JOHANNES KEPLER
UNIVERSITY LINZ

Experiments with and Development of Neural Additive Models as an Interpretable ML Method

Seminar in AI
Under the supervision of

Prof. Marc Streit
Christian Steinparz

INSTITUTE OF
COMPUTER GRAPHICS

Presented by

Prasil Adhikari
(k12049801)

23 January 2023

Previous Presentation

- Interpretable ML and its Importance
- Generalized Additive Models (GAM)
- Explainable Boosting Machines (EBM)
- **Neural Additive Models (NAM)**

Neural additive models (NAM) [1] Agarwal et al (2021)

as interpretable as Logistic regression, while achieving similar accuracy as standard DNNs.

Combination of Inherent intelligibility of GAMs and expressivity of DNNs

NAM

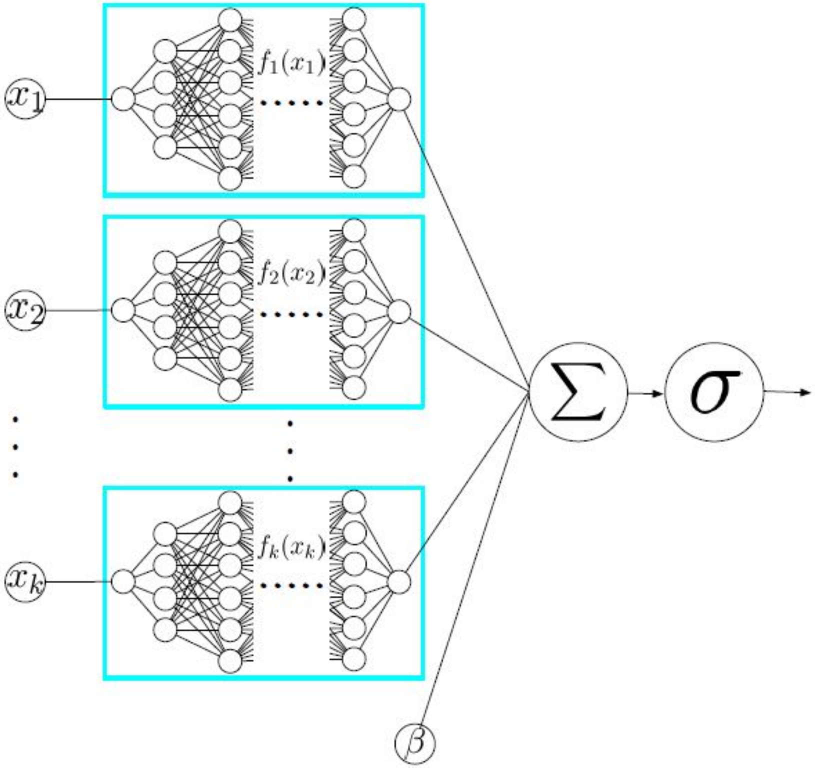
a NAM learns separate DNN subnets for each input feature

Learned in parallel and additively combined

Model is fully interpretable since each learned subnets can be visualized

Are differentiable and can be trained via backpropagation

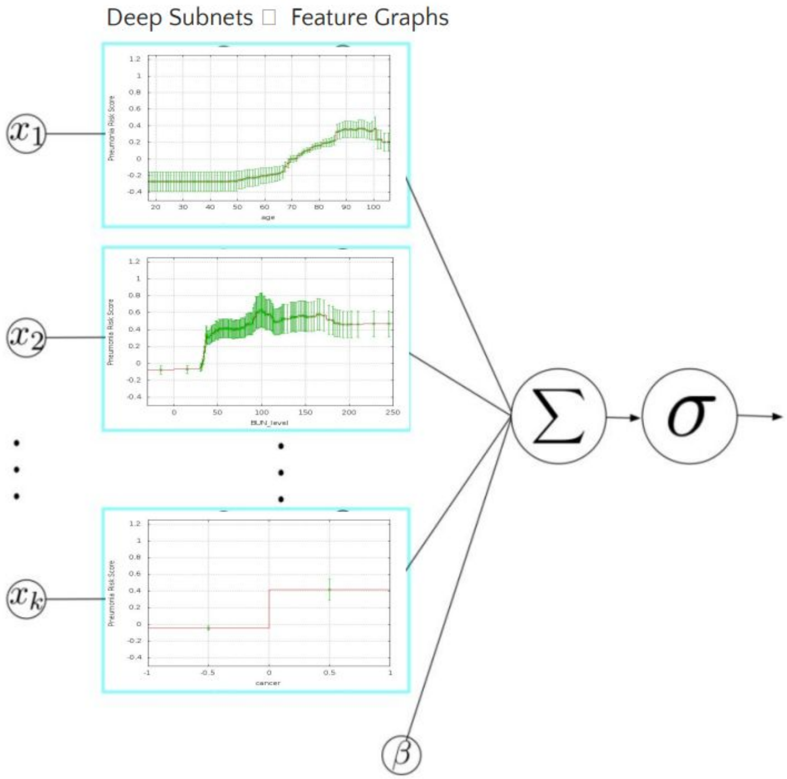
Sigmoid at output used for classification, not regression



NAM architecture for Binary Classification
[1] Agarwal et al. (2021)

NAM

After training, subnets can be replaced with feature graphs



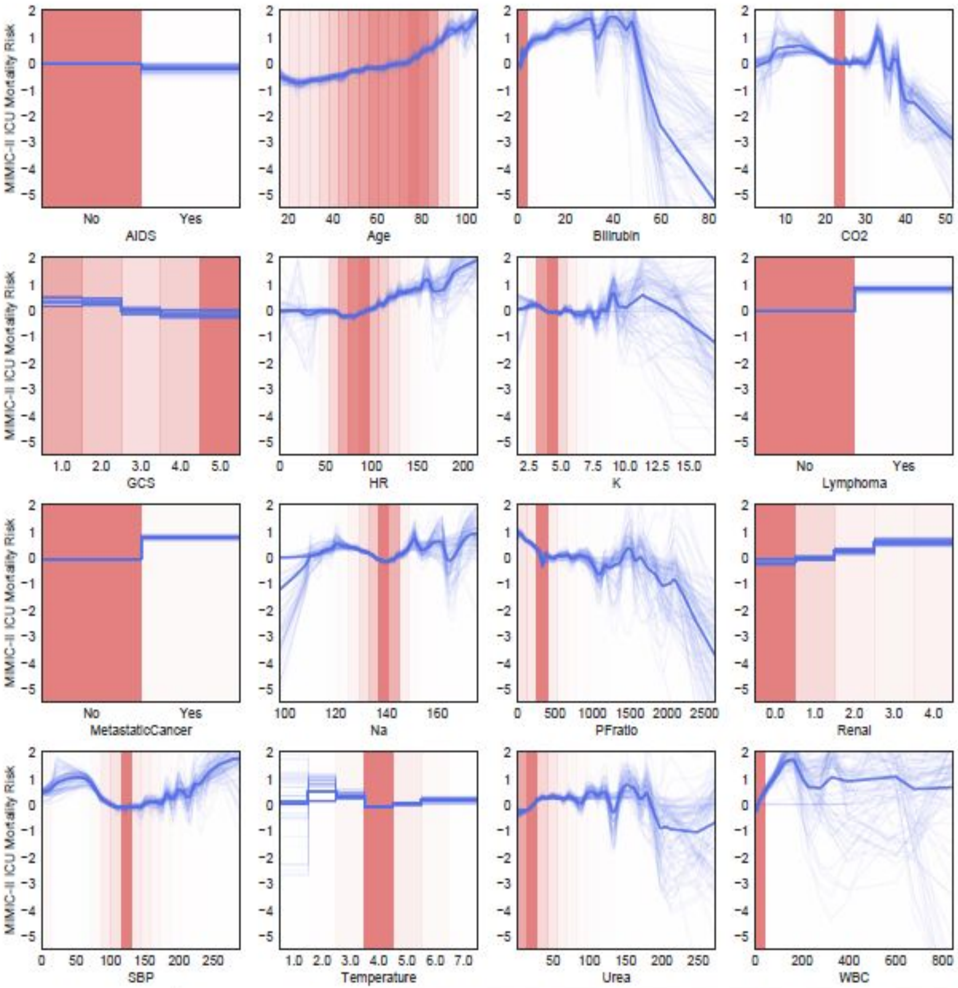
[2] NAMs Agarwal et al. (2021)

NAM

Self-interpretable - plotting the individual independent shape function which parametrizes individual feature

The plots are an exact description o how NAMs compute a prediction

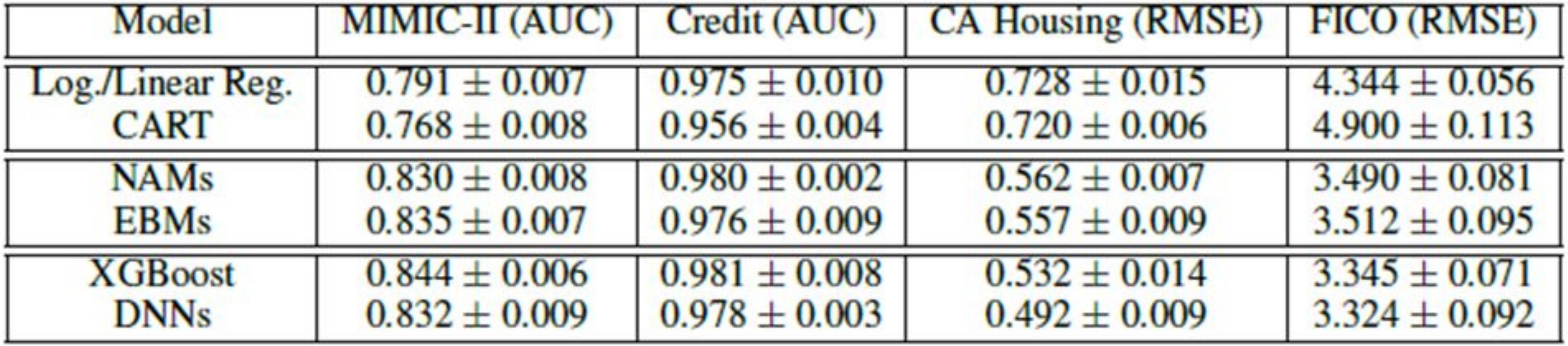
Note: NAMs are not causal models



[1] Shape functions on MIMIC II Dataset

NAM

an competitive edge to EBM with its differentiability and composability.



Model	MIMIC-II (AUC)	Credit (AUC)	CA Housing (RMSE)	FICO (RMSE)
Log./Linear Reg.	0.791 ± 0.007	0.975 ± 0.010	0.728 ± 0.015	4.344 ± 0.056
CART	0.768 ± 0.008	0.956 ± 0.004	0.720 ± 0.006	4.900 ± 0.113
NAMs	0.830 ± 0.008	0.980 ± 0.002	0.562 ± 0.007	3.490 ± 0.081
EBMs	0.835 ± 0.007	0.976 ± 0.009	0.557 ± 0.009	3.512 ± 0.095
XGBoost	0.844 ± 0.006	0.981 ± 0.008	0.532 ± 0.014	3.345 ± 0.071
DNNs	0.832 ± 0.009	0.978 ± 0.003	0.492 ± 0.009	3.324 ± 0.092

Single-task learning NAM results.
High AUCs and lower RMSEs are better

[1] Agarwal et al. (2021)

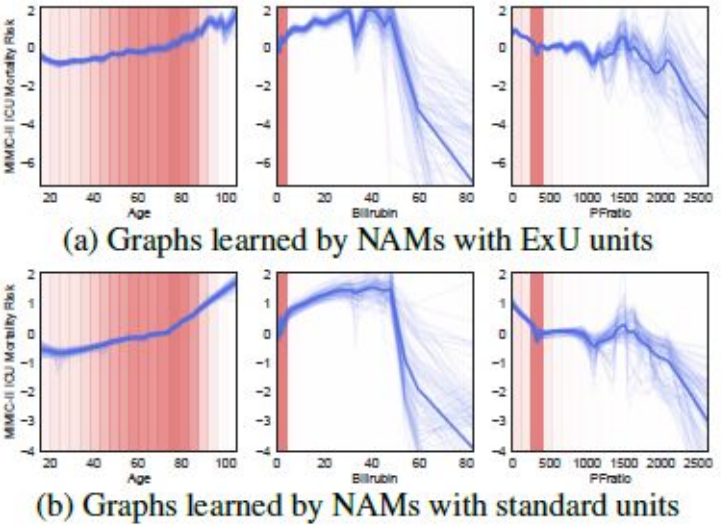
NAM

Modeling jagged shape functions is required to learn accurate additive models

A special activation function, exp-centered (ExU) hidden units

$$h(x) = f\,(e^{w} * (x - b))$$

slope of activation function can be very steep so small changes in input => large changes in output



(a) Graphs learned by NAMs with ExU units

(b) Graphs learned by NAMs with standard units

Experiments on MIMIC II Dataset

[1] Agarwal et al. (2021)

NAM

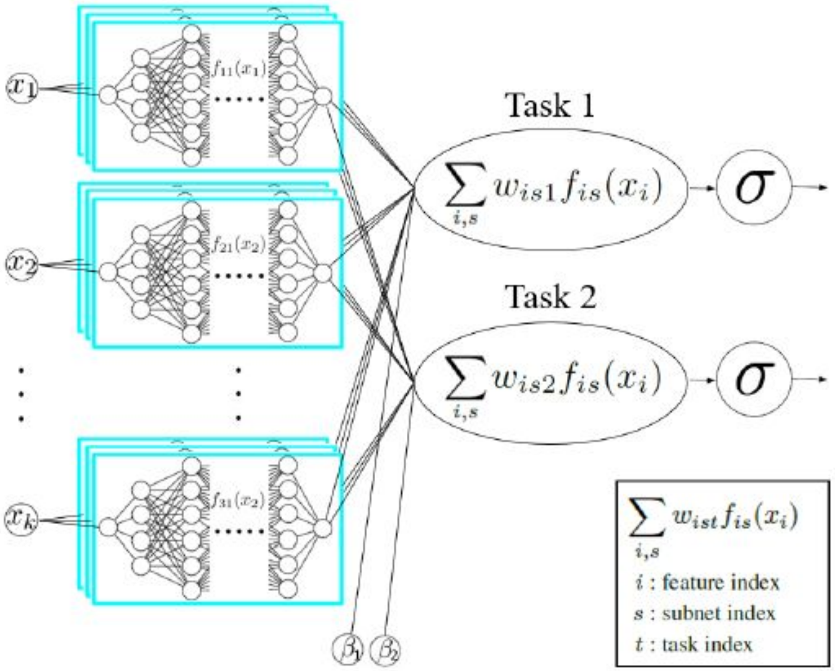
Multitask Learning

the composability of NNs makes it easy to train multiple subnets per feature

The model jointly learns a task specific weighted sum over their outputs

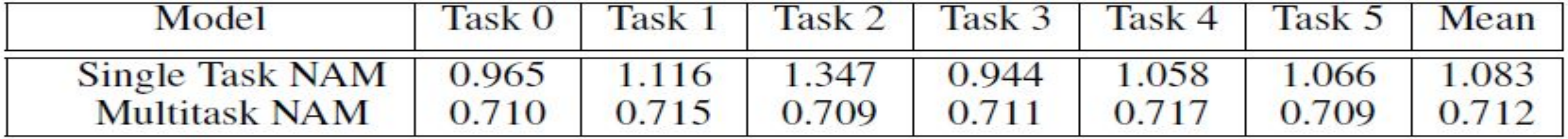
The outputs corresponding to each task are summed and a bias is added to obtain the final prediction score

learn different feature representations for each task while preserving the intelligibility and modularity

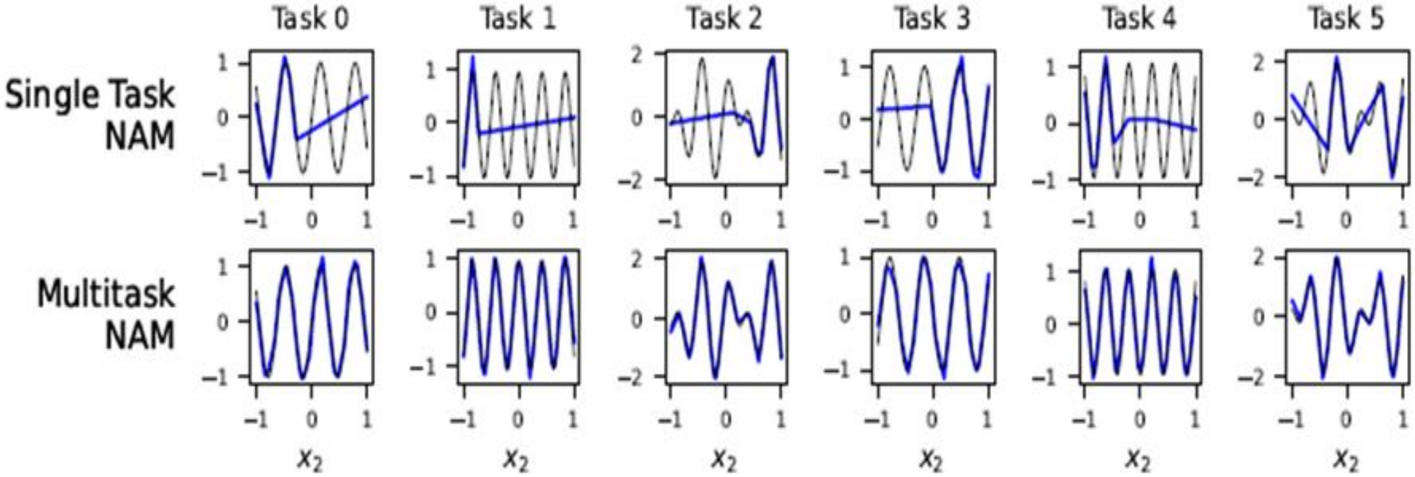


[1] Agarwal et al. (2021)

Multitask Learning on Synthetic Data



Model	Task 0	Task 1	Task 2	Task 3	Task 4	Task 5	Mean
Single Task NAM	0.965	1.116	1.347	0.944	1.058	1.066	1.083
Multitask NAM	0.710	0.715	0.709	0.711	0.717	0.709	0.712

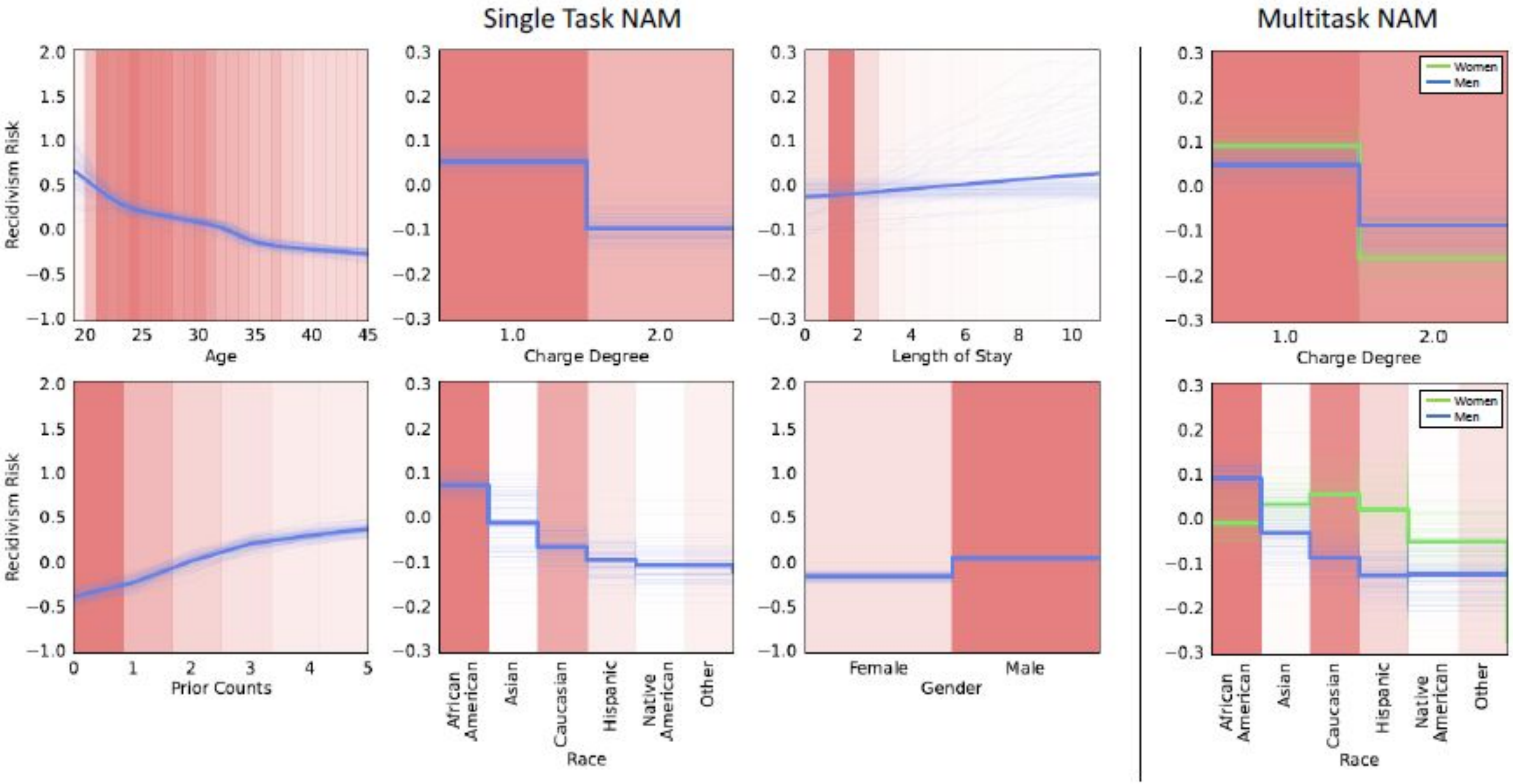


MSE for STL and MTL NAMs on synthetic data. Average of 20 runs. Lower MSEs are better. [1] Agarwal et al. (2021)

MLT NAMs achieve MSE 34% lower than SLT, and at least 25% lower on each individual task

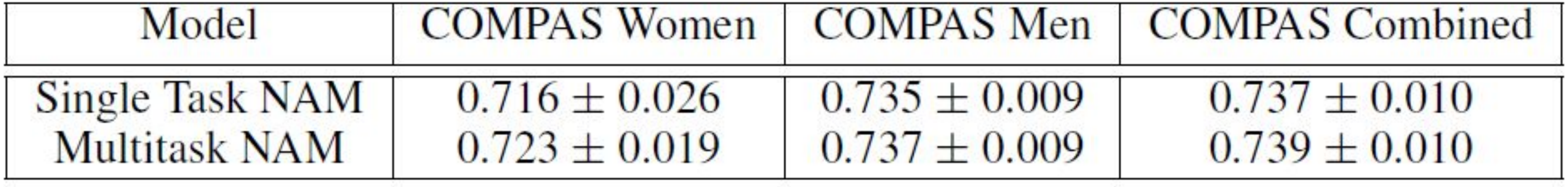
Multitask Learning on COMPAS Recidivism Data

(COMPAS is a proprietary score developed to predict recidivism risk, which is used to inform bail, sentencing and parole decisions and has been the subject of scrutiny for racial bias)



[1] Agarwal et al. (2021)

Multitask Learning on COMPAS Recidivism Data



Model	COMPAS Women	COMPAS Men	COMPAS Combined
Single Task NAM	0.716 ± 0.026	0.735 ± 0.009	0.737 ± 0.010
Multitask NAM	0.723 ± 0.019	0.737 ± 0.009	0.739 ± 0.010

ROC AUC for multi task and single task NAMs on COMPAS dataset, broken down by gender. Higher AUCs are better.

[1] Agarwal et al. (2021)

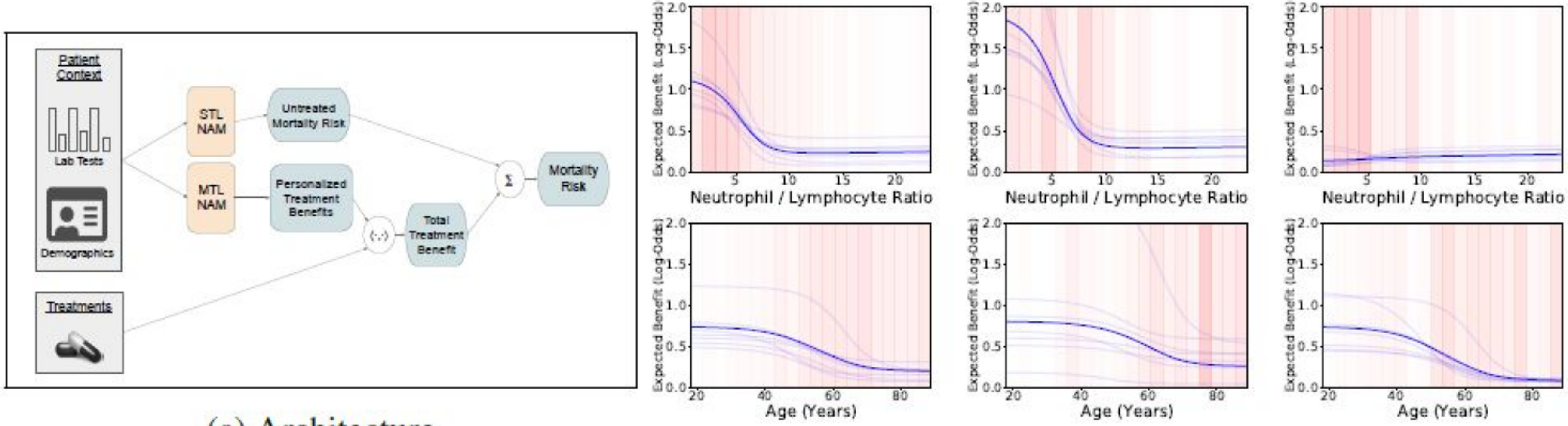
Transparency and modularity of NAM allows to detect unanticipated biases in data

makes it easier to correct the bias in the learned model

Differentiability of NAM

Allows to train more complex interpretable models

NAMs are the only nonlinear GAM suitable for this application because of their differentiability



(a) Architecture

(b) Anti-Coagulants

(c) NSAIDs

(d) Glucocorticoids

Estimating personalized treatment benefits for Covid19 patients
[1] Agarwal et al. (2021)

NAM

Globally & intrinsically interpretable

Model-specific (Linear Regression, Splines, Random Forests, and NNs)

Advantages

Complete description of the model

Allows the application of any NN architecture

Multitask prediction

Differentiability

Disadvantages

No Feature Interactions, even if, interpretation becomes difficult

Suffering a little loss in prediction accuracy when applied to tabular data

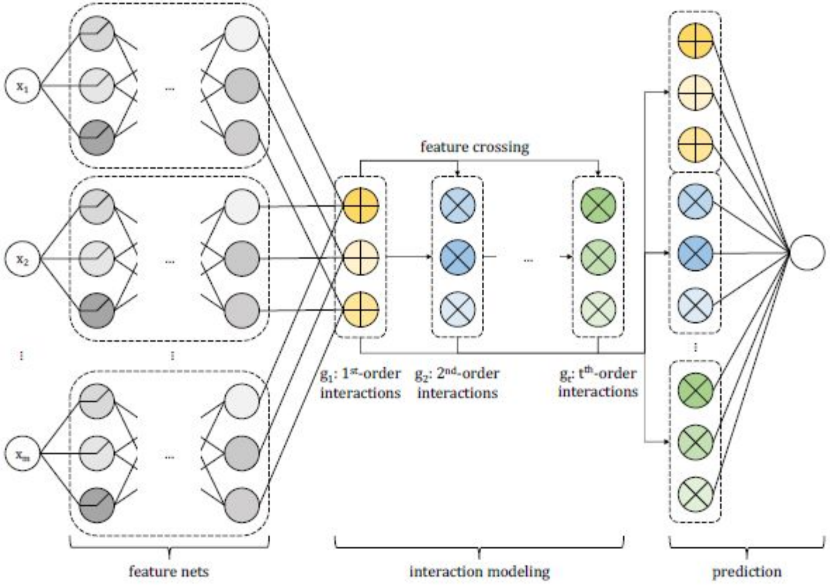
Higher-order Neural Additive Models (HONAM)

Can model arbitrary orders of feature interactions

redefined the existing NAM and applied CrossNe model high-order feature interactions

proposed the ExpDive unit, which overcomes the limitations of the ExU unit, to effectively learn sha shape functions

$$h(x) = \sigma((x - b) * (\exp(W) - \exp(-W)))$$


x_1
x_2
x_m
feature crossing
g_1: 1st-order interactions
g_2: 2nd-order interactions
g_t: tth-order interactions
feature nets
interaction modeling
prediction

Architecture of HONAM
[3] Minkyu et al. (2022)

Current Research

SurvNAM: The machine learning survival model explanation
[Lev V. Utkin, Egor D. Satyukov, Andrei V. Konstantinov] (December 2021)

Neural Additive Models for Nowcasting
[Wonkeun Jo, Dongil Kim] (May 2022)

Neural Additive Models for Explainable Heart Attack Prediction
[Ksenia Balabaeva, Sergey Kovalchuk] (June 2022)

Higher-order Neural Additive Models: An Interpretable Machine Learning Model with Feature Interactions
[Minkyu Kim, Hyun-Soo Choi, Jinho KIm] (Sep 2022)

Next Steps

- Gathering more insights on higher order NAMs
- Practical Work in AI: Experimentation on NAM with Image data and further integration

References

[1] Agarwal, R., Melnick, L., Frosst, N., Zhang, X., Lengerich, B., Caruana, R., & Hinton, G. (2021, October 24). *Neural Additive Models: Interpretable machine learning with neural nets*. arXiv.org. Retrieved from https://arxiv.org/abs/2004.13912

[2] Agarwal et al. NAMs, GAMs discussed in Molnar Chapter 5.3, (2021)

[3] Kim, M., Choi, H.-S., & Kim, J. (2022, September 30). *Higher-order Neural Additive Models: An interpretable machine learning model with feature interactions*. arXiv.org. Retrieved from https://arxiv.org/abs/2209.15409v1

[4] Nori, H., Jenkins, S., Koch, P., & Caruana, R. (2019, September 19). *InterpretML: A unified framework for machine learning interpretability*. arXiv.org., from https://arxiv.org/abs/1909.09223v1

[5] Oleszak, M. (2022, January 27). *Explainable boosting machines*. Medium. Retrieved, from https://pub.towardsai.net/explainable-boosting-machines-c71b207231b5

[6] Molnar Christoph “Interpretable Machine Learning” (2022) https://christophm.github.io/interpretable-ml-book/

[7] Dallanoce, F. (2022, August 20). *Explainable AI: A comprehensive review of the main methods*. Medium. Retrieved from https://medium.com/mlearning-ai/explainable-ai-a-complete-summary-of-the-main-methods-a28f9ab132f7

Thank you for your attention!