Measure of human-likelihood in tree-based ensemble model and artificial neural networks

Jong M. Shin, Leyla Isik, Joshua T. Vogelstein

Johns Hopkins University | Whiting School of Engineering | Baltimore, MD

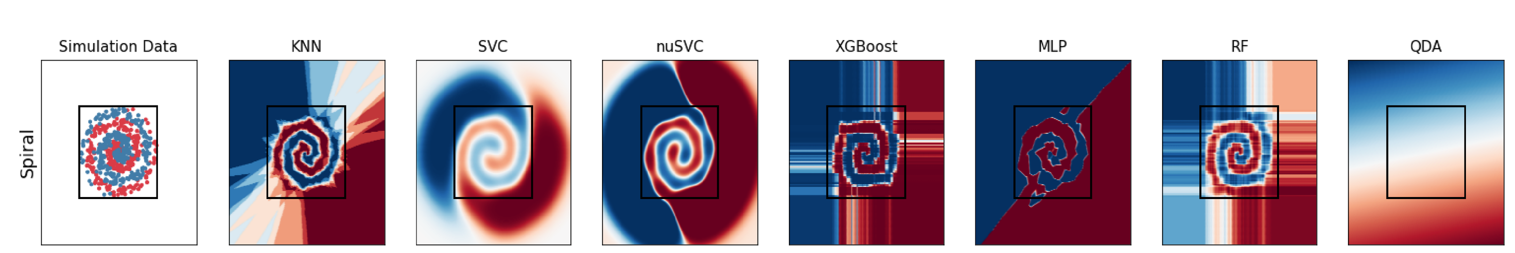
Abstract

Every machine learning (ML) algorithm operates under one form or another inductive bias that allows for generalization of training data. The problem arises when the bias of a model interferes in evaluating data that are not sampled from the similar distribution as the training data, which is often known as distributional shift. We observe that this is particularly the case for artificial neural networks (ANN) compared to the other ML models and even its human counterpart. Let's consider the simple examples of nonlinear data (e.g., Gaussian XOR and spiral simulations) distributed within a unit circle. We trained the classical artificially intelligent (AI) machines (e.g., random forests (RF), multilayer perceptron) on these data and had these machines inductively predict patterns in- and out-of-distribution (OOD) regions of the unit circle. We found that ANN became more confident and behaved less like humans as it extrapolated further away from the origin. In contrast, tree-based ensemble algorithms, such as RF, presented closer resemblance to the behavioral patterns of humans on these data. Of note, we report that tree-based ML models consistently outperformed in metrics that measure functional likelihood of human thereby rendering the most similar behavior to human among the classical AI machines challenged in the study. To our knowledge, this is the first study that demonstrates a sharp contrast between tree-based ensemble model and ANN in the outside of the convex hull using actual human-behavioral data that exhibits OOD generalization.

Problem Statement

- Artificial neural network models are touted as the model of human mind and powerful extrapolators but we are not quite there yet
- These shortfalls are particularly observable from its extrapolative patterns in the out of distribution region
- There are a number of studies suggesting the shortcomings but there has not been a study that explicitly demonstrating a direct comparison between machines and humans to measure these differences

Compare ML models to humans



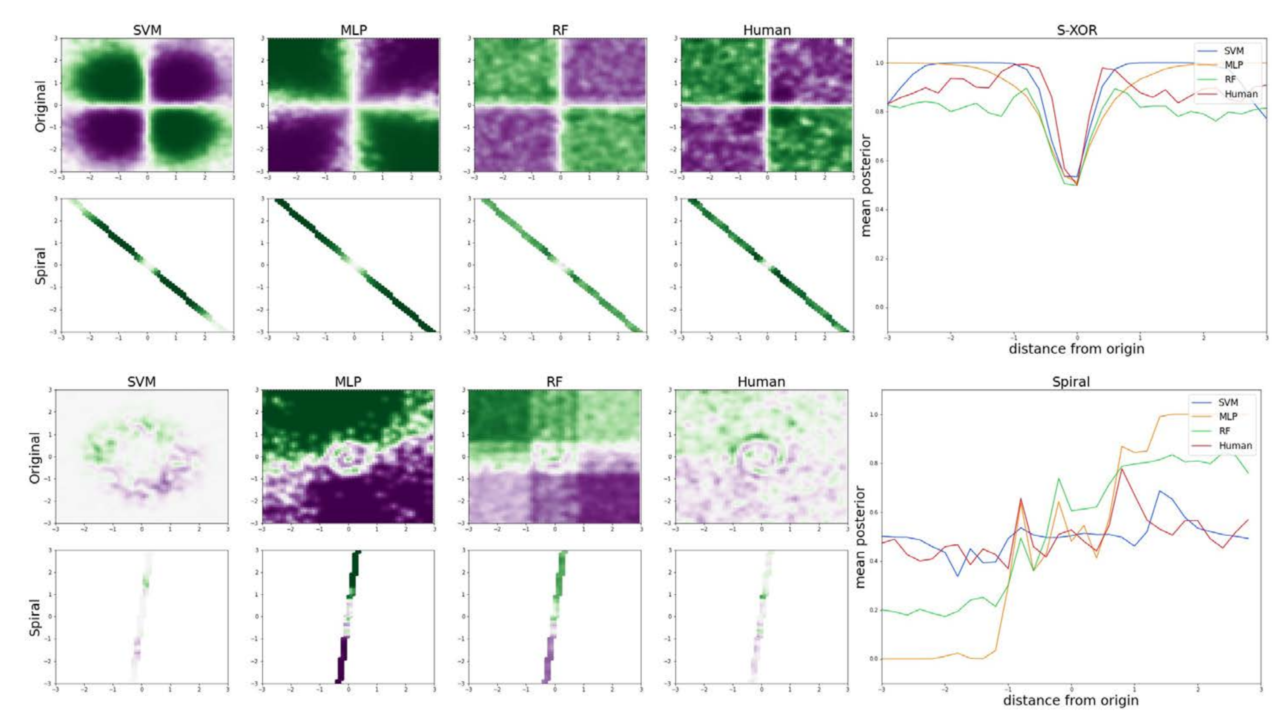
- We initially screened mutlple algorithms, then we noticed a general trend between tree-based ensemble models and artificial neural networks

Hellinger Distance

$$H(P,Q) = \frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{k}(\sqrt{p_i}-\sqrt{q_i})^2} \qquad P = (p_1,\ldots,p_k) \quad Q = (q_1,\ldots,q_k)$$

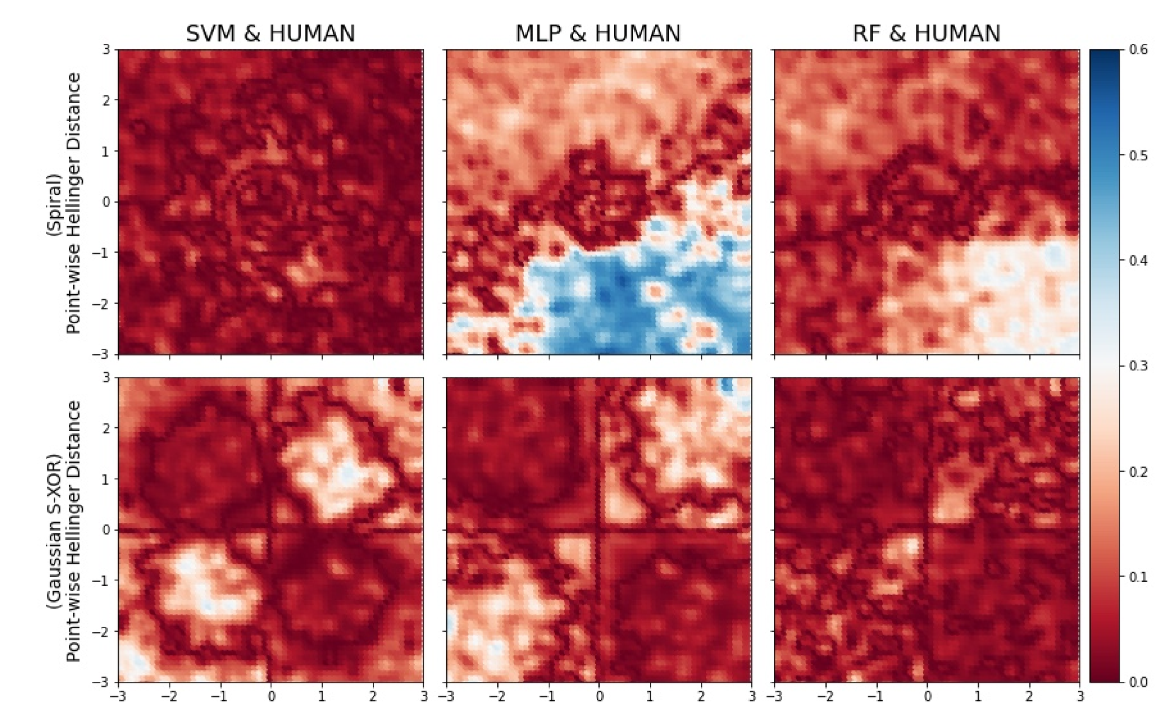
- We used point-wise discrete Hellinger distance between two smoothed posterior distributions (P, Q) over the range of [-3, 3] x [-3, 3] grid

Assessing position-variant posterior



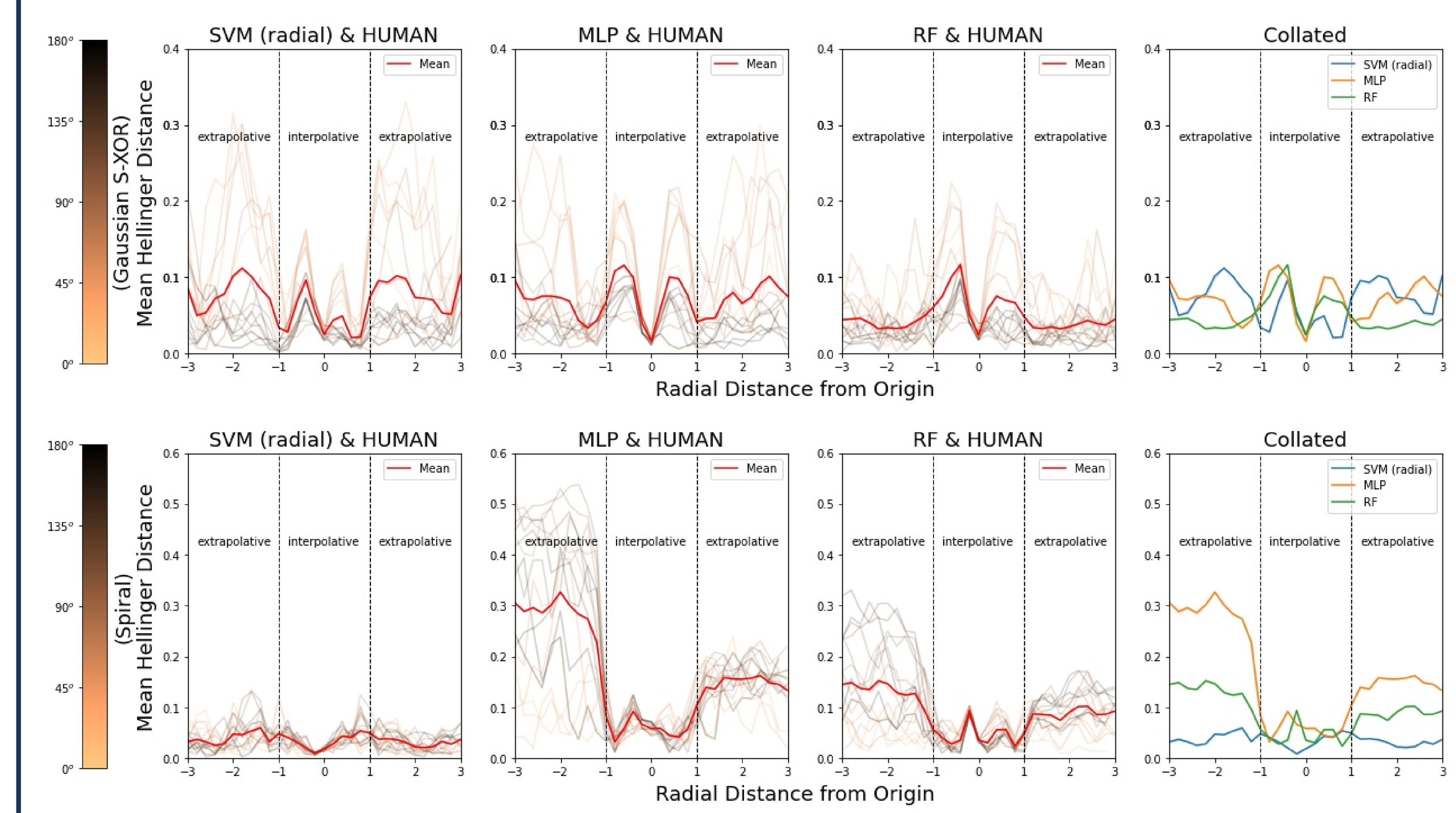
- Non-linear posteriors are not position-invariant. A general approach to capture changes w.r.t. radial distance would not work, thus we assessed posterior in a linear fashion as a function of angle

ANN worst with increasing non-linearity



- As non-linearity of simulation increases, we can see more dissimilarity in the OOD region where MLP is worse in human-likelihood measure

ANN least like humans



- We quantified and visualized human-likelihood from our posterior assessment in a position-invariant manner
- There is a significant difference in human-likelihood in the OOD region as non-linearity of simulation increases
- The confidence, the rate at which each algorithm achieves maximal human-likelihood, also increases as simulation became more complex

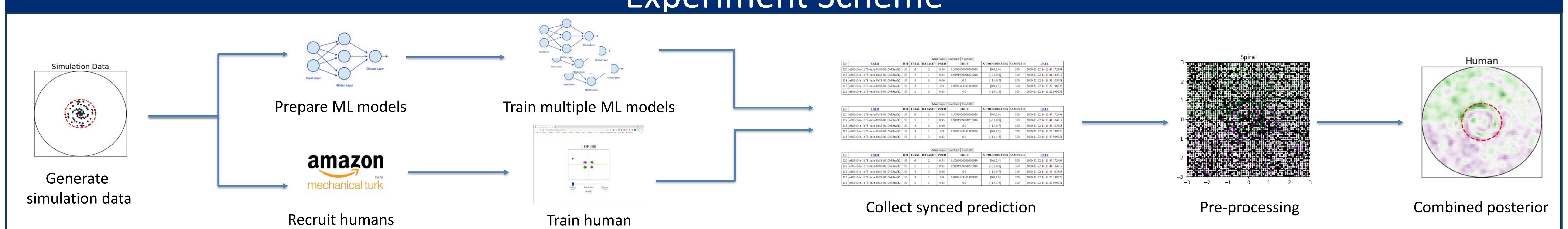
Conclusion

- Position invariant method to assess human-likelihood revealed that ANN behaves least like human especially in the OOD region
- ML models generally perform poorly as non-linearity increases while becoming more confident; this was particularly the case with ANN compared to tree-based ensemble model

Next Step

- Extend the complexity of the models. We can test our theory on more complex models of ANN and tree-based ensemble models
- Increase non-linearity of the simulation. We can corroborate our finding by testing on other simulation data with more non-linearity
- Assess the trend in real-world data. We can explore further into image/sound/text and assess human-likelihood
- Test the behavior of novel ML algorithm(s). We can test our novel ML algorithm(s) that combines ANN and tree-based ensemble models

Experiment Scheme



References

- Xu, Hao et al. *When are Deep Networks really better than Decision Forests at small sample sizes, and how?* arXiv preprint
- Krueger, David et al. *Out-of-Distribution Generalization via Risk Extrapolation.* PMLR

Let's connect!

jshinm@gmail.com | @jshinm | @jshinm