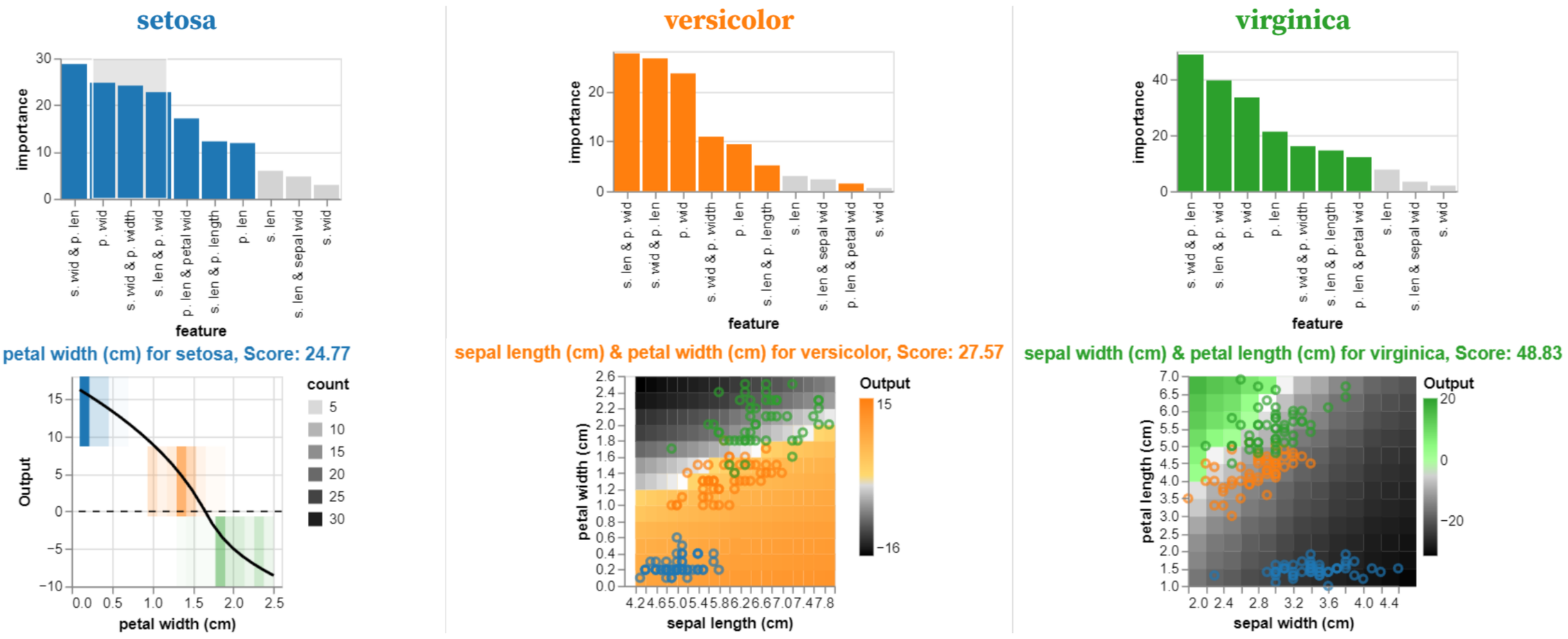
Visualizing Pairwise Feature Interactions in Neural Additive Models

Christian Steinparz, Andreas Hinterreiter, Marc Streit
Johannes Kepler University Linz

Part of the Institute of Computer Graphics at
Johannes Kepler University Linz, Austria



Neural Additive Models (NAMs) [1] are interpretable 'white-box' models, leveraging the expressiveness of neural networks. Each input feature is processed by a corresponding feature network submodule, learning a distinct response function. Their outputs are then aggregated, resulting in the final predictive output. We present an approach to integrate feature interactions into Neural Additive Models (NAMs) and their visualizations as heatmaps, building upon existing work in this area, to enhance their predictive capabilities while maintaining interpretability. Our contribution focuses on the visual exploration and management of the increased number of feature maps resulting from the addition of pairwise feature combinations to NAMs.

Figure 1 *Sorted 1-D and 2-D feature maps of a NAM trained to classify the Iris dataset are presented. In this case, the feature maps are scored by their output range, and then sorted and filtered to show the highest scoring per class. Note that the tool can be found at* `https://observablehq.com/@cursedseraphim/nams-vis`.

Addressing Pairwise Feature Interactions

Incorporating all possible features and pairwise feature interactions results in $n + n\,(n-1)/2$ features per class, where n denotes the number of features. Existing machine learning literature focuses on determining relevant interactions for model training and pruning the model accordingly [4] [3] [2]. In contrast to these approaches, we employ an interactive dashboard that enables users to simultaneously explore multiple feature maps based on their scores. We specifically enable the choice between the submodule output range and its permutation feature importance for scoring.

Visual Encoding of Feature Maps

For visualizing the feature maps, we employ line plots to represent main effects, where the x-axis represents the input feature for the corresponding submodule, while the line illustrates the submodule's output. We enhance the line plots by displaying class-wise data distributions. This is achieved by dividing the y-axis into l equal sections for l classes and representing histograms as heatmaps along the x-axis. The histograms use categorical colors to encode the classes and opacity for the counts.For pairwise feature interactions, we make use of heatmaps. The axes indicate the input features used for the submodule. The color map visualizes the submodule's output, with black areas denoting negative output and class-colored areas signifying positive output. A scatterplot demonstrates the class distribution based on the two input features.

NAM Architecture

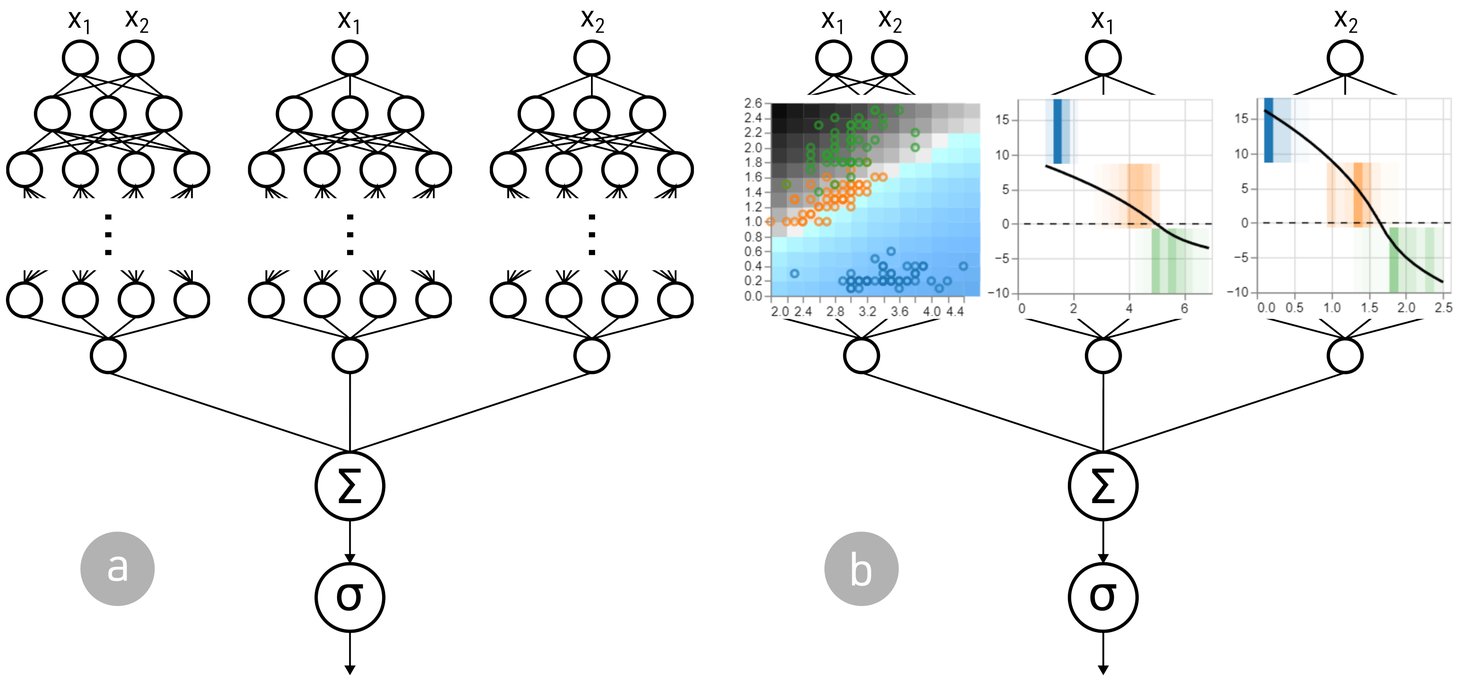


Figure 2 *The figure above schematically illustrates the architecture of a NAM. In this particular instance, the model incorporates two features, x1 and x2, along with their pairwise feature interaction. This interaction is captured through the integration of three corresponding submodules. We provide a representation of how the individual feature maps (b) correspond to these submodules (a). The output activation function depends on the specific classification or regression task.*

[1] R. Agarwal, L. Melnick, N. Frosst, X. Zhang, B. Lengerich, R. Caruana, and G. E. Hinton. Neural additive models: Interpretable machine learning with neural nets. *Advances in Neural Information Processing Systems*, 34:4699–4711, 2021.

[2] C.-H. Chang, R. Caruana, and A. Goldenberg. Node-gam: Neural generalized additive model for interpretable deep learning. *arXiv preprint arXiv:2106.01613*, 2021.

[3] Y. Lou, R. Caruana, J. Gehrke, and G. Hooker. Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 623–631, 2013.

[4] Z. Yang, A. Zhang, and A. Sudjianto. Gami-net: An explainable neural network based on generalized additive models with structured interactions. *Pattern Recognition*, 120:108192, 2021.

Contact
christian.steinparz@jku.at
https://jku-vds-lab.at/publications/#posters

Acknowledgements
Boehringer Ingelheim Regional Center Vienna
Austrian Science Fund (FWF DFH 23–N)