Interpretable Federated Learning via Neural Additive Models

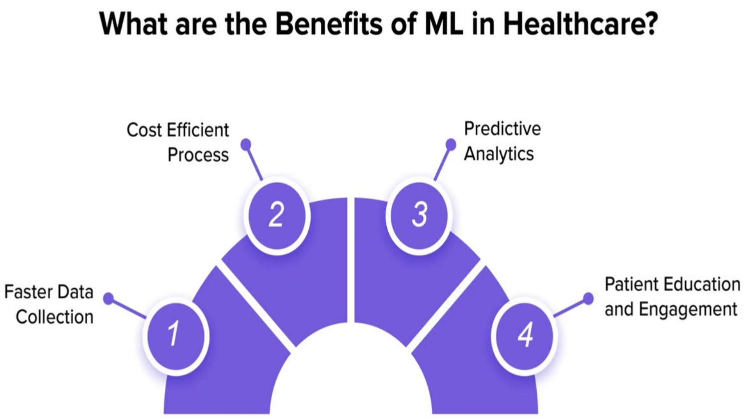
Sree Bhargavi Balija
University of California, San Diego

Motivation

- AI in drug discovery (DeepChem)
- Biomedical data visualizations (AlphaFold2)
- Improved diagnosis
- Offering accurate information
- Disease prediction and Enhanced care

Challenges:-

➢ Interpretability and Explainability
➢ Privacy
➢ Insufficient data availability
➢ Coronary Heart disease



Dataset

- Cardiovascular study is done on residents of the town of Framingham, Massachusetts.
- The classification goal is to predict whether the patient has 10-year risk of future coronary heart disease (CHD).
- It includes over 4,000 records and 15 attributes.

Vital features:-

Prevalent Hyp

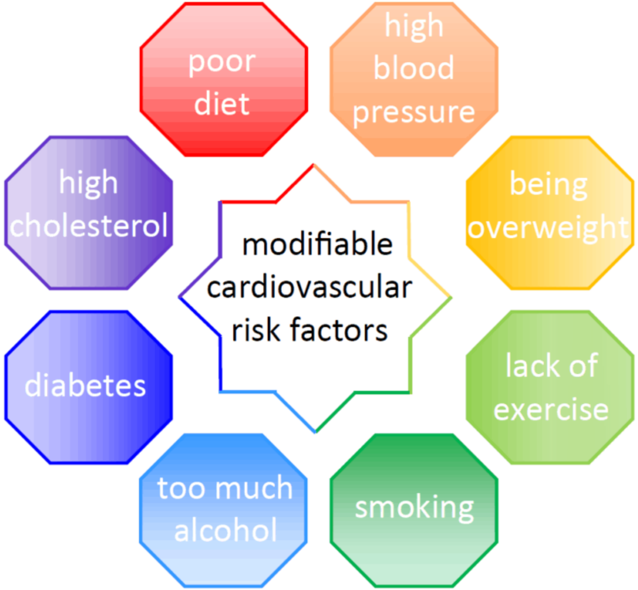
Heart Rate

Glucose level

Sys BP: systolic blood pressure

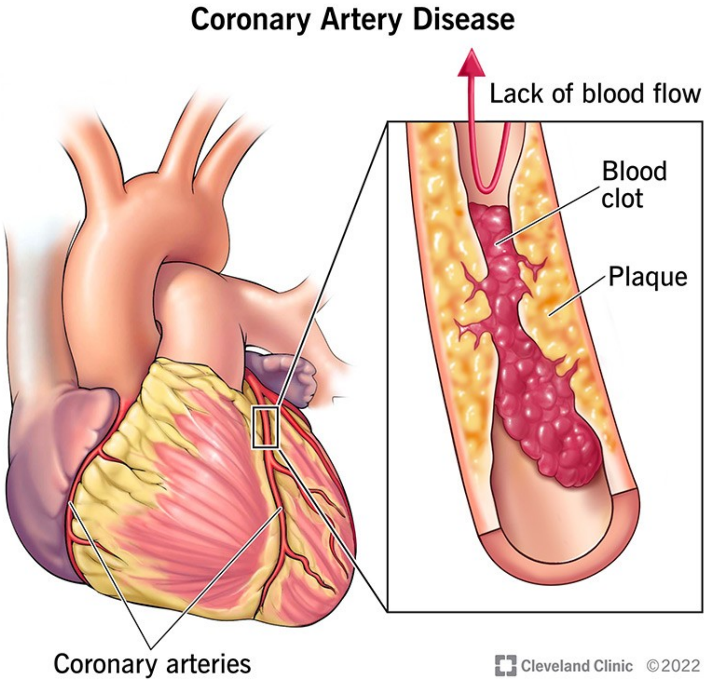
Dia BP: diastolic blood pressure

Tot Chol: total cholesterol level



Goal

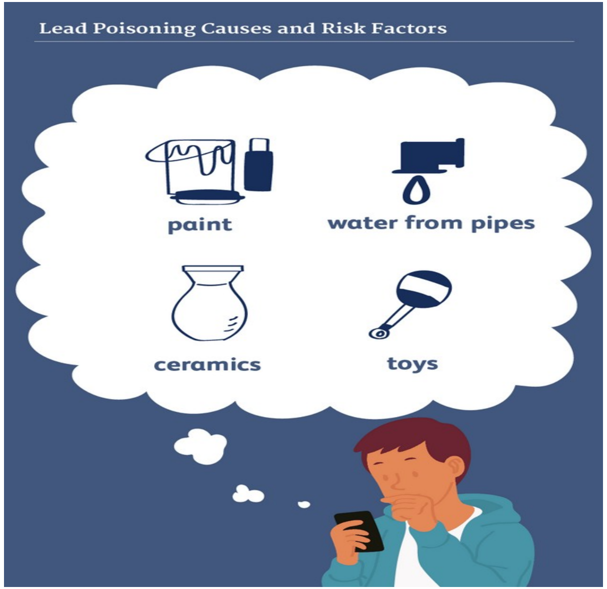
- Develop an Fed-Nam model to predict if a patient has a 10-Year Risk of future coronary heart disease (CHD) & Identify most relevant risk factors for heart disease
- Comparative Analysis of Interpretable Fnams models vs State of the Art Models
- Input feature functions

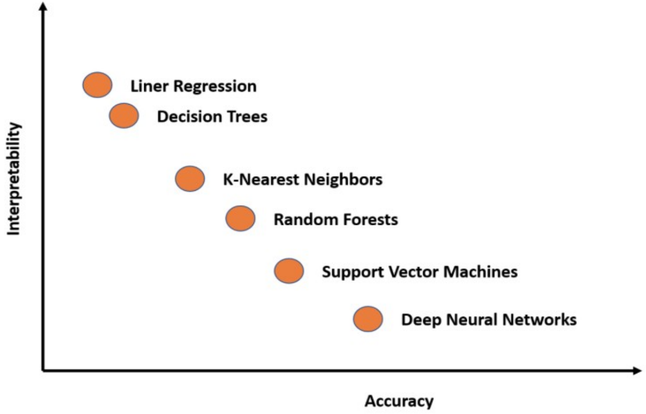


What is Interpretability?

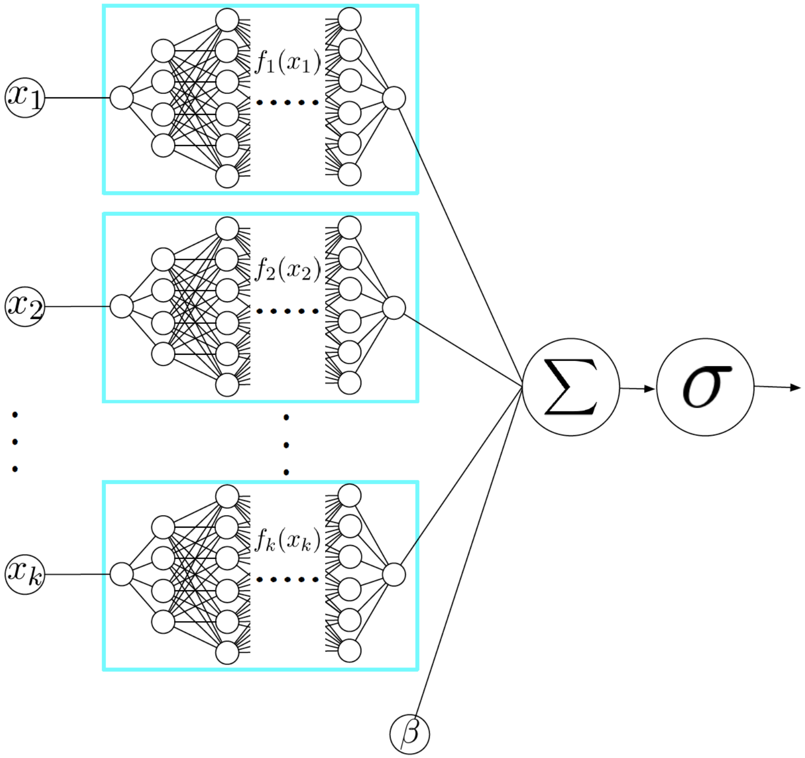
- Interpretability : Ability to explain how an ML model is making predictions and which factors are driving those predictions
- Crucial for complex models that are difficult to understand, for example LLM, DNNs
- Example
- Latest techniques
  - RMechRP – Radical reactions (link)
  - Discover – Interpretable technique for vision tasks
  - Interpretability in Gated Neural ODEs
  - Neural additive models
- Traditional techniques
  - Morris sensitivity analysis
  - Lime and shape

Source:- AI for Interpretable Chemistry: Predicting Radical Mechanistic Pathways via Contrastive Learning





What are Neural Additive Models?



- Every feature is handled by a different neural network
- We aggregate the final learned function for every feature & pass through a sigmoid layer to generate final prediction
- All networks are trained concurrently using backpropagation
- Can be trained at massive scale on GPU's

$$g(\mathbb{E}[y]) = \beta + f_1(x_1) + f_2(x_2) + \cdots + f_K(x_K)$$

Federated learning?

Collaboratively train the models across devices

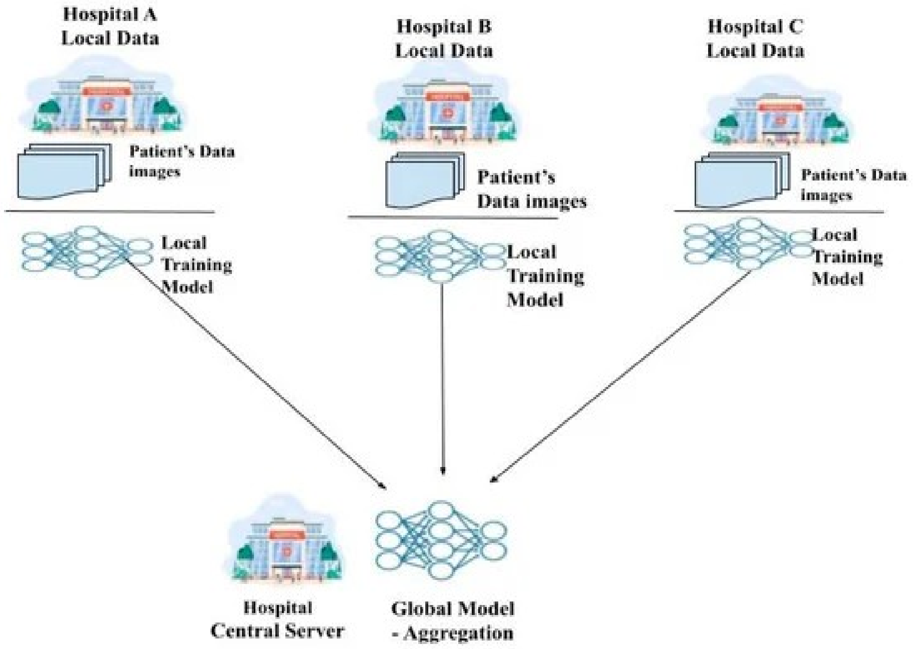
- Privacy
- Robustness
- Effective and data-driven healthcare solutions

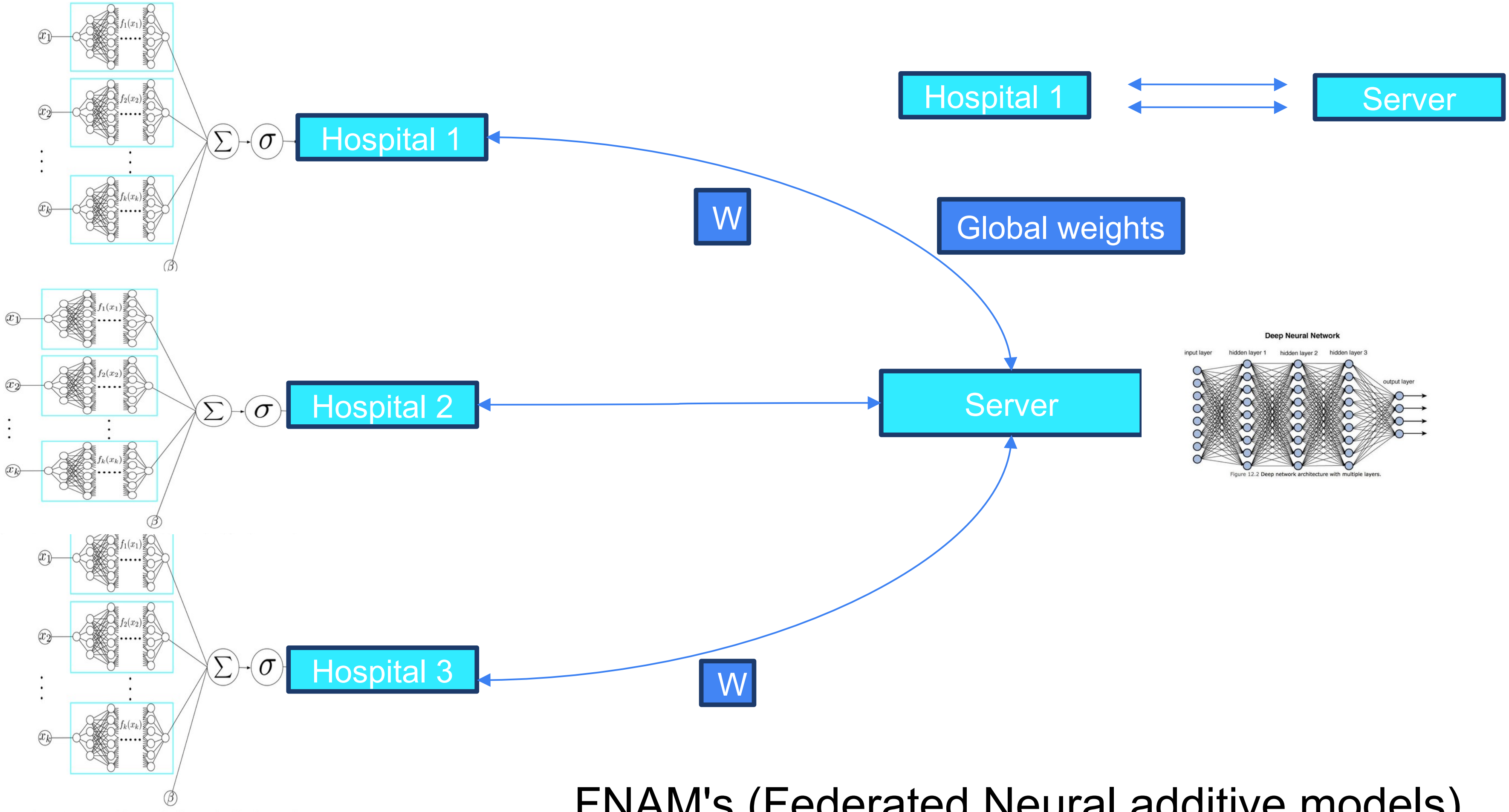
Challenges:-

- Data heterogeneity
- Convergence

Latest Techniques:-

- FedAvg
- FedScaffold

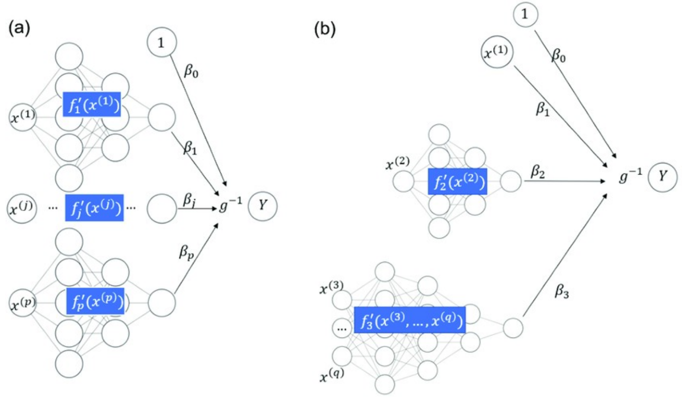




FNAM's (Federated Neural additive models)

Implementation

- Innovative architecture
- Extends NAMs to a federated learning context
- Network optimization problem
- Each input variable is handled by a separate neural network
- Maintains individual neural networks for each feature,
- Model that balances interpretability and accuracy.
- Preserving the interpretability of additive models while leveraging the representational power of neural networks for higher predictive performance.
- Relationships between each input feature to the output



Methods (Network optimization problem)

$$\theta_i^{(t+1)} = \theta_i^{(t)} - \eta \nabla L(\theta_i^{(t)})$$

$$\theta^{(t+1)} = \sum_{i=1}^{N} \frac{n_i}{n} \theta_i^{(t+1)}$$

Optimization at First stage

$g(E\,[y\text{client1}]) = \beta + f11(x1) + f12(x2) + \cdots + f1K\,(xK\,)$

$g(E\,[y\text{client2}]) = \beta + f21(x1) + f22(x2) + \cdots + f2K\,(xK\,)$

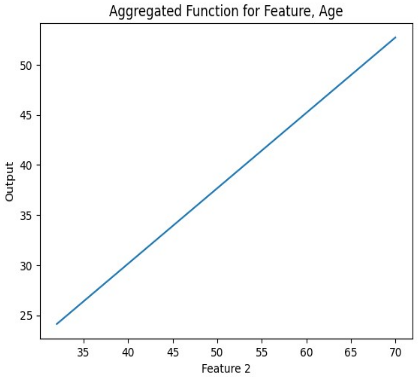
$g(E\,[y\text{client3}]) = \beta + f31(x1) + f32(x2) + \cdots + f3K\,(xK\,)$

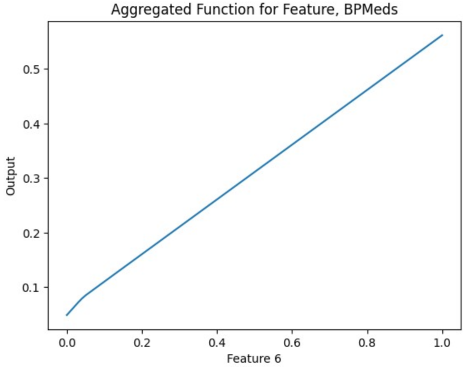
Optimization at Second stage

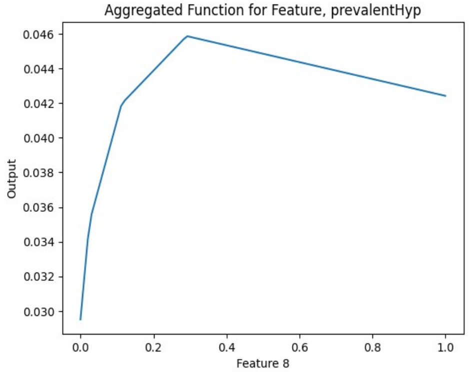
$f(x1\ ) = f11\ (x1) + f21\ (x1) + f31\ (x1) + \cdots + fn1\ (x1)/n$

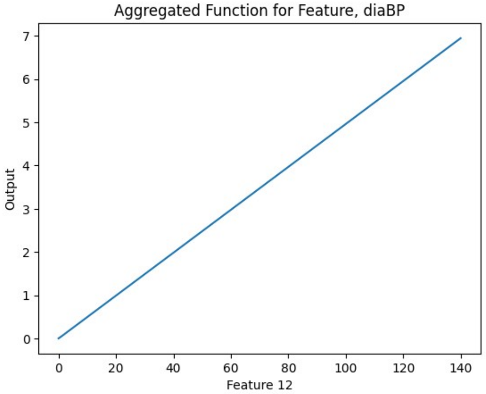
$f(x2\ ) = f12\ (x2) + f22\ (x2) + f32\ (x2) + \cdots + fn2\ (x2)/\mathrm{n}$

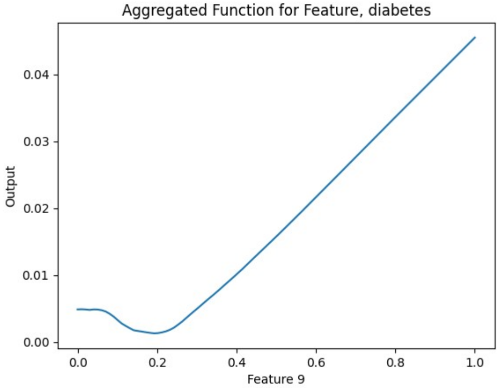
Results

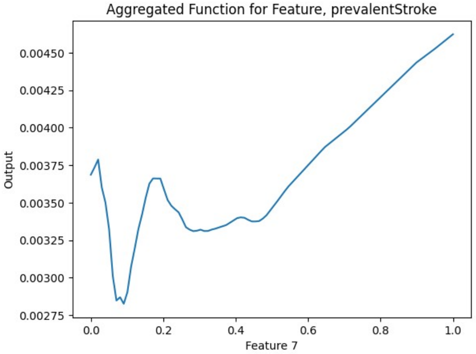

Aggregated Function for Feature, Age
Output
25 30 35 40 45 50
Feature 2
35 40 45 50 55 60 65 70


Aggregated Function for Feature, BPMeds
Output
0.1 0.2 0.3 0.4 0.5
Feature 6
0.0 0.2 0.4 0.6 0.8 1.0


Aggregated Function for Feature, prevalentHyp
Output
0.030 0.032 0.034 0.036 0.038 0.040 0.042 0.044 0.046
Feature 8
0.0 0.2 0.4 0.6 0.8 1.0


Aggregated Function for Feature, diaBP
Output
0 1 2 3 4 5 6 7
Feature 12
0 20 40 60 80 100 120 140


Aggregated Function for Feature, diabetes
Output
0.00 0.01 0.02 0.03 0.04
Feature 9
0.0 0.2 0.4 0.6 0.8 1.0

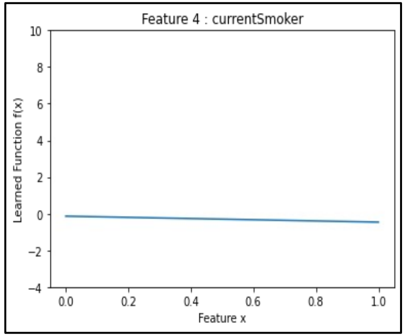

Aggregated Function for Feature, prevalentStroke
Output
0.00275 0.00300 0.00325 0.00350 0.00375 0.00400 0.00425 0.00450
Feature 7
0.0 0.2 0.4 0.6 0.8 1.0

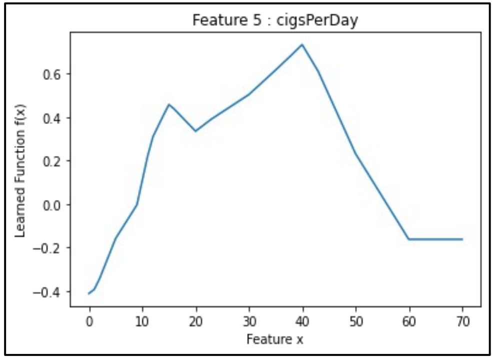
Auc scores

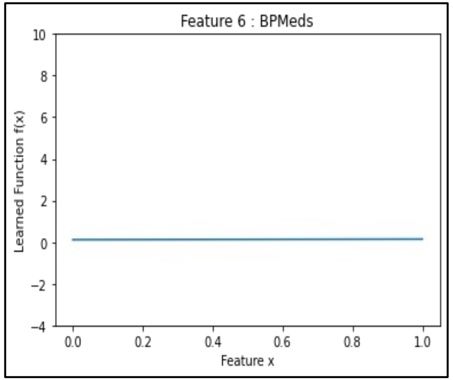
Method	Train	Val	Test
Logistic Regression	0.7	N/A	0.72
Fed-DNN's	0.852	0.8192	0.885
Fed-Nam's	0.82	0.84	0.82

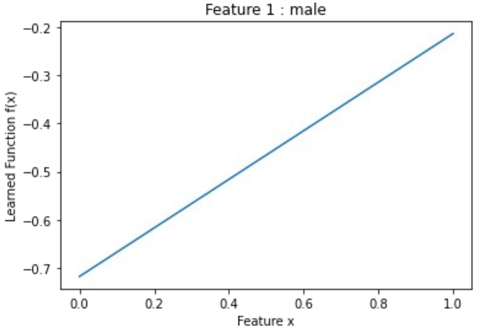
A little loss in accuracy of FedNams compared to Federated deep neural networks.

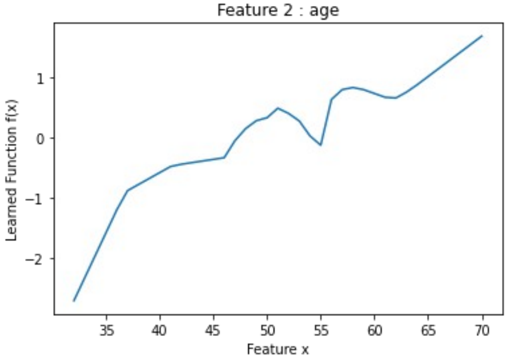
Results from one hospital

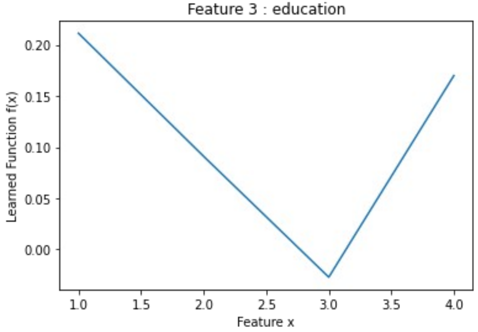












Summary

- F-NAM's allow us to train state of art GAMS with deep neural nets
  - Accurate
  - Interpretable
  - Differentiable as well as flexible
- Features like resting electrocardiographic results, Sex, Cholesterol are positively correlated with output
- Features like fasting blood sugar, Age are negatively correlated with risk.
- Blood glucose, age are highly correlated factors for diabetes dataset
- Building easy to use toolkits so everyone can train FNAM's
- Exploring other ways to combine FNAMS with CNN'S

Ongoing and future work

- Interpretability of large language models
- Extending IID setup to Non IID setup
- Robustness to Diverse Datasets
  - Test the robustness of FNAMS across a broader range of datasets
  - Across different data distributions.
- Mixed precision quantization of Large language models
- Image segmentation of colon cancer images using yolo v8
- Glaucoma detection using deep learning models

Questions ?

sbalija@ucsd.edu