Toward Robust Uncertainty Estimation with Random Activation Functions

Yana Stoyanova, Soroush Ghandi, Maryam Tavakol

Eindhoven University of Technology, Eindhoven, The Netherlands
y.stoyanova@student.tue.nl, s.ghandi@tue.nl, m.tavakol@tue.nl

Abstract

Deep neural networks are in the limelight of machine learning with their excellent performance in many data-driven applications. However, they can lead to inaccurate predictions when queried in out-of-distribution data points, which can have detrimental effects especially in sensitive domains, such as healthcare and transportation, where erroneous predictions can be very costly and/or dangerous. Subsequently, quantifying the uncertainty of the output of a neural network is often leveraged to evaluate the confidence of its predictions, and ensemble models have proved to be effective in measuring the uncertainty by utilizing the variance of predictions over a pool of models. In this paper, we propose a novel approach for uncertainty quantification via ensembles, called *Random Activation Functions (RAFs) Ensemble*, that aims at improving the ensemble diversity toward a more robust estimation, by accommodating each neural network with a different (random) activation function. Extensive empirical study demonstrates that RAFs Ensemble outperforms state-of-the-art ensemble uncertainty quantification methods on both synthetic and real-world datasets in a series of regression tasks.

Introduction

Recent advances in deep neural networks have demonstrated remarkable performance in a wide variety of applications, ranging from recommendation systems and improving user experience to natural language processing and speech recognition (Abiodun et al. 2018). Nevertheless, blindly relying on the outcome of these models can have harmful effects, especially in high-stake domains such as healthcare and autonomous driving, as models can provide inaccurate predictions when queried in out-of-distribution data points (Amodei et al. 2016). Consequently, correctly quantifying the uncertainty of models' predictions is an admissible mechanism to distinguish where a model can or cannot be trusted, and thus, increases the transparency of models about their capabilities and limitations (Abdar et al. 2021). Uncertainty Quantification (UQ) is important for a variety of reasons. For instance, in order to preserve the model's credibility, it is essential to report and communicate the encountered uncertainties regularly (Volodina and Challenor 2021). Additionally, models' predictions are inevitably uncertain in most cases, which has to be addressed to increase their transparency, trustworthiness, and reliability.

In the machine learning literature, uncertainty is usually decomposed into two different types, namely aleatoric uncertainty and epistemic uncertainty (Kiureghian and Ditlevsen 2009). *Aleatoric* uncertainty, aka data uncertainty, refers to the inherent uncertainty that stems from the data itself, e.g., noise. On the other hand, *epistemic* uncertainty, also called model uncertainty, is the type of uncertainty that occurs due to the lack of sufficient data. While data uncertainty *cannot* be alleviated, model uncertainty can be addressed by e.g., acquiring more data. Let $\boldsymbol{\sigma}_a^2$ and $\boldsymbol{\sigma}_e^2$ denote the aleatoric and epistemic uncertainties, respectively. Since the distinction between the two is imprecise to some degree (Sullivan 2015), we focus on the predictive (total) uncertainty, which is defined as the sum of the two

$$\boldsymbol{\sigma}_p^2 = \boldsymbol{\sigma}_a^2 + \boldsymbol{\sigma}_e^2. \tag{1}$$

Accordingly, the approaches developed for uncertainty estimation can be categorized into three groups: Bayesian UQ methods, ensemble UQ methods, and a combination of both, i.e., Bayesian ensemble UQ (Abdar et al. 2021). In this paper, we focus on ensemble UQ techniques, either Bayesian or non-Bayesian, as this group is less explored compared to the solely Bayesian techniques. An ensemble model aggregates the predictions of multiple individual base-learners (or ensemble members), which in our case are neural networks (NNs), and the empirical variance of their predictions gives an approximate measure of uncertainty. The idea behind this heuristic is highly intuitive: the more the base-learners disagree on the outcome, the more uncertain they are. Therefore, the goal of ensemble members is to have a great level of disagreement (variability) in the areas where little or no data is available, and to have a high level of agreement in regions with abundance of data (Pearce et al. 2018).

In this paper, we propose a novel method, called *Random Activation Functions Ensemble (RAFs Ensemble)*, for a more robust uncertainty estimation in (deep) neural networks. RAFs Ensemble is developed on top of Anchored Ensemble technique, proposed by (Pearce et al. 2018), however, instead of initializing each NN member in the ensemble with the same activation function, the NNs in RAFs Ensemble are accommodated with different (random) activation functions in the hidden layers. This simple, yet crucial, mod-

ification greatly improves the overall diversity of the ensemble, which is one of the most important components in forming a successful ensemble. We empirically show that RAFs Ensemble provides high quality uncertainty estimates compared to five state-of-the-art ensemble methods, that is Deep Ensemble (Lakshminarayanan, Pritzel, and Blundell 2017), Neural Tangent Kernel Gaussian Process Parameter Ensemble (He, Lakshminarayanan, and Teh 2020), Anchored Ensemble (Pearce et al. 2018), Bootstrapped Ensemble of NNs Coupled with Random Priors (Osband, Aslanides, and Cassirer 2018), and Hyperdeep Ensemble (Wenzel et al. 2020). The comparisons are performed in a wide range of regression tasks on both synthetic and real-world datasets in terms of negative log-likelihood and root mean squared error.

Related Work

Uncertainty Quantification (UQ) is an active field of research and various methods have been proposed to efficiently estimate the uncertainty of machine learning models (see Abdar et al. 2021 for an extensive overview). While most research focuses on Bayesian deep learning (Srivastava et al. 2014; Blundell et al. 2015; Sensoy, Kandemir, and Kaplan 2018; Fan et al. 2020; Järvenpää, Vehtari, and Marttinen 2020; Charpentier, Zügner, and Günnemann 2020), deep ensemble methods, which benefit from the advantages of both deep learning and ensemble learning, have been recently leveraged for empirical uncertainty quantification (Egele et al. 2021; Hoffmann, Fortmeier, and Elster 2021; Brown, Bhuiyan, and Talbert 2020; Althoff, Rodrigues, and Bazame 2021). Although Bayesian UQ methods have solid theoretical foundation, they often require significant changes to the training procedure and are computationally expensive compared to non-Bayesian techniques such as ensembles (Egele et al. 2021; Rahaman and Thiery 2021; Lakshminarayanan, Pritzel, and Blundell 2017).

Lakshminarayanan, Pritzel, and Blundell (2017) are among the first to challenge Bayesian UQ methods by proposing Deep Ensemble, a simple and scalable technique, that demonstrates superb empirical performance on a variety of datasets. However, one of the challenges of ensemble techniques when quantifying uncertainty is that they tend to give overconfident predictions (Amodei et al. 2016). To address this, Pearce et al. (2018) propose to also regularize the model's parameters w.r.t. the initialization values, instead of zero, leading to Anchored Ensembles, which additionally allows for performing Bayesian inference in NNs. He, Lakshminarayanan, and Teh (2020) relate Deep Ensembles to Bayesian inference using neural tangent kernels. Their method, i.e., Neural Tangent Kernel Gaussian Process Parameter Ensemble (NTKGP-param), trains all layers of a finite width NN, obtaining an exact posterior interpretation in the infinite width limit with neural tangent kernel parameterization and squared error loss. They prove that NTKGP-param is always more conservative than Deep Ensemble, yet, its advantages are generally not clear in practice.

A prominent advance to the Bayesian ensemble UQ methods is the bootstrapped ensemble of NNs coupled with random priors, proposed by (Osband, Aslanides, and Cassirer 2018), in which, the random prior function and neural models share an input and a summed output, but the networks are the only trainable parts, while the random prior remains untrained throughout the whole process. Furthermore, Wenzel et al. (2020) exploit an additional source of randomness in ensembles by designing ensembles not only over weights, but also over hyperparameters. Their method, called Hyperdeep Ensemble, demonstrates high accuracy for a number of different classification tasks. Nevertheless, despite the recent contributions in ensemble UQ methods, the research in this direction still needs further advancement.

Toward Robust Uncertainty Estimation

Preliminaries

Following the notations of (Lakshminarayanan, Pritzel, and Blundell 2017), let S_{train} be a training dataset consisting of n independently and identically drawn (i.i.d.) data points, $S_{train} = \{\boldsymbol{x}_i, y_i\}_{i=1}^{n}$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ denotes a d-dimensional feature vector and $y_i \in \mathbb{R}$ is a scalar output. Similarly, S_{test} indicates the test set. Subsequently, X represents the design matrix and $\boldsymbol{y}$ indicates the output vector, where $(S_{train}.X, S_{train}.\boldsymbol{y})$ and $(S_{test}.X, S_{test}.\boldsymbol{y})$ represent the train and test sets, respectively. Without the loss of generality, we consider the regression tasks of the form

$$\boldsymbol{y} = f(X) + \epsilon,$$

where ϵ is a normally distributed constant noise, i.e., $\epsilon \sim \mathcal{N}(0, \boldsymbol{\sigma}_a^2)$, and is assumed to be known. The goal is hence to quantify the predictive uncertainty $\boldsymbol{\sigma}_p^2$ associated with $S_{test}.\boldsymbol{y}$, while optimizing f on the training data.

We adapt the regularized loss function from the Anchored Ensemble technique (Pearce et al. 2018), in which, the regularization of the models' parameters are carried out w.r.t. their initialization values instead of zero. Consequentially, given $\boldsymbol{\theta}_j$ as the parameters of the j_{th} base-learner, the objective function is as follows

$$\mathcal{L}(\boldsymbol{\theta}_j) = \frac{1}{n}||\boldsymbol{y} - \hat{\boldsymbol{y}}_j||_2^2 + \frac{1}{n}||\Gamma^{1/2}(\boldsymbol{\theta}_j - \boldsymbol{\theta}_{0,j})||_2^2, \qquad (2)$$

where $\boldsymbol{\theta}_{0,j}$ is derived from the prior distribution, $\boldsymbol{\theta}_{0,j} \sim \mathcal{N}(\boldsymbol{\mu}_0, \Sigma_0)$, and Γ is the regularization matrix. Furthermore, minimizing this objective allows for performing Bayesian inference in NNs. However, this technique only models the epistemic uncertainty, while aleatoric uncertainty is assumed to be constant (Pearce et al. 2018), which is a limitation, as it is not always possible to distinguish the different origins or types of uncertainty in practice (see Equation 1).

Therefore, in this paper, we aim at enhancing the performance of the ensemble toward a more robust uncertainty estimation. The literature suggests that diversifying the ensembles is effective in improving their predictive performance both theoretically and empirically (Zhou 2012; Zhang and Ma 2012; Hansen and Salamon 1990; Krogh and Vedelsby 1994). Ideally, diversity is achieved when the predictions made by each model in the ensemble are independent and uncorrelated. However, generating diverse ensemble members is not a straightforward task. The main impediment is the fact that each neural network is trained on the same training data to solve the same problem, which usually results in

a high correlation among the individual base-learners (Zhou 2012). In the subsequent section, we introduce a simple technique to efficiently improve the overall diversity of the ensemble for a more reliable uncertainty quantification.

RAFs Ensemble

In this section, we present Random Activation Functions (RAFs) Ensemble for uncertainty estimation, which can be extended to all ensemble methods in terms of methodological modification. When a (Bayesian) ensemble is leveraged to estimate the uncertainty of a deep neural network model, we propose to increase the diversity of predictions among the ensemble members using varied activation functions (AFs), in addition to the random initialization of the parameters. To do so, instead of initializing the neural networks with the same activation function, each NN is accommodated with a different (random) activation function. Subsequently, distinct activation functions account for different non-linear properties introduced to each ensemble member, therewith improving the overall diversity of the ensemble.

As mentioned previously, the ensemble diversity is one of the most important building blocks when it comes to creating a successful ensemble (Hansen and Salamon 1990). Hence, it might be preferable to combine the predictions of top-performing base-learners with the predictions of weaker ones (Zhou 2012). Otherwise, stacking only strong models will likely result in a poor ensemble as the predictions made by the models will be highly correlated, and thus, the ensemble diversity will be greatly limited. Therefore, the choice of activation functions should be motivated purely by their variability and not their appropriateness for the task at-hand.

Let $\boldsymbol{\mu}_0$ be the prior means, Σ_0 be the prior covariance, $\hat{\boldsymbol{\sigma}}_a^2$ be an estimate of data noise, m denote the number of base-learners, and NN_j indicate the j_{th} member, the entire procedure for both training and prediction is summarized in Algorithm 1. In this algorithm, a regularization matrix is first created and a set of activation functions is defined (line 1-2). Then, the NNs in the ensemble are trained to minimize the loss function defined in Equation 2 with stochastic gradient descent, using arbitrary optimizer and no early stopping (line 3-13). Note that if the size of the ensemble m is smaller or equal to the cardinality of the AFs set k, then each NN is trained with a different activation function, and with random functions from the set, otherwise (line 7-11). Consequently, predictions are made with each ensemble member (line 14-16), which are then averaged and an estimate of the predictive uncertainty is computed (line 17-19).

Algorithm 1: RAFs Ensemble

Input: S_{train}, S_{test}, priors $\boldsymbol{\mu}_0$ and Σ_0, m, $\hat{\boldsymbol{\sigma}}_a^2$
Output: Estimate of predictive mean $\hat{\boldsymbol{y}}$ and variance $\hat{\boldsymbol{\sigma}}_p^2$

1: $\Gamma \leftarrow \hat{\boldsymbol{\sigma}}_a^2 \Sigma_0^{-1}$ ▷ Regularization matrix
2: $\mathbb{A} \leftarrow \{a_1, \ldots, a_k\}$ ▷ Set of k AFs
3: **for** j in $1 : m$ **do** ▷ Train the ensemble
4: Create NN_j with $\boldsymbol{\theta}_{j,0} \leftarrow \mathcal{N}(\boldsymbol{\mu}_0, \Sigma_0)$
5: **if** $j \leq k$ **then**
6: $\alpha_j = a_j$
7: **else**
8: $\alpha_j \leftarrow$ Randomly selected from $\mathbb{A}$
9: **end if**
10: $NN_j.train(S_{train}, \Gamma, \boldsymbol{\theta}_{j,0}, \alpha_j)$ using loss in Eq. 2
11: **end for**
12: **for** j in $1 : m$ **do** ▷ Predict with the ensemble
13: $\hat{\boldsymbol{y}}_j = NN_j.predict(S_{test}.X)$
14: **end for**
15: $\hat{\boldsymbol{y}} = \frac{1}{m} \sum_{j=1}^{m} \hat{\boldsymbol{y}}_j$ ▷ Mean predictions
16: $\hat{\boldsymbol{\sigma}}_e^2 = \frac{1}{m-1} \sum_{j=1}^{m} (\hat{\boldsymbol{y}}_j - \hat{\boldsymbol{y}})^2$ ▷ Epistemic variance
17: $\hat{\boldsymbol{\sigma}}_p^2 = \hat{\boldsymbol{\sigma}}_e^2 + \hat{\boldsymbol{\sigma}}_a^2$ ▷ Total variance Eq. 1
18: **return** $\hat{\boldsymbol{y}}, \hat{\boldsymbol{\sigma}}_p^2$

Empirical Study

Experimental Setups

In the experiments, the base-learners of RAFs Ensemble are multilayer perceptrons that consist of one hidden layer of 100 neurons. The ensemble size m is set to five. This is standard for the implementations of all methods in this paper, as $m = 5$ proved to be empirically sufficient for obtaining predictive uncertainty estimates in the experiments. In addition, we choose a set of seven activation functions which is comprised of (i) Gaussian Error Linear Unit (GELU) (Hendrycks and Gimpel 2016), (ii) Softsign (Turian, Bergstra, and Bengio 2009), (iii) Swish (Ramachandran, Zoph, and Le 2018), (iv) Scaled Exponential Linear Unit (SELU) (Klambauer et al. 2017), (v) hyperbolic tangent (tanh), (vi) error activation function, and (vii) linear (identity) activation function. Furthermore, the number of testing samples is set to be always larger than the number of training points n to detail the uncertainty. Moreover, to account for epistemic uncertainty, the synthetic testing feature vectors $\boldsymbol{x} \in S_{test}$ range over wider intervals compared to $\boldsymbol{x} \in S_{train}$ and both are sampled uniformly at random. The code is available at https://github.com/YanasGH/RAFs for reproducibility purposes.

Baselines. We include five state-of-the-art methods as baselines for empirical comparison with RAFs Ensemble as follows. (i) DE (Lakshminarayanan, Pritzel, and Blundell 2017), (ii) AE (Pearce et al. 2018), (iii) HDE (Wenzel et al. 2020), (iv) RP-param (Osband, Aslanides, and Cassirer 2018), and (v) NTKGP-param (He, Lakshminarayanan, and Teh 2020), on both synthetic and real-world datasets with different dimensionalities. To ensure fair comparison between the UQ techniques, roughly the same amount of time has been put into hyperparameter tuning for each method.

Synthetic Data. We generate multiple synthetic datasets that fall into four categories: physical models (PM), many local minima (MLM), trigonometric (T), and others (O). Each set in the PM category is generated from a physical mathematical model, such that all values in S_{train} and S_{test} are achievable in the real world. Generally, the PM datasets

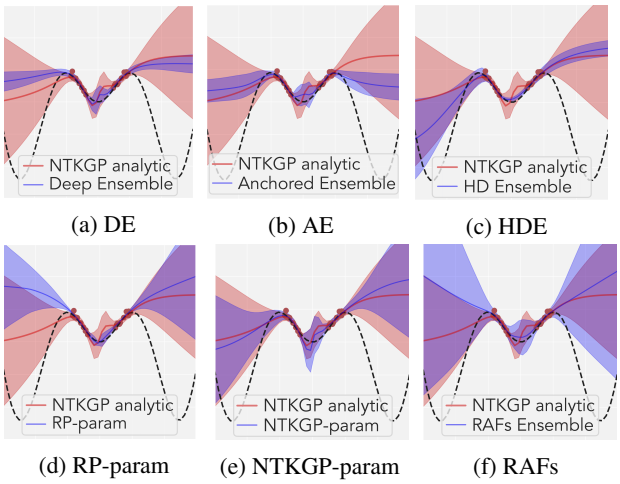
have complex modeling dynamics and can be characterized as having predominant epistemic uncertainty due to the considerably wider testing sampling regions by design. Similarly, the MLM data, generated from functions with many local minima, are also designed so that the model uncertainty is higher than the aleatoric one. These datasets are usually hard to approximate due to their inherent high-nonlinearity and multimodality. Another category with higher epistemic uncertainty is trigonometric, such as data generated by (He, Lakshminarayanan, and Teh 2020) and (Forrester, Sobester, and Keane 2008), where the training data is partitioned into two equal-sized clusters in order to detail uncertainty on out-of-distribution data (see Figure 1). In contrast, the predominant type of uncertainty in the O category is aleatoric. This category includes datasets generated from various functions such as rational and product integrand functions. It is distinguished from the rest of the categories by its high interaction effects. The dimensionality of all datasets can range from one to ten and we consider two datasets per dimension, thus, the total number of synthetic data is 20. More detail on how the data is created can be found in the Appendix at https://arxiv.org/abs/2302.14552.

Real-world Data. Additionally, we use five real-world datasets for evaluation: Boston housing, Abalone shells (Nash et al. 1994), Naval propulsion plant (Coraddu et al. 2014), Forest fire (Cortez and de Jesus Raimundo Morais 2007), and Parkinson's disease dataset (Little et al. 2007). To account for aleatoric uncertainty (some) context factors are disregarded, such that this type of uncertainty is characteristically high (see Appendix for more details).

Evaluation Criteria. We employ two evaluation criteria to gauge the overall performance of the trained models, namely calibration and robustness to the distribution shift. Both measures are inspired by the practical applications of NNs, as generally there is no theoretical evidence for evaluating uncertainty estimates (Abdar et al. 2021). Calibration is defined as the analytical process of adjusting the inputs with the purpose of making the model to predict the actual observations as precisely as possible (Bijak and Hilton 2021). The quality of calibration can be measured by proper scoring rules such as negative log-likelihood (NLL). NLL is a common choice when it comes to evaluating UQ estimates, as it depends on predictive uncertainty (Lakshminarayanan, Pritzel, and Blundell 2017). Additionally, due to its practical applicability in a wide spectrum of regression tasks, root mean squared error (RMSE) is measured, although it does not depend on the estimated uncertainty (Lakshminarayanan, Pritzel, and Blundell 2017), but serves as a proxy and a secondary assessor of the performance. Moreover, to measure the robustness/generalization of methods to distributional shift, we test the models in out-of-distribution settings, such as the synthetic datasets by (Forrester, Sobester, and Keane 2008; He, Lakshminarayanan, and Teh 2020).

Performance Results

Qualitative Comparison. Figure 1 shows the performance of different methods compared to a Gaussian process



(a) DE (b) AE (c) HDE

(d) RP-param (e) NTKGP-param (f) RAFs

Figure 1: Uncertainty quantification of different methods on He et al. dataset. Gaussian process with neural tangent kernel (NTKGP analytic) is included as a reference.

with a neural tangent kernel (NTKGP analytic) as a reference, on a 1D toy dataset generated from $\boldsymbol{y} = \boldsymbol{x}\sin(\boldsymbol{x}) + \boldsymbol{\epsilon}$ (dashed line). The plots demonstrate that DE, HDE, and AE provide narrow uncertainty bounds in areas where no data has been observed by the model, which translates to high confidence in OOD data. On the contrary, NTKGP-param, RP-param, and RAFs Ensemble bound their uncertainty estimates with wider intervals in areas with no data, accounting for adequate quantification of epistemic uncertainty, while also indicating robustness to OOD data. Among these methods, RAFs Ensemble provides the widest uncertainty which is reasonable considering the amount of data that is available to the methods over each area. Moreover, this observation is quantitatively validated as RAFs Ensemble achieves the lowest NLL compared to the other methods (see Table 1).

Overall Performance. We evaluate the overall performance of all methods in terms of both NLL and RMSE. The outcomes of comparing RAFs Ensemble with five baseline methods on twenty synthetic and five real-world datasets are outlined in Table 1 and Table 2. The results illustrate that our approach outperforms the competitors in most scenarios. Furthermore, Table 3 summarizes the obtained results in terms of ranking, in which the methods are ranked based on their performance for a particular dataset. The left integer corresponds to NLL, while the right one points to RMSE, and the bold values indicate the best-performing method.

Discussion. The obtained results in this section illustrate that DE has good uncertainty estimates with respect to NLL for the real-world datasets, and it takes the first place for Naval propulsion and Parkinson's datasets. For the rest of the data categories, when compared to the other methods, DE fails to provide strong performance, usually scoring a very low NLL rank. Therefore, this indicates that Deep Ensemble might have difficulty quantifying epistemic uncertainty in general as displayed by the experiments in this paper, but seemingly manages to capture aleatoric uncertainty well.

	NLL					
	DE	HDE	AE	NTKGP-p.	RP-p.	RAFs
He et al. 1D	>100 ± 0.18	71.31 ± 0.51	38.75 ± 0.12	4.48 ± 0.18	13.05 ± 0.43	**2.21 ± 0.18**
Forrester et al. 1D	>100 ± 0.53	>100 ± 0.51	50.82 ± 0.52	>100 ± 0.50	13.7 ± 0.58	**0.64 ± 0.74**
Schaffer N.4 2D	0.29 ± 0.01	-0.71 ± 0.01	2.15 ± 0.01	-0.55 ± 0.01	-0.36 ± 0.01	**-0.79 ± 0.01**
Double pendulum 2D	2.95 ± 0.05	2.18 ± 0.84	-0.36 ± 0.05	**-0.58 ± 0.05**	**-0.47 ± 0.05**	**-0.49 ± 0.04**
Rastrigin 3D	29.24 ± 1.30	**3.09 ± 1.15**	35.94 ± 0.74	28.38 ± 0.64	**4.35 ± 1.24**	**3.44 ± 1.05**
Ishigami 3D	6.01 ± 0.08	>100 ± 0.08	8.73 ± 0.08	1.53 ± 0.08	**-0.01 ± 0.08**	**0.06 ± 0.07**
Environmental 4D	64.72 ± 0.23	7.84 ± 0.13	1.65 ± 0.20	4.5 ± 0.27	3.94 ± 0.21	**0.81 ± 0.17**
Griewank 4D	28.29 ± 2.43	**5.50 ± 1.62**	**4.64 ± 3.06**	10.21 ± 2.37	**4.29 ± 2.93**	**4.79 ± 2.40**
Roos & Arnold 5D	-2.02 ± 0.01	**-2.21 ± 0.00**	-1.89 ± 0.01	-1.71 ± 0.01	-1.70 ± 0.01	-2.1 ± 0.01
Friedman 5D	96.94 ± 0.41	>100 ± 0.51	15.04 ± 0.50	41.69 ± 0.39	4.22 ± 0.44	**1.78 ± 0.39**
Planar arm torque 6D	9.58 ± 0.07	4.11 ± 0.08	3.07 ± 0.05	**-0.32 ± 0.08**	-0.05 ± 0.07	-0.16 ± 0.06
Sum of powers 6D	>100 ± 0.41	>100 ± 0.62	55.03 ± 0.43	>100 ± 0.41	41.59 ± 0.40	**35.22 ± 0.35**
Ackley 7D	7.11 ± 0.23	**1.38 ± 0.16**	2.50 ± 0.36	3.11 ± 0.27	2.09 ± 0.26	**1.16 ± 0.08**
Piston simulation 7D	**-2.19 ± 0.00**	14.06 ± 0.00	3.50 ± 2.40	2.87 ± 2.93	2.67 ± 0.42	3.63 ± 0.57
Robot arm 8D	10.71 ± 0.03	6.87 ± 0.01	7.11 ± 0.01	**0.27 ± 0.03**	0.80 ± 0.06	**0.25 ± 0.02**
Borehole 8D	>100 ± 1.01	>100 ± 1.01	**4.89 ± 1.87**	**5.48 ± 3.54**	**4.06 ± 1.20**	**4.36 ± 1.26**
Styblinski-Tang 9D	>100 ± 3.05	>100 ± 0.00	40.80 ± 5.33	>100 ± 3.03	**15.82 ± 6.31**	**25.23 ± 4.12**
PUMA560 9D	6.59 ± 0.15	**1.62 ± 0.14**	4.24 ± 0.14	5.93 ± 0.08	6.40 ± 0.14	2.14 ± 0.13
Adapted Welch 10D	>100 ± 0.81	>100 ± 0.75	>100 ± 0.55	>100 ± 0.75	>100 ± 0.57	**78.53 ± 0.67**
Wing weight 10D	>100 ± 0.00	27.31 ± 4.37	**5.46 ± 4.36**	67.30 ± 0.53	**5.54 ± 4.15**	**5.39 ± 1.69**
Boston housing	74.54 ± 1.06	>100 ± 1.04	71.53 ± 1.06	70.82 ± 1.06	>100 ± 1.10	**40.67 ± 1.00**
Abalone	>100 ± 0.10	>100 ± 0.10	47.67 ± 0.10	>100 ± 0.10	**28.37 ± 0.10**	28.90 ± 0.10
Naval propulsion	**-2.27 ± 0.00**	>100 ± 0.00	3.92 ± 0.10	2.28 ± 1.51	2.16 ± 0.16	1.91 ± 0.07
Forest fire	15.71 ± 0.05	**3.14 ± 0.02**	**2.66 ± 0.69**	**3.10 ± 1.11**	4.68 ± 0.14	**2.15 ± 0.28**
Parkinson's	**26.74 ± 0.02**	>100 ± 0.10	>100 ± 0.16	>100 ± 0.03	>100 ± 0.16	45.69 ± 0.16

Table 1: Performance of methods on all datasets w.r.t. NLL, including 95% confidence intervals. The best scores are in bold.

Unlike DE, the HDE provides outwardly reliable estimates for datasets with many local minima, despite its unimpressive overall results when compared to the other methods. However, both DE and HDE can produce uncertainty bounds that are unreasonably narrow in areas with unobserved data, as shown in Figure 1 and noted by (Heiss et al. 2021).

Nonetheless, AE demonstrates good performance in the dataset categories that exhibit higher epistemic uncertainty such as the physical models. This is due to the fact that AE is designed for capturing model uncertainty, while aleatoric uncertainty is assumed to be constant. Accordingly, AE achieves inferior performance on the real-world datasets, as those generally have more data uncertainty appropriated.

On the other hand, NTKGP-param achieves its finest performance for datasets in the physical model category, which is normally associated with substantial model uncertainty. A credible rationale to explain this insight is the fact that NTKGP-param tends to be more conservative than Deep Ensemble. However, it is generally unclear in which situations this is beneficial since the ensemble members of NTKGP-param will always be misspecified in practice according to (He, Lakshminarayanan, and Teh 2020).

Furthermore, RP-param manages to rank comparatively high for real-world datasets as well as trigonometric data, that contain vast amounts of aleatoric and epistemic uncertainty, respectively, indicating that it does not quantify either type of uncertainty better than the other. This observation serves as a demonstration that RP-param generalizes well for different types of datasets that exhibit broad characteristics. However, this technique fails to deliver low NLL scores on some occasions, which might be attributed to the fact that RP-param is based on bootstrapping. While bootstrapping can be a successful strategy for inducing diversity, it can sometimes harm the performance when the base-learners have multiple local optima, as is a common case with NNs (Lakshminarayanan, Pritzel, and Blundell 2017).

Nevertheless, RAFs Ensemble outperforms RP-param, and every other method in the comparisons, for 13 out of 25 datasets. In terms of NLL, our approach does not rank below the second place for any data, which is consistent with the strong results from Table 1. Meanwhile, the RMSE scores of this method are altogether satisfactory, although not as prominent compared to the NLL scores. In agreement with the overall outstanding results, RAFs Ensemble holds the highest NLL rank for all data from *MLM* and *T* categories, which can be contemplated as a concluding statement regarding its capabilities to estimate epistemic uncertainty and challenging multimodality. Among all categories, the real-world datasets are least favored by RAFs Ensemble, primarily due to their high level of aleatoric uncertainty. This indicates that RAFs Ensemble captures model uncertainty slightly better than aleatoric uncertainty. Nonetheless, the empirical superiority of this technique is due to the exhaustively exploited added source of randomness via random activation functions, combined with method simplicity and Bayesian behavior, resulted from the anchored loss (Equation 2). This successful combination leads to greatly improved diversity among ensemble members, which can be further confirmed by a direct comparison between RAFs Ensemble and AE. Note that even though RAFs Ensemble does

	RMSE					
	DE	HDE	AE	NTKGP-p.	RP-p.	RAFs
He et al. 1D	3.71 ± 0.18	5.70 ± 0.51	$\mathbf{3.15 \pm 0.12}$	3.64 ± 0.18	5.24 ± 0.43	3.80 ± 0.18
Forrester et al. 1D	5.00 ± 0.53	4.12 ± 0.51	4.09 ± 0.52	6.05 ± 0.50	5.70 ± 0.58	$\mathbf{2.8 \pm 0.74}$
Schaffer N.4 2D	$\mathbf{0.23 \pm 0.01}$	0.34 ± 0.01	0.30 ± 0.01	$\mathbf{0.24 \pm 0.01}$	0.31 ± 0.01	0.27 ± 0.01
Double pendulum 2D	$\mathbf{0.46 \pm 0.05}$	2.22 ± 0.84	0.71 ± 0.05	$\mathbf{0.51 \pm 0.05}$	0.74 ± 0.05	$\mathbf{0.58 \pm 0.04}$
Rastrigin 3D	18.41 ± 1.30	$\mathbf{10.96 \pm 1.15}$	25.58 ± 0.74	18.10 ± 0.64	$\mathbf{12.87 \pm 1.24}$	$\mathbf{14.85 \pm 1.05}$
Ishigami 3D	$\mathbf{0.69 \pm 0.08}$	1.05 ± 0.08	0.88 ± 0.08	$\mathbf{0.69 \pm 0.08}$	$\mathbf{0.58 \pm 0.08}$	$\mathbf{0.57 \pm 0.07}$
Environmental 4D	$\mathbf{2.04 \pm 0.23}$	2.51 ± 0.13	$\mathbf{1.83 \pm 0.20}$	$\mathbf{2.34 \pm 0.27}$	$\mathbf{2.03 \pm 0.21}$	$\mathbf{1.68 \pm 0.17}$
Griewank 4D	83.97 ± 2.43	$\mathbf{45.68 \pm 1.62}$	$\mathbf{42.12 \pm 3.06}$	78.47 ± 2.37	$\mathbf{38.62 \pm 2.93}$	78.79 ± 2.40
Roos & Arnold 5D	0.07 ± 0.01	$\mathbf{0.01 \pm 0.00}$	0.07 ± 0.01	0.09 ± 0.01	0.08 ± 0.01	0.08 ± 0.01
Friedman 5D	$\mathbf{3.17 \pm 0.41}$	$\mathbf{3.63 \pm 0.51}$	$\mathbf{2.95 \pm 0.50}$	$\mathbf{3.39 \pm 0.39}$	$\mathbf{2.74 \pm 0.44}$	$\mathbf{3.1 \pm 0.39}$
Planar arm torque 6D	$\mathbf{0.65 \pm 0.07}$	$\mathbf{0.62 \pm 0.08}$	$\mathbf{0.71 \pm 0.05}$	$\mathbf{0.71 \pm 0.08}$	1.08 ± 0.07	$\mathbf{0.74 \pm 0.06}$
Sum of powers 6D	$\mathbf{22.81 \pm 0.41}$	$\mathbf{21.19 \pm 0.62}$	$\mathbf{21.87 \pm 0.43}$	$\mathbf{22.79 \pm 0.41}$	$\mathbf{22.22 \pm 0.40}$	$\mathbf{22.24 \pm 0.35}$
Ackley 7D	8.92 ± 0.23	2.43 ± 0.16	7.28 ± 0.36	8.58 ± 0.27	4.03 ± 0.26	$\mathbf{1.33 \pm 0.08}$
Piston simulation 7D	$\mathbf{0.02 \pm 0.00}$	0.04 ± 0.00	29.1 ± 2.40	$>100 \pm 2.93$	5.78 ± 0.42	7.40 ± 0.57
Robot arm 8D	0.92 ± 0.03	$\mathbf{0.80 \pm 0.01}$	0.88 ± 0.01	0.93 ± 0.03	1.09 ± 0.06	$\mathbf{0.83 \pm 0.02}$
Borehole 8D	$\mathbf{32.11 \pm 1.01}$	$\mathbf{32.12 \pm 1.01}$	48.75 ± 1.87	$>100 \pm 3.54$	38.60 ± 1.20	41.35 ± 1.26
Styblinski-Tang 9D	$>100 \pm 3.05$	$>100 \pm 0.00$	$\mathbf{>100 \pm 5.33}$	$>100 \pm 3.03$	$\mathbf{>100 \pm 6.31}$	$>100 \pm 4.12$
PUMA560 9D	3.93 ± 0.15	$\mathbf{3.23 \pm 0.14}$	$\mathbf{3.40 \pm 0.14}$	3.93 ± 0.08	$\mathbf{3.24 \pm 0.14}$	$\mathbf{3.4 \pm 0.13}$
Adapted Welch 10D	$\mathbf{99.51 \pm 0.81}$	$\mathbf{99.4 \pm 0.75}$	$>100 \pm 0.55$	$\mathbf{99.79 \pm 0.75}$	$>100 \pm 0.57$	$\mathbf{100.00 \pm 0.67}$
Wing weight 10D	$>100 \pm 0.00$	$\mathbf{58.16 \pm 4.37}$	$\mathbf{63.1 \pm 4.36}$	$>100 \pm 0.53$	$\mathbf{63.35 \pm 4.15}$	$>100 \pm 1.69$
Boston housing	$\mathbf{11.28 \pm 1.06}$	$\mathbf{11.36 \pm 1.04}$	$\mathbf{11.42 \pm 1.06}$	$\mathbf{11.28 \pm 1.06}$	$\mathbf{11.56 \pm 1.10}$	$\mathbf{11.31 \pm 1.00}$
Abalone	$\mathbf{2.06 \pm 0.10}$	$\mathbf{2.09 \pm 0.10}$	$\mathbf{2.08 \pm 0.10}$	$\mathbf{2.05 \pm 0.10}$	$\mathbf{2.09 \pm 0.10}$	$\mathbf{2.08 \pm 0.10}$
Naval propulsion	$\mathbf{0.02 \pm 0.00}$	$\mathbf{0.02 \pm 0.00}$	38.86 ± 0.60	62.61 ± 1.51	9.40 ± 0.16	3.45 ± 0.08
Forest fire	1.97 ± 0.05	$\mathbf{1.87 \pm 0.02}$	6.43 ± 0.69	10.43 ± 1.11	2.32 ± 0.14	3.32 ± 0.28
Parkinson's	12.17 ± 0.02	12.40 ± 0.10	12.49 ± 0.16	$\mathbf{11.97 \pm 0.03}$	12.60 ± 0.16	12.78 ± 0.16

Table 2: Performance of methods on all datasets w.r.t. RMSE, including 95% confidence intervals. The best scores are in bold.

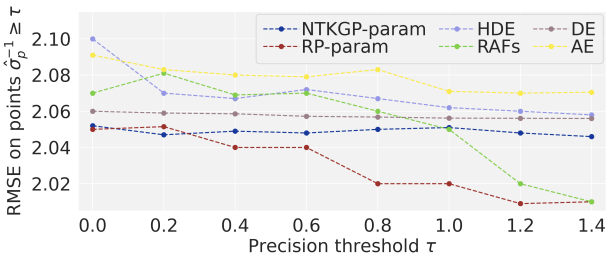
	DE	HDE	AE	NTKGP-p.	RP-p.	RAFs
He et al. 1D	6,2	5,3	4,1	2,2	3,3	**1**,2
Forrester et al. 1D	4,2	5,2	3,2	6,2	2,2	**1**,1
Schaffer N.4 2D	5,1	2,4	6,3	3,1	4,3	**1**,2
Double pendulum 2D	3,1	3,3	2,2	**1**,1	**1**,2	**1**,1
Rastrigin 3D	2,2	**1**,1	3,3	2,2	**1**,1	**1**,1
Ishigami 3D	3,1	5,3	4,2	2,1	**1**,1	**1**,1
Environmental 4D	6,1	5,2	2,1	4,1	3,1	**1**,1
Griewank 4D	3,3	**1**,1	**1**,1	2,2	**1**,1	**1**,2
Roos & Arnold 5D	3,1	**1**,1	4,3	5,1	5,1	2,2
Friedman 5D	5,1	6,1	3,1	4,1	2,1	**1**,1
Planar arm torque 6D	5,1	4,1	3,1	**1**,1	2,2	2,1
Sum of powers 6D	5,1	6,1	3,1	4,1	2,1	**1**,1
Ackley 7D	4,5	**1**,2	2,4	3,5	2,3	**1**,1
Piston simulation 7D	**1**,1	3,2	2,5	2,6	2,3	2,4
Robot arm 8D	5,3	3,1	4,2	**1**,3	2,4	**1**,1
Borehole 8D	2,1	3,1	**1**,4	**1**,5	**1**,2	**1**,3
Styblinski-Tang 9D	3,2	5,5	2,1	4,3	**1**,1	**1**,4
PUMA560 9D	5,2	**1**,1	3,1	4,2	4,1	2,1
Adapted Welch 10D	6,1	2,1	4,3	5,1	3,2	**1**,1
Wing weight 10D	4,4	2,1	**1**,1	3,3	**1**,1	**1**,2
Boston housing	3,1	5,1	2,1	2,1	5,1	**1**,1
Abalone	5,1	6,1	3,1	4,1	**1**,1	2,1
Naval propulsion plant	**1**,1	4,1	3,4	2,5	2,3	2,2
Forest fire	3,2	**1**,1	1,5	1,6	2,3	1,4
Parkinson's	**1**,2	6,3	4,3	5,1	3,3	2,3

Table 3: Rank of the methods corresponding to NLL (left) and RMSE (right). The best overall score is in bold (ties are possible in case of an overlap in confidence intervals).

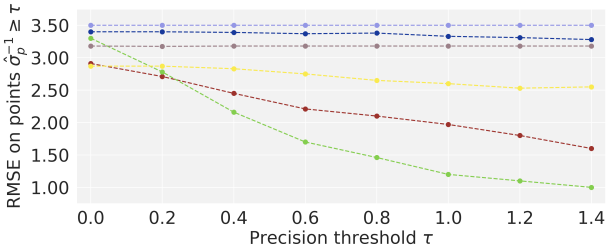
not provide as prominent results with respect to RMSE in the higher dimensional datasets as it does in datasets of lower dimensions, it still achieves better or on par results compared to the state-of-the-art methods. In addition, RAFs Ensemble can be deployed in both complex and straightforward settings. On a related note, while DE struggles when dealing with high multimodality and RP-param underperforms when the dataset has interaction effects (from "others" category), RAFs excels in both such settings.

Scalability to higher dimensions and larger networks. To test the scalability of RAFs Ensemble, we compare it with the strongest baseline, RP-param, on two additional real-world datasets, i.e., a 65-dimensional data with around 20k samples and a 40-dimensional data with almost 40k samples. The former is the superconductivity dataset, where the goal is to predict the critical temperature of superconductors (Hamidieh 2018). The latter summarizes features about articles, where the target is the number of shares in social networks (Fernandes, Vinagre, and Cortez 2015). Both methods utilize the same neural architecture for their base-learners, that is two hidden layers of 128 hidden neurons each, which is more complex than the previous experiments. The conclusion of these experiments is conclusive in favor of our approach. RAFs Ensemble scores NLL of 5.49 and 25.89 on the first and second dataset, respectively, while RP-param scores NLL of over 100 on both datasets.

Confidence vs. Error. We further analyze the relation between the RMSE and the precision thresholds in order to examine the confidence of each method in the prediction



(a) Abalone (real-world)

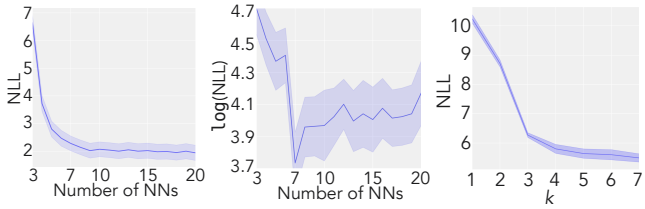


(b) Friedman (synthetic)

Figure 2: Confidence versus error of estimations.

task. Figure 2 displays the confidence versus error plots for one synthetic and one real-world dataset, i.e., Friedman and Abalone. In this figure, for each precision threshold τ, the RMSE is plotted for examples where the predicted precision $\boldsymbol{\sigma}_p^{-2}$ is larger than the threshold τ, demonstrating confidence. In general, reliable estimates are expected to have decreasing error when the confidence is increasing. For Friedman dataset, it is clear that RAFs Ensemble delivers well-calibrated estimates, which is especially in contrast with DE, NTKGP-param, and HDE (Figure 2b). However, for the Abalone data, RP-param demonstrates the most reliable behavior, although RAFs Ensemble meets its performance at the last precision threshold (Figure 2a). Overall, our approach sustains lower error over most precision thresholds compared to the majority of the other methods, and this contrast in performance is emphasized as the predictions get more confident.

Ablation. We study the effect of number of base-learners in the ensemble on the quality of UQ, which also measures the sensitivity of the results to the cardinality of the set of AFs k. We conduct an experiment on two different datasets, one synthetic (PUMA590) and one real-world (Abalone), where the results in terms of NLL are represented in Figure 3. Note that Figure 3b is shown in log-scale for better visibility. According to the theory, in the limit of infinite number of ensemble members, the ensemble error converges to zero (Hansen and Salamon 1990). However, practically speaking, five NNs in the ensemble provide optimal results regarding the trade-off between empirical performance and computational time (Lakshminarayanan, Pritzel, and Blundell 2017), which is also the case in our experiments. This is further confirmed by the plot in Figure 3a. In addition, for the PUMA590 dataset, it seems that RAFs Ensemble's performance is not impacted negatively by the number of NNs



(a) PUMA560 (b) Abalone (c) Superconductivity

Figure 3: The effect of number of NNs in the ensemble in terms of NLL, including the 95% confidence interval.

in the ensemble. Moreover, an interesting observation is the steep through for seven NNs (equal to k) in Figure 3b, which is an indication that there might be a correlation between k and the performance in some cases. A plausible reason for this is the fact that the additional source of randomness is utmostly exploited via a different activation function.

To further confirm the effectiveness of the random activation functions, we evaluate the performance of RAFs Ensemble (of five NNs) in terms of NLL w.r.t. different cardinalities k of the set of AFs. The dataset used for this experiment is the superconductivity data. As the results in Figure 3c clearly suggest, by increasing the cardinality k, NLL has a decreasing pattern, which shows that having more random AFs significantly improves the performance of the ensemble.

Moreover, we combine our approach with RP-param instead of AE to show that RAFs can be methodologically applied to any ensemble technique. We evaluate the performance of this combination on the Parkinson's dataset, using the same network architecture for fair comparison. The obtained results demonstrate that applying RAFs to RP-param leads to reducing the original NLL score of > 100 to 48.66, which is in line with the results we get when comparing AE with RAFs Ensemble and is a further proof that the methodology indeed increases the performance.

Conclusions

We introduced a novel method, Random Activation Functions Ensemble, for a more robust uncertainty estimation in approaches based on neural networks, in which, each network in the ensemble is accomodated with a different (random) activation function to increase the diversity of the ensemble. The empirical study illustrates that our approach achieves excellent results in quantifying both epistemic and aleatoric uncertainty compared to five state-of-the-art ensemble uncertainty quantification methods on a series of regression tasks across 25 datasets, which proved there does not have to be a trade-off between simplicity and strong empirical performance. Furthermore, the properties of datasets such as dimensionality or complexity of modeling dynamics do not appear to affect RAFs Ensemble negatively, which also demonstrates robustness in out-of-distribution settings.

References

Abdar, M.; Pourpanah, F.; Hussain, S.; Rezazadegan, D.; Liu, L.; Ghavamzadeh, M.; Fieguth, P. W.; Cao, X.; Khosravi, A.; Acharya, U. R.; Makarenkov, V.; and Nahavandi, S. 2021. A Review of Uncertainty Quantification in Deep Learning: Techniques, Applications and Challenges. *Inf. Fusion*, 76: 243–297.

Abiodun, O. I.; Jantan, A. B.; Omolara, A. E.; Dada, K. V.; Mohamed, N.; and Arshad, H. 2018. State-of-the-art in artificial neural network applications: A survey. *Heliyon*, 4.

Althoff, D.; Rodrigues, L. N.; and Bazame, H. C. 2021. Uncertainty quantification for hydrological models based on neural networks: the dropout ensemble. *Stochastic Environmental Research and Risk Assessment*, 35: 1051 – 1067.

Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P. F.; Schulman, J.; and Mané, D. 2016. Concrete Problems in AI Safety. *ArXiv*, abs/1606.06565.

Bijak, J.; and Hilton, J. 2021. Uncertainty Quantification, Model Calibration and Sensitivity. *Towards Bayesian Model-Based Demography*.

Blundell, C.; Cornebise, J.; Kavukcuoglu, K.; and Wierstra, D. 2015. Weight Uncertainty in Neural Networks. *ArXiv*, abs/1505.05424.

Brown, K. E.; Bhuiyan, F. A.; and Talbert, D. A. 2020. Uncertainty Quantification in Multimodal Ensembles of Deep Learners. In *FLAIRS Conference*.

Charpentier, B.; Zügner, D.; and Günnemann, S. 2020. Posterior Network: Uncertainty Estimation without OOD Samples via Density-Based Pseudo-Counts. *ArXiv*, abs/2006.09239.

Coraddu, A.; Oneto, L.; Ghio, A.; Savio, S.; Anguita, D.; and Figari, M. 2014. Machine Learning Approaches for Improving Condition? Based Maintenance of Naval Propulsion Plants. *Journal of Engineering for the Maritime Environment*.

Cortez, P.; and de Jesus Raimundo Morais, A. 2007. A data mining approach to predict forest fires using meteorological data. *EUROSIS-ETI*.

Egele, R.; Maulik, R.; Raghavan, K.; Balaprakash, P.; and Lusch, B. 2021. AutoDEUQ: Automated Deep Ensemble with Uncertainty Quantification. *ArXiv*, abs/2110.13511.

Fan, X.; Zhang, S.; Chen, B.; and Zhou, M. 2020. Bayesian Attention Modules. *ArXiv*, abs/2010.10604.

Fernandes, K.; Vinagre, P.; and Cortez, P. 2015. A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News. In *Portuguese Conference on Artificial Intelligence*.

Forrester, A. I. J.; Sobester, A.; and Keane, A. J. 2008. *Engineering Design via Surrogate Modelling - A Practical Guide*. Wiley.

Hamidieh, K. 2018. A data-driven statistical model for predicting the critical temperature of a superconductor. *Computational Materials Science*.

Hansen, L. K.; and Salamon, P. 1990. Neural Network Ensembles. *IEEE Trans. Pattern Anal. Mach. Intell.*, 12: 993–1001.

He, B.; Lakshminarayanan, B.; and Teh, Y. W. 2020. Bayesian Deep Ensembles via the Neural Tangent Kernel. *ArXiv*, abs/2007.05864.

Heiss, J.; Weissteiner, J.; Wutte, H.; Seuken, S.; and Teichmann, J. 2021. NOMU: Neural Optimization-based Model Uncertainty. *ArXiv*, abs/2102.13640.

Hendrycks, D.; and Gimpel, K. 2016. Gaussian Error Linear Units (GELUs). *arXiv: Learning*.

Hoffmann, L.; Fortmeier, I.; and Elster, C. 2021. Uncertainty quantification by ensemble learning for computational optical form measurements. *Machine Learning: Science and Technology*, 2.

Järvenpää, M.; Vehtari, A.; and Marttinen, P. 2020. Batch simulations and uncertainty quantification in Gaussian process surrogate approximate Bayesian computation. In *UAI*.

Kiureghian, A. D.; and Ditlevsen, O. 2009. Aleatory or epistemic? Does it matter? *Structural Safety*, 31: 105–112.

Klambauer, G.; Unterthiner, T.; Mayr, A.; and Hochreiter, S. 2017. Self-Normalizing Neural Networks. *ArXiv*, abs/1706.02515.

Krogh, A.; and Vedelsby, J. 1994. Neural Network Ensembles, Cross Validation, and Active Learning. In *NIPS*.

Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In *NIPS*.

Little, M. A.; McSharry, P. E.; Roberts, S. J.; Costello, D.; and Moroz, I. M. 2007. Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection. *BioMedical Engineering OnLine*, 6: 23 – 23.

Nash, W.; Sellers, T.; Talbot, S.; Cawthorn, A.; and Ford, W. 1994. 7he Population Biology of Abalone (Haliotis species) in Tasmania. I. Blacklip Abalone (H. rubra) from the North Coast and Islands of Bass Strait. *Sea Fisheries Division, Technical Report No*, 48.

Osband, I.; Aslanides, J.; and Cassirer, A. 2018. Randomized Prior Functions for Deep Reinforcement Learning. In *NeurIPS*.

Pearce, T.; Zaki, M.; Brintrup, A.; and Neely, A. D. 2018. Uncertainty in Neural Networks: Bayesian Ensembling. *ArXiv*, abs/1810.05546.

Rahaman, R.; and Thiery, A. H. 2021. Uncertainty Quantification and Deep Ensembles. In *NeurIPS*.

Ramachandran, P.; Zoph, B.; and Le, Q. V. 2018. Searching for Activation Functions. *ArXiv*, abs/1710.05941.

Sensoy, M.; Kandemir, M.; and Kaplan, L. M. 2018. Evidential Deep Learning to Quantify Classification Uncertainty. *ArXiv*, abs/1806.01768.

Srivastava, N.; Hinton, G. E.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15: 1929–1958.

Sullivan, T. J. 2015. *Introduction to Uncertainty Quantification*. Springer.

Turian, J. P.; Bergstra, J.; and Bengio, Y. 2009. Quadratic Features and Deep Architectures for Chunking. In *NAACL*.

Volodina, V.; and Challenor, P. 2021. The importance of uncertainty quantification in model reproducibility. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379(2197).

Wenzel, F.; Snoek, J.; Tran, D.; and Jenatton, R. 2020. Hyperparameter Ensembles for Robustness and Uncertainty Quantification. *ArXiv*, abs/2006.13570.

Zhang, C.; and Ma, Y. 2012. *Ensemble Machine Learning: Methods and Applications*. Springer.

Zhou, Z.-H. 2012. *Ensemble Methods: Foundations and Algorithms*. Chapman & Hall/CRC, 1st edition. ISBN 1439830037.