Federated Learning Between Monoethnic Clinical Institutes

ANNIKA L. SCHAVEMAKER ALEXANDER MO
schav | alexamo @kth.se

January 15, 2023

Abstract

Leveraging federated learning for clinical decision support across clinical institutes can have many advantages due to the privacy-friendly and efficient nature of federated learning. Nevertheless, clinical decision support models might function differently for different ethnicities because of differences in physical body structure. As a result, it may cause the application of federated learning to be ineffective or unfavorable for monoethnic clinical institutes. This research investigated the performance of a federated forest model on the MIMIC-IV dataset split on ethnicity and found within the scope of the study that some local models outperform the federated models by a significant margin, while other models seemed to improve. Overall, little difference (+- 1.5%) can be seen between the average accuracy, precision, recall and area under the receiver operating characteristic curve of the local models and the federated model. However, it should be taken into account that this study features various limitations which require further investigation to generalize the statements made.

Contents

List of Acronyms and Abbreviations

AI Artificial Intelligence

AUC Area Under the Curve

BIDMC Beth Israel Deaconess Medical Center

EHR Electronic Health Record

eMAR electronic Medication Administration Records

GDPR General Data Protection Regulation

HIPAA Health Insurance Portability and Accountability Act

PHI Protected Health Information

ROC Receiver Operating Characteristic

1 Introduction

1.1 Context

With the rise of machine learning and AI, there has been an increase in medical digitization around the globe [1]. Hospitals and other caregiving facilities keep EHR for monitoring their patients, which can be used to automate processes or support decision making for professional care physicians [2].

The difficulty with creating reliable models for decision support is the limited amount of data available to hospitals and other caregiving facilities. Clinical data contains highly Protected Health Information (PHI), which makes it virtually impossible for hospitals to share their EHR with other hospitals [2].

Multiple researchers have investigated the use of federated learning to solve this problem due to the privacy-friendly nature of this machine learning approach. The researchers cut up their initial dataset into several subsets, which were used to imitate training data for the local models. The local parameters would then be exchanged and updated with the global parameters [3].

However, as has been pointed out by several papers researching the bias between medical data among different regions in the world, the quality of the data varies significantly in completeness and accuracy [4]. Additionally, the skewed data caused by geolocation and its respective ethnic population distribution directly affects the model, where physical differences in body structure might introduce undesired bias in certain models [5]. Reflecting on federated learning models in healthcare, the aggregation system attempts to increase the number of contributing samples in sheer quantity to combat overfitting. Based on this approach, the aforementioned issues motivate this work to debate the quality of the federated contributions to respective local models.

1.2 Aim

If different local models train on patients with different ethnicities, with potentially very different physical body structures, updating with the global model parameters might not be beneficial in terms of performance, see figure 1. Nevertheless, how these ethnical differences can impact the overall performance of the federated learning approach is currently unknown. This is why we aim to solve the following research question:

What is the impact of monoethnic data on the performance of a federated learning approach to clinical decision support?

To answer this question, this research will divide the MIMIC-IV clinical dataset into subsets, where each subset will consist of one specific ethnicity. Each of these subsets will represent data for the local models. The local models will consist of a random forest classifier to classify whether the diagnoses of the patient is related to kidney issues, such as kidney failure and chronic kidney disease. The goal is to provide performance metrics for each local model individually, as well as providing performance metrics for the federated learning approach on the local test data.

2 Theoretical Framework

2.1 Federated Learning

As mentioned previously, this project aims to assess the performance of federated learning algorithms between monoethnic healthcare facilities to investigate potential improvements in clinical decision support algorithms.

Federated learning is a machine learning framework first proposed by Google in 2017. It allows collaborative learning without centralized training data. Essentially, every node trains a local model based on its own training data and sends the local model parameters to a global model. The global model receives parameter updates from several nodes and uses them to update the global parameters. The local parameters

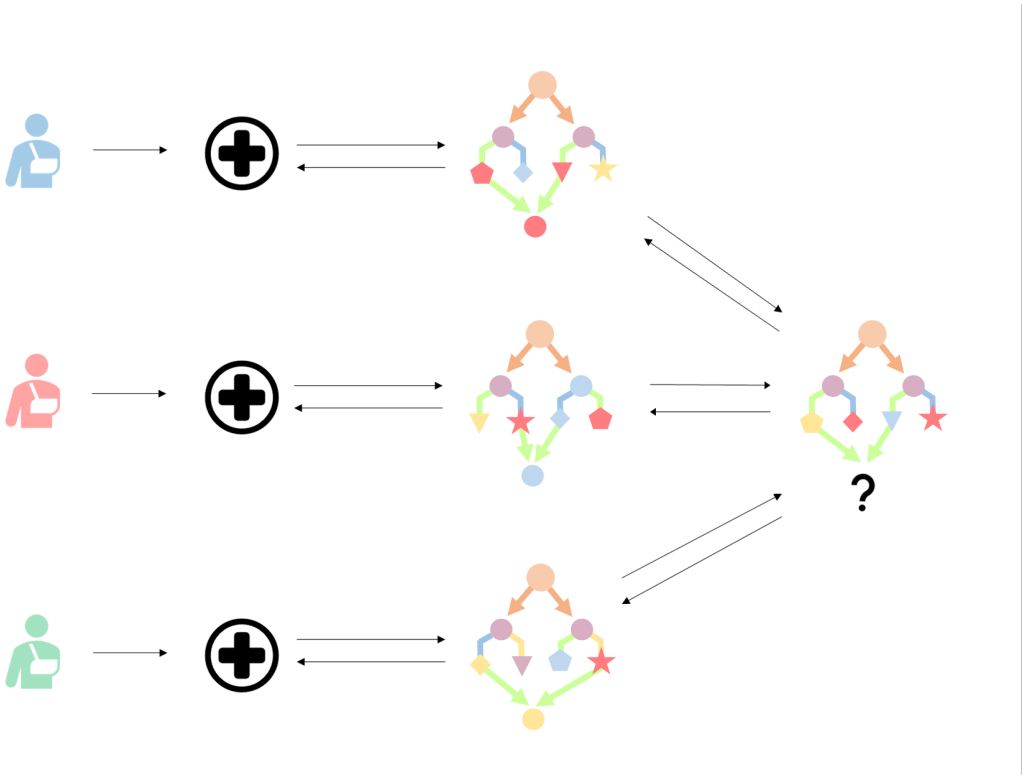

Figure 1: Problem illustration: what is the effect of monoethnic local models in a federated learning approach?

will in turn be updated with the global parameters. The major advantage of this approach is that the nodes never have to send their data to the global model. This has many benefits in terms of latency, power consumption and privacy [6].

Ever since the introduction of this principle, multiple different algorithms have been developed using different machine learning approaches for the local models, thus resulting in different federations [7].

2.2 Federated Forest

Many medical researchers have pointed out the effectiveness of applying random forests for predicting medical diagnoses due to their interpretability [8]. Thus, this research has opted to implement the local models using a random forest that predicts whether a patient is diagnosed with a kidney related issue. The global model will therefore be a federated forest, as described by Liu et al. [9] and Li et al. [10].

Specifically, this research will mainly be based on the approaches described by Liu et al. [9], with some adjustments to turn the proposed vertical federated learning problem into a horizontal problem resembling the available data. Deviating from the problem statement, the data used in this research has the same features across hospitals, but contains different patients. The problem described in the study by Liu et al. features clients who all possess the same data subjects but different feature spaces.

2.3 Ethnic Bias

As previously described, physical differences among patients from contrasting ethnic groups, as well as skewed data categories, cause a bias in the diagnosis of patient conditions in many cases [5]. Most prominently, this phenomenon is evident in clinical decision support systems, which motivates the baseline model target of this work. Clinical decision support systems have been shown to reflect ethnic bias despite omitting this parameter as a training feature in its respective algorithm [4]. Although similar in its objective, the mentioned work directs the cause of bias toward three external aspects of the data as opposed to the personal factors of the patients. Namely, the cause of the classification bias was suggested to lie with:

1. Missing data

2. Misclassification or measurement error

3. Sample size

Regarding these sources of bias, measurement errors and missing data are expected to be uniform across ethnicities in the MIMIC-IV dataset as all data subjects have been recorded at the same institute, Beth Israel Deaconess Medical Center (BIDMC). On the other hand, this caused the ethnic sets of data to be skewed, i.e., sample size representation per ethnicity makes the model training prone to underfitting. Using regularization techniques or resampling methods [11, 12], such bias could be reduced to shift the focus towards physical differences in data values among different ethnic groups.

2.4 Ethical Considerations

In recent years AI has taken a role in improving living standards in what is the start of smart societies [13]. According to the paper by Thwal et al., advances are catered towards key sectors of education, transportation, healthcare, economy, and environmental management. Relevantly, in healthcare it has been evident that recent AI innovations allow for enhanced and more efficient handling of real-life situations, e.g. the COVID-19 pandemic [14]. Despite great societal benefits, the controversy around private data has been an obstacle the AI sector has struggled with, especially in the healthcare application sector [15]. Based on this topic of discussion, two main ethical problems arise. Firstly, the ownership of personal data is weighted against improving algorithms that could potentially save lives or make large-scale processes more efficient and environmentally friendly in other sectors. Second, the usage of personal data features has been found to result in underlying bias and discrimination among social groups [16].

In the first ethical debate, the classical conflict of interest between society and individuals is laid bare. Commonly it has been argued in favor of societal impact along utilitarianism [17], a philosophy that contends for the outcome that brings the greatest good to the greatest number of individuals. Conversely, virtue ethics have been proposed to find a better middle ground in the privacy paradox.

Regardless of the philosophical trends applied to the case, it appears real consequences could be attached to the integration of data which can be distinguished based on discriminatory features such as geolocation and ethnicity. The presence of bias in a given model is controversial in its justification, which makes it essential to research the model in an explainable fashion [18].

2.5 Sustainability Considerations

Since the slowly increasing adoption of AI in the healthcare sector, many care institutions are keeping track of patients' PHI. Clinical decision support tools can be trained on the locally available data, but are likely to be more accurate when trained on more data. Federated learning poses an excellent solution to this problem, not just in terms of securely transferring protected information, but also in terms of efficiency, latency and energy consumption. As merely the model parameters are exchanged, rather than all the data points, communication between entities becomes faster and more efficient. Additionally, updating the local models with the newly communicated information does not require the models to rerun, but the parameters can simply be updated, therewith saving computing power and energy. For these reasons, federated learning could offer a good solution with minimal burden on the environment.

3 Hypotheses

Based on previously mentioned research, it is expected that applying federated learning approaches on monoethnic local models perform worse than the local models itself.

4 Methods

4.1 Data Acquisition

4.1.1 MIMIC-IV

For this research, the MIMIC-IV v0.4 dataset is used. MIMIC-IV contains, amongst others, patient demographics, hospitalization records and EHR of over 40,000 patients at the Beth Israel Deaconess Medical Center (BIDMC) between the years 2008 and 2019. The data is deidentified according to the Health Insurance Portability and Accountability Act (HIPAA) Safe Harbour provision. Access to the dataset is granted upon completion of the CITI Data or Specimens Only Research training [19].

4.1.2 Ethical Considerations

Multiple privacy and security laws aim to protect clinical health data from being abused. The world's toughest privacy and security law is the General Data Protection Regulation (GDPR), which sets clearly defined legal terms for the processing of personal data, where personal data refers to any information that relates to an individual who can directly or indirectly be identified [20]. For the healthcare industry specifically, the Health Insurance Portability and Accountability Act (HIPAA) also applies. This requires the data subjects to consent to the use and processing of their PHI and ensures the confidentiality, integrity and availability of the PHI [21].

As briefly mentioned above, the MIMIC-IV dataset ensures to comply with these privacy and security laws by deidentifying the data. Patient identifiers as stipulated by HIPAA were replaced by random ciphers, structured data was filtered, and dates and times were shifted [19]. Additionally, this research followed the guidelines specified by CITI Data and Specimens Only Research in their respective training and conditions. The four principles of Beauchamp and Childress - autonomy, non-maleficence, beneficence and justice - are also taken into serious consideration before conducting this research. This allowed the intended research to utilize the clinical data in a responsible manner.

4.2 Software and Packages

This research has been conducted using Jupyter Notebook on Python 3.10.6 and the packages depicted in table 1. The code can be found in the public GitHub repository of the project [22].

	Package	Version
preprocessing	pandas dask numpy	1.4.4 2022.2.1 1.21.5
local models	pandas numpy scikit-learn	1.4.4 1.21.5 1.1.2
federated model	pandas numpy	1.4.4 1.21.5
visualizations	pandas seaborn tensorflow matplotlib	1.4.4 0.11.2 2.8.2 3.5.2
performance metrics	pandas scikit-learn	1.4.4 1.1.2

Table 1: The packages and their version number per research step

4.3 Data Exploration

The MIMIC-IV v0.4 data is divided into three folders: `core, hosp` and `icu`. The `core` folder contains all the data tables related to patients' demographics and admissions. The `hosp` folder includes various data tables related to regular hospitalizations, such as diagnoses and medications. The `icu` folder specifically contains data for intensive care admissions.

Between the different ethnicities featured in the data set, the number of samples varies greatly as the 'white' class represents more than half of the samples. As a result, the local models for the large classes have much more data to train on, which in theory often results in better models. In order to mask such imbalance, the models were trained to incorporate and dissuade misclassification through the Area Under the Curve (AUC) metric in the hyperparameter tuning phase.

4.4 Data Preprocessing

4.4.1 Preprocessing-I: Filtering

In the first step of the data preprocessing, the MIMIC-IV data was filtered. Table 2 presents what columns were selected from the MIMIC-IV tables. Generally, this research filtered out the columns concerning regular hospitalizations, diagnoses, electronic Medication Administration Records (eMAR) and lab findings on, for instance, urine and blood samples. This is the data deemed essential to build a simple classifier for the scope of this research.

Additionally, this step of the preprocessing aggregated values by `hadm_id` for certain data tables. This ensures that data is represented as an array of values, rather than multiple rows with one value. Intuitively, this can be seen as having one hospitalization record with a list of 10 lab results, for example, instead of having ten separate records with one lab result each for the same hospitalization. For convenience, a column containing the length of this array is added as well. The name of these columns ends on `_count`.

As a final side note on the first step of the preprocessing, it can be seen from table 2 that `charttime` often appears in the selected columns, but not in the aggregated values. The reason for this is that all values for `charttime` aggregated by `hadm_id` are the same. Thus, the first value of `charttime` is taken to represent all charttimes for that specific `hadm_id`.

The selected and added columns were merged by `hadm_id` and split and saved by `ethnicity`.

	Selected Columns	Aggregated Values	Added Columns
/core/admissions.csv.gz	['subject_id', 'hadm_id', 'admittime', 'admission_type', 'admission_location']		
/core/patients.csv.gz	['subject_id', 'insurance', 'marital_status', 'ethnicity', 'edregtime', 'gender', 'anchor_age', 'anchor_year']		
/hosp/diagnoses_icd.csv.gz	['hadm_id', 'icd_code']	'icd_code'	'icd_code_count'
/hosp/emar.csv.gz	['hadm_id', 'charttime', 'medication', 'event_txt']	['emar_charttime', 'emar_medications']	'emar_count'
/hosp/labevents.csv.gz	['hadm_id', 'charttime', 'flag', 'priority', 'comments']	['lab_flag', 'lab_priority', 'lab_comments']	'lab_count'

Table 2: The columns selected and adapted from MIMIC-IV in the first preprocessing step

4.4.2 Preprocessing-II: Labeling

This step of the preprocessing focused on determining and creating a column for the dependent variable. Merging `mimic-iv-0.4/hosp/d_icd_diagnoses.csv.gz` and `mimic-iv-0.4/hosp/diagnoses_icd.csv.gz` and counting word occurrences in the column `long_title` revealed that the word 'kidney' appeared most often (ignoring words such as 'of' and 'and'), with 189,696 occurrences out of 4,694,786 diagnoses. Such diagnoses can range from acute kidney failure to (diabetic) chronic kidney disease.

Based on this, all hospitalizations that contained at least one kidney related diagnosis were assigned the boolean value `True` in the newly created column `has_kidney_issue`, and `False` otherwise. The column `icd_code` was removed and CSV files were outputted again for each ethnic group. The distribution of patients with kidney issues across ethnic groups is presented in table 3.

	White	Black / African American	Hispanic / Latino	Other	Asian	Unknown	Unable to Obtain	American Indian / Alaska Native
Data Subjects	338,044	80,526	29,887	26,844	24,522	19,419	3,742	1,536
Patients with Kidney Issues	61,841	18,318	4,610	3,917	2,776	2,867	279	339
	18.3%	22.7%	15.4%	14.6%	11.3%	14.8%	7.5%	22.1%

Table 3: The distribution of patients with kidney issues across the ethnic groups

4.4.3 Preprocessing-III: Converting Data Types

This part of the data preprocessing focuses on making the data columns interpretable by the machine learning model. This includes typecasting of columns and encoding of categorical variables. On this part, some columns featured lists as entries that required additional processing and interpretation. Specifically, the code distinguishes between 4 types of columns:

- Float to integer: `['icd_code_count', 'emar_count', 'lab_count']`
- Object to datetime: `['admittime', 'edregtime', 'emar_charttime', 'lab_charttime']`
- String to one-hot encoding: `['admission_type', 'admission_location', 'insurance', 'marital_status', 'gender']`
- List of strings to one-hot encoding: `['emar_medications', 'emar_events', 'lab_flag', 'lab_comments', 'lab_priority']`

For the latter, data-specific algorithms are developed to extract the necessary data.

For `emar_medications`, the list of medications was converted into an integer column `emar_medicine_count` representing the amount of distinct medicine this patient is prescribed. Additionally, three boolean columns `emar_contains_insulin`, `emar_contains_ace_inhabitors`, `emar_contains_calcium_blockers` are added, indicating whether at least one of the patient's medicine is insulin, an ACE inhabitor or a calcium blocker respectively. These factors may contribute to kidney related problems, as they implicate underlying diseases that can contribute to kidney failure, according to Mayo Clinic [23].

For `emar_events`, the list was partially one-hot encoded based on whether the drug was administered or not. In most cases, the medication was administered. Therefore the exceptional cases may have an impact on the patient's treatment. This partial one-hot encoding resulted in the following columns: `emar_contains_not_given`, `emar_contains_not_flushed`, `emar_contains_stopped`, `emar_contains_not_started`.

The column `lab_flag` was replaced with the integer column `abnormal_lag_flags`, which contains the number of lab observations that were considered 'abnormal'. Abnormal doses of specific substances in the blood or urine, for example, may be indicators of health problems possibly related to the kidneys.

The column `lab_comments` is based on a similar reasoning and encodes whether the clinical laboratory scientists left a comment about the measurements. Most measurements do not get a comment, but rare or concerning cases do. Therefore, the new boolean column `has_lab_comment` represents if there is at least one comment present.

Finally, the column `lab_priority` is replaced by the boolean columns `lab_priority_stat`, `lab_priority_routine`. This indicates the purpose of the lab investigation. A patient could have both STAT and ROUTINE lab samples.

After these steps, each dataset consisted of the columns depicted in table 4.

Column	**Type**
hadm_id	int64
subject_id	int64
admittime	datetime64[ns]
edregtime	datetime64[ns]
anchor_age	int64
anchor_year	int64
icd_code_count	int64
emar_count	int64
emar_charttime	datetime64[ns]
lab_count	int64
lab_charttime	datetime64[ns]
has_kidney_issue	bool
admission_type_ew_emer.	bool
admission_type_observation_admit	bool
admission_type_eu_observation	bool
admission_type_surgical_same_day_admission	bool
admission_type_direct_emer.	bool
admission_type_urgent	bool
admission_type_direct_observation	bool
admission_type_elective	bool
admission_type_ambulatory_observation	bool
admission_location_information_not_available	bool
admission_location_internal_transfer_to_or_from_psych	bool
admission_location_walk-in/self_referral	bool
admission_location_clinic_referral	bool
admission_location_ambulatory_surgery_transfer	bool
admission_location_procedure_site	bool
admission_location_transfer_from_hospital	bool
admission_location_pacu	bool
admission_location_transfer_from_skilled_nursing_facility	bool
admission_location_emergency_room	bool
admission_location_physician_referral	bool
admission_location_unknown	bool
insurance_medicaid	bool
insurance_other	bool
insurance_medicare	bool

marital_status_married	bool
marital_status_single	bool
marital_status_widowed	bool
marital_status_divorced	bool
marital_status_unknown	bool
gender_m	bool
gender_f	bool
emar_medicine_count	int64
emar_contains_insulin	bool
emar_contains_ace_inhabitors	bool
emar_contains_calcium_blockers	bool
emar_contains_not_given	bool
emar_contains_not_flushed	bool
emar_contains_stopped	bool
emar_contains_not_started	bool
abnormal_lab_flags	int64
has_lab_comment	bool
lab_priority_stat	bool
lab_priority_routine	bool

Table 4: The ready-made data after the first three preprocessing steps

4.4.4 Preprocessing-IV: Splitting

In this step of the preprocessing, the ready-made data was split into a train set, validation set and test set for each ethnicity. Firstly, the data was sorted by `admittime`, which represents the date and the time the patient was admitted into the hospital. Additionally, the indexing was reset to match the new sorting. The first 80% of the admissions were assigned to the train set. The second 10% was assigned to the validation set. The final 10% was assigned to the test set. This was the preferred splitting algorithm, as it mimics real-life applications in which the model is trained on previous patients and applied to future patients.

4.5 Federated Forest

As the data is made ready to be interpretable by the machine learning models, there are two frameworks to evaluate. First local models will be trained for each ethnicity respectively and validated for optimal parameter configurations. Then these models contribute to forming the global model.

4.5.1 Local Models

Here, each preprocessed ethnicity represents a local model. The classification is performed based on Scikit-Learn's implementation of the random forest classifier [24]. This implementation allows for tuning the maximum depth and the number of trees the forest consists of. In the process of finding the optimal parameter configurations, the number of estimators and the maximum depth are therefore evaluated against the Area Under the Curve (AUC) of the model on the validation set. Ideally, the AUC is chosen as high as possible without showing signs of overfitting.

4.5.2 Global Model

The initial inspiration for this approach is based on the method proposed in the paper by Liu et al. [9], as mentioned before. However, as the problem statement in that work featured vertical federation, i.e., same

samples, different feature spaces, it does not reflect the problem in this work. To reiterate the approach, the feature selection training has been removed for the first formulation. As presented in algorithm 1, the algorithm follows a naive implementation that extends the contributing trees by adding the other random forest models. This aggregation-based format is deemed more applicable due to the horizontal similarity, i.e., the feature space is equal across all clients. In order to account for the inconsistency in model size, i.e., the number of decision trees in each local model, the addition of new trees is normalized. As a result, each model is represented evenly in the federated model.

Algorithm 1: Federated forests naive aggregation

Input: *Set of client models C*
Output: *Aggregated federated forest model*
max_trees $\leftarrow$ *# of trees in largest model* $c \in C$
for *Client* $c \in C$ **do**
  federated_model $\leftarrow$ normalize(*c.estimators, max_trees*)
end
Return: federated_model

4.6 Performance Metrics

The local models and the global model are evaluated on their accuracy, precision, recall and AUC. A brief motivation for each of these metrics is given below.

The performance of the federated model is obtained by referencing the global model configurations on the test sets of each ethnicity. This mimics real-life situations and sketches a complete image of the model performance per ethnicity.

4.6.1 Accuracy

The output of the models is a binary value, predicting whether the diagnosis of the patient is related to kidney issues. This has to be evaluated by comparing it to the actual diagnosis. The accuracy represents the fraction of correctly predicted values. Accuracy is a useful metric for determining the effectiveness of the model. Nevertheless, a final conclusion about model effectiveness cannot be made on accuracy alone, as it can mask class imbalances.

4.6.2 Precision

Precision is also known as the positive predictive value and represents how many positive predictions were correctly classified out of all positive predictions. This value is especially important when the cost of a false positive is relatively high, which can be the case in medical decision support, as further investigation into the potential kidney issue can become time consuming and require costly equipment.

4.6.3 Recall

The recall is also known as the sensitivity of the data and represents how many relevant elements were correctly classified. This value should be maximized when false negatives are relatively costly, which can definitely be the case for this specific medical application, as overlooked diseases can worsen if not treated in time. Thus, this metric plays an essential role in the evaluation of the model performance for this specific application.

4.6.4 Area Under the Curve

The Receiver Operating Characteristic (ROC) illustrates the true and false positive rates of the model predictions. The area under the Receiver Operating Characteristic, i.e., Area Under the Curve (AUC), provides a measure of the separability of the labels. Especially in the given problem domain where misclassification could be costly, the Area Under the Curve is an important metric.

5 Results and Analysis

5.1 Local Models

After tuning the `max_depth` and `n_estimators` in the random forest, the plots presented in figure 2 depict the optimal parameter configurations. Reading the configurations at their peak results in the data shown in table 5.

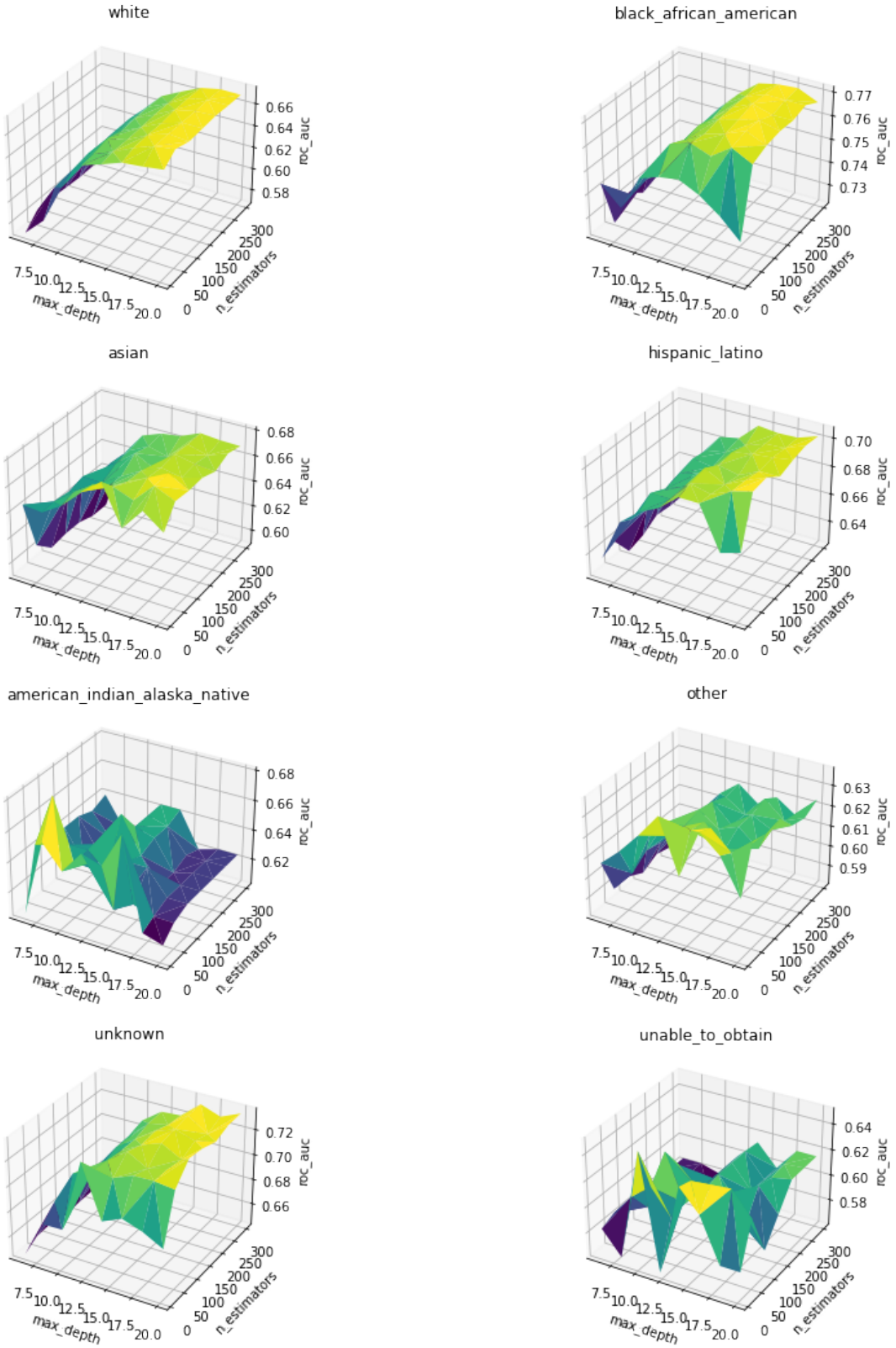

Figure 2: The area under the curve of the ethnic group against the maximum depth of the trees `max_depth` and the number of trees `n_estimators`

ethnicity	max_depth	n_estimators	accuracy
unknown	16.0	300.0	0.90
other	12.0	10.0	0.88
unknown	16.0	300.0	0.90
asian	18.0	100.0	0.90
hispanic_latino	20.0	300.0	0.88
black_african_american	18.0	100.0	0.86
unable_to_obtain	8.0	50.0	0.94
american_indian_alaska_native	8.0	50.0	0.87

Table 5: The optimal parameter configurations obtained by hyperparameter tuning

One observation that can be made from the results presented in figure 2 is that the smaller datasets have more inconsistent performances across parameter configurations compared to the larger datasets. Namely the 'American Indian / Alaska Native' and the 'unable to be obtained' ethnicities compared to the 'white' and the 'black / African American' ethnicities show this contrast evidently. Additionally, we can see from table 5 that these smaller datasets tend to have maximum performance at lower depths when using fewer estimators.

5.2 Local vs. Federated model metrics

Upon model testing with each model's respective local testing set, i.e., only samples of the ethnicity it was initially trained for, the results presented in table 6 were found for both the local and federated models. Table 7 depicts the equally weighted average of the federated model over all ethnicities.

Model	Ethnicity	Accuracy	Precision	Recall	AUC
Local	unknown	0.8852	0.6856	0.5097	0.7328
Federated	unknown	0.8749	0.6006	0.6299	0.7755
Local	white	0.8145	0.704	0.3957	0.6715
Federated	white	0.8156	0.6779	0.4454	0.6891
Local	other	0.832	0.5949	0.2843	0.6202
Federated	other	0.8439	0.6222	0.3952	0.6704
Local	asian	0.8715	0.6263	0.456	0.7026
Federated	asian	0.8752	0.6695	0.4093	0.6858
Local	hispanic_latino	0.8341	0.6885	0.3013	0.6338
Federated	hispanic_latino	0.8391	0.7116	0.3199	0.6439
Local	black_african_american	0.823	0.7173	0.6278	0.7645
Federated	black_african_american	0.7936	0.8097	0.3626	0.6643
Local	unable_to_obtain	0.9144	0.4667	0.2258	0.6012
Federated	unable_to_obtain	0.9144	0.4783	0.3548	0.6599
Local	american_indian_alaska_native	0.9156	0.75	0.7778	0.8613
Federated	american_indian_alaska_native	0.8961	0.7895	0.5556	0.762

Table 6: Resulting performance metrics of the local and federated models respectively on the test data

Model	Accuracy	Precision	Recall	AUC
Local	0.8612	0.6541	0.4473	0.6985
Federated	0.8566	0.6699	0.4341	0.6939

Table 7: Equally weighted average of performance metrics of the the federated models across all ethnicities

The first remark that can be made based on the results in table 6 is that the recall seems to be much lower than the other metrics for most of the ethnicities. Another very noticeable observation is the discrepancy in AUC performance between the local and federated models for the 'black / African American' and 'American Indian / Alaska native' ethnicities. In both cases, the model performs about 10% worse when aggregating.

From table 7, we can see that the federated model on average achieves an accuracy that is about 0.45% lower than the average local model accuracy. The precision is approximately 1.5% higher. The recall is about 1.3% lower and the AUC is approximately 0.5% lower.

6 Discussion

6.1 On Results

The results found that, on average, the aggregated and federated models perform nearly identically. However, when comparing specific local models and their respective test sets, a large performance gap is observed in some ethnic groups. In the two cases of the 'black / African American' and 'American Indian / Alaska native' ethnicities, the initial hypothesis is not rejected. Namely, the federated learning approach on monoethnic data did perform worse than the respective local model. For other ethnicities, such as the 'unable to obtain', 'other' and 'Hispanic / Latino' ethnicities, however, the federated model seemed to improve the predictive performance slightly.

Furthermore, we can see that most of the performance improvements are made on the precision of the model, while for medical applications such as this one, the recall is of relatively greater importance. The recall, however, decreases on average when a federated forest is applied.

Nevertheless, it must be noted that the obtained precisions, recalls and AUCs are rather low compared to other studies. For instance, a study conducted by Khalilia et al. obtained an average AUC of 88.79% over eight disease categories with a random forest [8]. This research performed significantly worse, with an average AUC of 69.85% across all local models. As a result, it should be accounted for that the model and feature qualities might have affected the results and will therefore be considered as a limitation to the study.

6.2 Limitations

In hindsight, the study could feature some risks and limitations influencing any results found. Mainly, the data does not reflect real-world situations as the data source is a single institute. Moreover, as described in the theory, the aspects contributing to the bias of the model might vary from our expectations and cloud any conclusions.

Furthermore, the model performance could be improved by adding more encodings and potentially field-specific information. For instance, `emar_medications` could be extended to categorize or one-hot encode more medicine. Potentially, deep-going medical knowledge is needed to select the medicine of importance if not all medicine can be encoded due to overhead. This was unfortunately outside of the scope of this research. Additionally, given more extensive parallel computing power, one ought to validate the hyperparameters of the models along a larger range to evaluate more optimal settings for training. Potentially cross-validation could be used for this, as not all ethnicities have enough registered administrations to give up 20% of their data for testing and validating.

Ultimately, the foundation of this research rests on the federated learning framework. As this work's federated forests algorithm 1 implementation follows a naive and simple model, it is bound to result in a limitation of the study.

6.3 Future Work

As covered in this paper, leveraging federated learning can hugely benefit the development of machine learning algorithms in clinical contexts. Due to the limitations mentioned above, further investigation ought

to be done in several fields.

First of all, the impact on monoethnic clinical institutes should be further investigated such that the limitations mentioned above are decreased. Moreover, it could be interesting to examine the effect of partial federation, where only a subset of the local models contributes to the global model and only a subset updates its model with the global parameters. With this, a more elaborate ethical debate should also take place if partial federation obtains promising results.

Secondly, the impact of polyethnic federation of models could be further looked into. Specifically, the feature and training pipelines offer opportunities for improvement. On that notion, the exploration of other varied data sources might offer more completeness which should aid in the feature engineering step. Lastly, the federated forest framework would require a dedicated study in terms of identifying an optimal formulation for the given problem domain and feature space.

6.4 Final Words

This research found that applying a federated approach to monoethnic data might hurt the performance of the model compared to its local model. Therefore, despite the numerous positives offered by the federated learning framework, the effectiveness of such a broad generalization should be questioned. Based on the briefness of this study, further in-depth research should be done to provide the necessary evidence that federated learning models are lucrative in clinical application.

References

[1] N. H. S. Stephen R. Pfohl, Agata Foryciarz, "An empirical characterization of fair machine learning for clinical risk prediction," *Journal of Biomedical Informatics*, vol. 113, p. 103621, 2021. doi: https://doi.org/10.1016/j.jbi.2020.103621. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1532046420302495

[2] J. Xu, B. S. Glicksberg, C. Su, P. Walker, J. Bian, and F. Wang, "Federated learning for healthcare informatics," *Journal of Healthcare Informatics Research*, vol. 5, no. 1, pp. 1–19, Mar 2021. doi: 10.1007/s41666-020-00082-4. [Online]. Available: https://doi.org/10.1007/s41666-020-00082-4

[3] S. Bharati, M. R. H. Mondal, P. Podder, and V. B. S. Prasath, "Federated learning: Applications, challenges and future directions," 2022. doi: 10.3233/HIS-220006. [Online]. Available: https://dl.acm.org/doi/abs/10.3233/HIS-220006

[4] M. A. Gianfrancesco, S. Tamang, J. Yazdany, and G. Schmajuk, "Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data," *JAMA Internal Medicine*, vol. 178, no. 11, pp. 1544–1547, 11 2018. doi: 10.1001/jamainternmed.2018.3763. [Online]. Available: https://doi.org/10.1001/jamainternmed.2018.3763

[5] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, 2019. doi: 10.1126/science.aax2342. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.aax2342

[6] D. R. Brendan McMahan, "Federated learning: Collaborative machine learning without centralized training data," 2017. [Online]. Available: https://ai.googleblog.com/2017/04/federated-learning-collaborative.html

[7] B. Pfitzner, N. Steckhan, and B. Arnrich, "Federated learning in a medical context: A systematic literature review," *ACM Trans. Internet Technol.*, vol. 21, no. 2, jun 2021. doi: 10.1145/3412357. [Online]. Available: https://doi.org/10.1145/3412357

[8] C. S. . P. M. Khalilia, M., "Predicting disease risks from highly imbalanced data using random forest," *BMC Medical Informatics and Decision Making*, vol. 11, 2011. doi: 10.1186/1472-6947-11-51. [Online]. Available: https://doi.org/10.1186/1472-6947-11-51

[9] Y. Liu, Y. Liu, Z. Liu, J. Zhang, C. Meng, and Y. Zheng, "Federated forest," 2019, article in review. [Online]. Available: https://arxiv.org/pdf/1905.10053.pdf

[10] Q. Li, Y. Cai, Y. Han, C. M. Yung, T. Fu, and B. He, "FedTree: A fast, effective, and secure tree-based federated learning system," jun 2022. [Online]. Available: https://github.com/Xtra-Computing/FedTree/blob/main/FedTree_draft_paper.pdf

[11] J. Kukacka, V. Golkov, and D. Cremers, "Regularization for deep learning: A taxonomy," *CoRR*, vol. abs/1710.10686, 2017. [Online]. Available: http://arxiv.org/abs/1710.10686

[12] R. Mohammed, J. Rawashdeh, and M. Abdullah, "Machine learning with oversampling and undersampling techniques: Overview study and experimental results," in *2020 11th International Conference on Information and Communication Systems (ICICS)*, 2020. doi: 10.1109/ICICS49469.2020.239556 pp. 243–248.

[13] C. M. Thwal, K. Thar, Y. L. Tun, and C. S. Hong, "Attention on personalized clinical decision support system: Federated learning approach," in *2021 IEEE International Conference on Big Data and Smart Computing (BigComp)*, 2021. doi: 10.1109/BigComp51126.2021.00035 pp. 141–147.

[14] I. Dayan, H. R. Roth, A. Zhong, A. Harouni, A. Gentili, A. Z. Abidin, A. Liu, A. B. Costa, B. J. Wood, C.-S. Tsai, C.-H. Wang, C.-N. Hsu, C. K. Lee, P. Ruan, D. Xu, D. Wu, E. Huang, F. C. Kitamura, G. Lacey, G. C. de Antônio Corradi, G. Nino, H.-H. Shin, H. Obinata, H. Ren, J. C. Crane, J. Tetreault, J. Guan, J. W. Garrett, J. D. Kaggie, J. G. Park, K. Dreyer, K. Juluru, K. Kersten, M. A. B. C. Rockenbach, M. G. Linguraru, M. A. Haider, M. AbdelMaseeh, N. Rieke, P. F. Damasceno, P. M. C. e Silva, P. Wang, S. Xu, S. Kawano, S. Sriswasdi, S. Y. Park, T. M. Grist, V. Buch, W. Jantarabenjakul, W. Wang, W. Y. Tak, X. Li, X. Lin, Y. J. Kwon, A. Quraini, A. Feng, A. N. Priest, B. Turkbey, B. Glicksberg, B. Bizzo, B. S. Kim, C. Tor-Díez, C.-C. Lee, C.-J. Hsu, C. Lin, C.-L. Lai, C. P. Hess, C. Compas, D. Bhatia, E. K. Oermann, E. Leibovitz, H. Sasaki, H. Mori, I. Yang, J. H. Sohn, K. N. K. Murthy, L.-C. Fu, M. R. F. de Mendonça, M. Fralick, M. K. Kang, M. Adil, N. Gangai, P. Vateekul, P. Elnajjar, S. Hickman, S. Majumdar, S. L. McLeod, S. Reed, S. Gräf, S. Harmon, T. Kodama, T. Puthanakit, T. Mazzulli, V. L. de Lavor, Y. Rakvongthai, Y. R. Lee, Y. Wen, F. J. Gilbert, M. G. Flores, and Q. Li, "Federated learning for predicting clinical outcomes in patients with covid-19," *Nature Medicine*, vol. 27, no. 10, pp. 1735–1743, Oct 2021. doi: 10.1038/s41591-021-01506-3. [Online]. Available: https://doi.org/10.1038/s41591-021-01506-3

[15] B. Murdoch, "Privacy and artificial intelligence: challenges for protecting health information in a new era," *BMC Medical Ethics*, vol. 22, no. 1, p. 122, Sep 2021. doi: 10.1186/s12910-021-00687-3. [Online]. Available: https://doi.org/10.1186/s12910-021-00687-3

[16] X. Ferrer, T. v. Nuenen, J. M. Such, M. Coté, and N. Criado, "Bias and discrimination in ai: A cross-disciplinary perspective," *IEEE Technology and Society Magazine*, vol. 40, no. 2, pp. 72–80, 2021. doi: 10.1109/MTS.2021.3056293

[17] J. S. Mill, *Utilitarianism*, s. Parker and Bourn, Eds. London: Oxford University Press UK, 1861.

[18] T. Hellström, V. Dignum, and S. Bensch, "Bias in machine learning what is it good (and bad) for?" vol. abs/2004.00686, 2020. doi: 10.48550/2004.00686. [Online]. Available: https://doi.org/10.48550/arXiv.2004.00686

[19] A. Johnson, L. Bulgarelli, T. Pollard, S. Horng, L. A. Celi, and R. Mark, "Mimic-iv," 2022, last updated: 2022. [Online]. Available: https://physionet.org/content/mimiciv/2.0/

[20] B. Wolford, "What is gdpr, the eu's new data protection law?" 2022. [Online]. Available: https://gdpr.eu/what-is-gdpr/

[21] L. Center for State, Tribal and P. H. L. P. Territorial Support, "Health insurance portability and accountability act of 1996 (hipaa)," 6 2022. [Online]. Available: https://www.cdc.gov/phlp/publications/topic/hipaa.html

[22] A. Schavemaker and A. Mo, "Mimic-iv," 2023, github repository. [Online]. Available: https://github.com/AnnikaLarissa/MIMIC-IV

[23] M. C. Staff, 06 2022. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/kidney-failure/symptoms-causes/syc-20369048

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

*

* All references valid and last accessed 6 October 2022