used by all the devices. Despite the convenience behind creating a general model that can be broadly applied, there are many issues associated with the methodology. Many concerns involve the amount of battery consumption centralized training incurs. Transferring data to a centralized cloud server and downloading the model updates continuously can quickly drain the device battery life and be slow for devices with an unstable network connection. An additional concern with centralized training is the lack of user data privacy. Data sent to the cloud can be at risk for being misused or damaged.

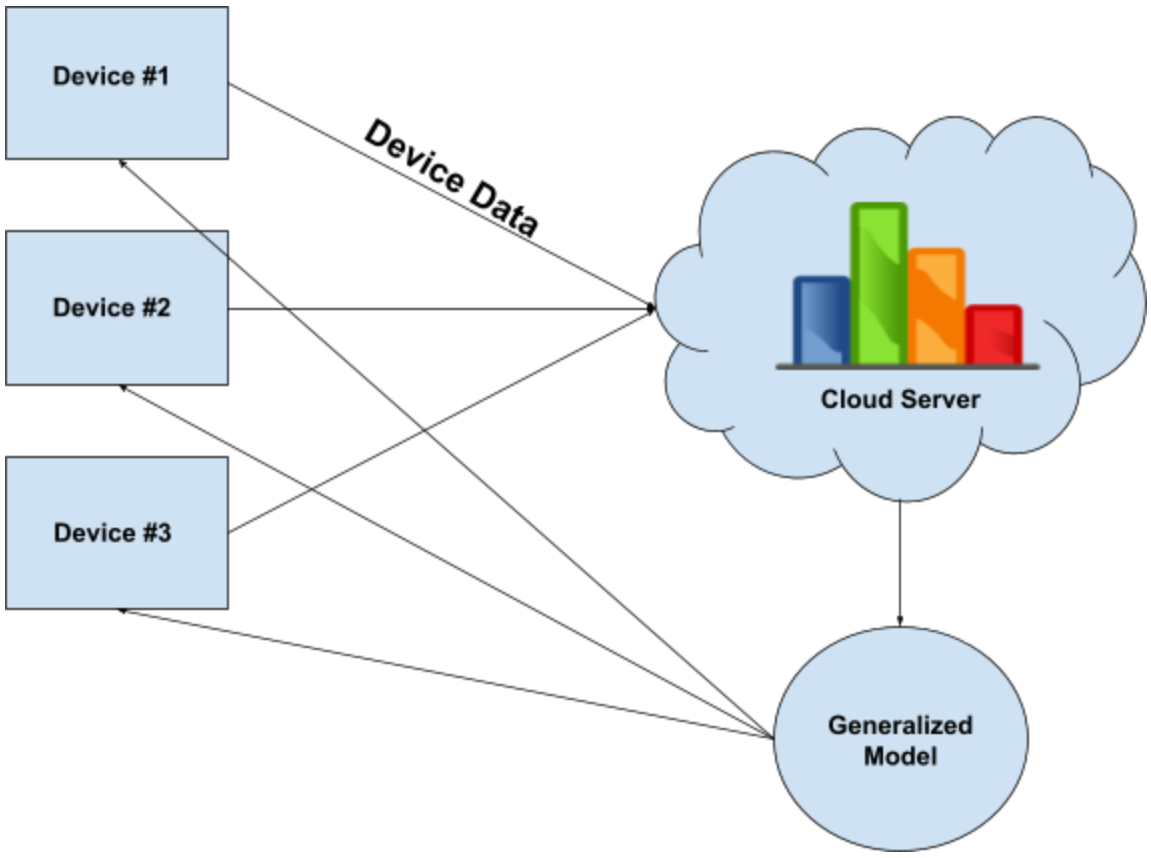

Fig. 11. A diagram of centralized machine learning. Each device sends data directly to the cloud server. A generalized model is generated and sent to all devices to be applied.

Decentralized machine learning is a more privatized and newer approach to training a machine learning model. Decentralized machine learning prevents personal data from being exposed to the cloud server by creating a user specific model using the streaming data on each device. The individually trained models are then sent to the cloud server to be compared for variations and sent back to each device to be improved on over time. Decentralized machine learning learns from data stored locally on each device which keeps user data confidential and safe from the cloud. The amount of battery consumption is lowered since only trained models are sent to the cloud opposed to all user data. Keeping data localized and only sending models to the cloud decreases the dependency for constant internet connection.

A more specialized, decentralized machine learning method that allows for a general training model to be created is called federated learning. Federated learning allows for a single model to be trained and used among all devices while maintaining the privacy and efficiency of decentralized learning. It operates by distributing an initial model to all devices that trains and enhances the model with local user data. The locally trained models are then collected in the cloud to be averaged into a new generalized model that is received by each device to continue the learning cycle. The application of federated learning for mobile machine learning has been simplified with the open-source framework called TensorFlow Federated (TFF) made by Google. This framework can be installed with a pip package manager and has the option for developers to choose from existing available models.

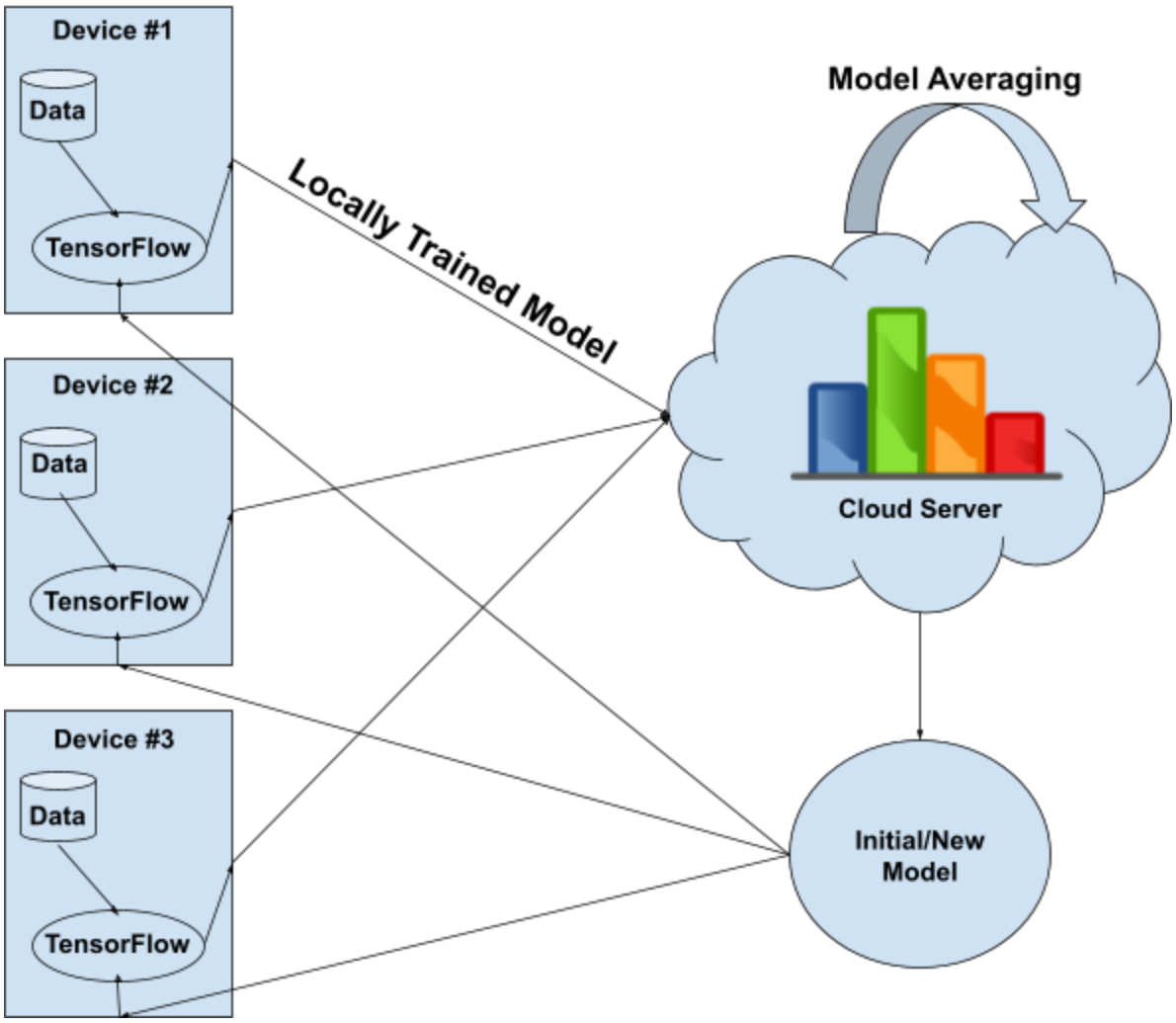

Fig. 12. A diagram of decentralized federated machine learning. An initial model is sent to all devices and trained locally by the TensorFlow Federated application using local device data. A locally modified model is then produced and received by the cloud server. All models are averaged to produce a new single model to be learned by all devices.

The disadvantage with applying a generalized learning model to all user data is that it might give inaccurate results when trying to measure personalized metrics like mood, behavior, or habits. A more user specific learning approach is federated multi-task learning. This approach factors in the varying heterogeneity in each client by learning a personalized model for each client. Individualized user models are first learned using client data. An improved personal model is then received by a cloud server for model relationship learning and sent back to the client to continue the learning cycle.

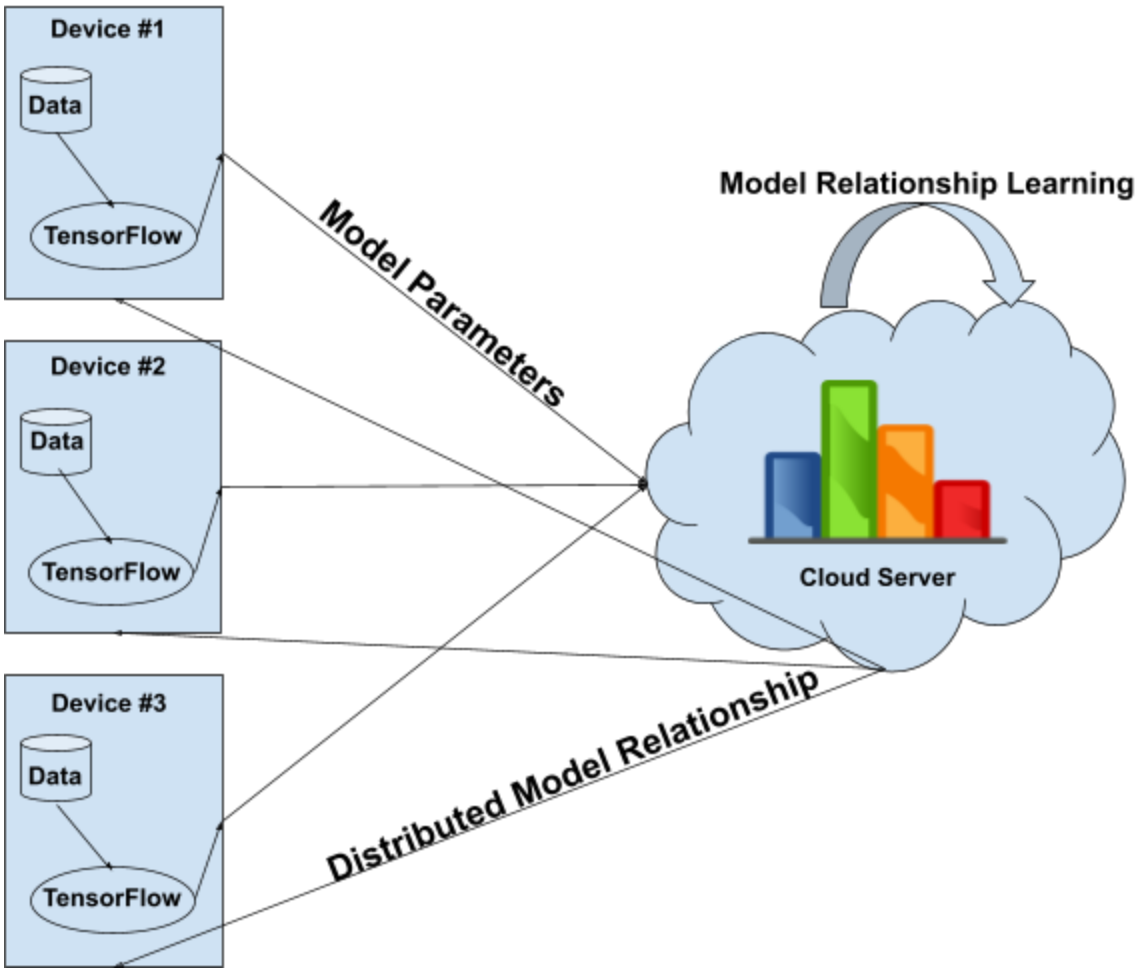

Fig. 13. A diagram of decentralized federated multi-task machine learning. Each device locally trains an individual model through the TensorFlow Federated application. The device model parameters are sent to the cloud. A model relationship is learned and distributed to devices to improve personalized model learning.

IV. CONCLUSION

For the duration of this research, our goal was to understand how patient data can be protected alongside the application of mobile machine learning. We wanted to explore this by using the existing StudentLife Study Mobile Dataset. We used High Performance Computing (HPC), Matlab, and Excel to organize the data points for 15 users. Each user had sufficient data for the variables of activity, exercise, sleep, study space, and stress. These variables are used to train a model predicting user stress levels. We would then compare the resulting stress levels

predicted by centralized machine learning, decentralized federated learning, and decentralized federated multi-task learning. Due to time constraints, we did not have the opportunity to apply the user data to train models for the three different learning methods. The project hopes to be continued to further our understanding of how patient data can be protected in the machine learning process.

REFERENCES

Arun. (2020, October 6). *Understanding federated learning*. Medium. Retrieved November 19, 2021, from https://towardsdatascience.com/understanding-federated-learning-99bc86a0d026.

Federated multi-task learning under a mixture ... - arxiv.org. (n.d.). Retrieved November 19, 2021, from https://arxiv.org/pdf/2108.10252.

Liu1, J. C., Goetz1, J., Sen2, S., Tewari1, A., Statistics, 1D. of, & Liu, C. A. J. C. (n.d.). *Learning from others without sacrificing privacy: Simulation comparing centralized and Federated Machine Learning on Mobile Health Data*. JMIR mHealth and uHealth. Retrieved November 19, 2021, from https://mhealth.jmir.org/2021/3/e23728.

On multiplicative multitask feature learning. (n.d.). Retrieved November 19, 2021, from https://cs.odu.edu/~jsun/doc/pub/nips-wang-2014.pdf.

SWP federated learning final version - ekkono solutions ab. (n.d.). Retrieved November 19, 2021, from https://www.ekkono.ai/wp-content/uploads/2020/12/SWP_Federated_Learning_Ekkono_Solutions_May_2020.pdf.