Full Bayesian Significance Testing for Neural Networks

Zehua Liu[1], Zimeng Li[1], Jingyuan Wang[1,2,3*], Yue He[4]

[1]School of Computer Science and Engineering, Beihang University, Beijing, China
[2]School of Economics and Management, Beihang University, Beijing, China
[3]Key Laboratory of Data Intelligence and Management (Beihang University), Ministry of Industry and Information Technology, Beijing, China
[4]Department of Computer Science and Technology, Tsinghua University, Beijing, China

Abstract

Significance testing aims to determine whether a proposition about the population distribution is the truth or not given observations. However, traditional significance testing often needs to derive the distribution of the testing statistic, failing to deal with complex nonlinear relationships. In this paper, we propose to conduct Full Bayesian Significance Testing for neural networks, called *n*FBST, to overcome the limitation in relationship characterization of traditional approaches. A Bayesian neural network is utilized to fit the nonlinear and multi-dimensional relationships with small errors and avoid hard theoretical derivation by computing the evidence value. Besides, *n*FBST can test not only global significance but also local and instance-wise significance, which previous testing methods don't focus on. Moreover, *n*FBST is a general framework that can be extended based on the measures selected, such as Grad-*n*FBST, LRP-*n*FBST, DeepLIFT-*n*FBST, LIME-*n*FBST. A range of experiments on both simulated and real data are conducted to show the advantages of our method.

Introduction

Significance testing aims to determine whether a proposition about the population distribution[1] is true or false given observations, which is widely used in many scientific fields, such as social sciences(Orlitzky 2012; Ortega and Navarrete 2017) and medical research(Matthews et al. 1990; Rutledge and Loh 2004). For example, it is often used to evaluate the efficacy of new treatments or drugs. First, clinical trials are performed to compare the response of patients treated with a new therapy against a control group. Then, significance testing is used as an analytical tool to determine whether the observed improvement in the treatment group is significant, which provides evidence that the new therapy is effective.

To attain a proper testing result, a golden standard is to recover the true data generation model f_0 behind the population distribution, then justify the proposition according to f_0 directly. For this purpose, a number of significance testing approaches are proposed under different assumptions about f_0 (Gozalo 1993; Lavergne and Vuong 1996; Racine 1997). However, simple assumptions can hardly fit the real situation, while complex assumptions make it hard to derive the theoretical distribution of the testing statistic. Recently, (Horel and Giesecke 2020) provides a provable solution of significance testing in nonlinear cases, but it suffers from the computational difficulty in statistics, only addressing a limited function space. There is still a gap to be solved for significance testing in the general correlations.

Existing significance testing methods only focus on global propositions. However, some propositions that are invalid globally can still hold on and contribute to a certain sub-population. For example, clinical trials have shown that a drug is effective in treating cancer, but not in some individuals. For a better justification in nonlinear cases, a significance testing approach should verify the correctness of a proposition in population distribution and sub-population distribution respectively.

To deal with the complicated real data in wide applications(He et al. 2020; Wang et al. 2018, 2020a, 2023), we introduce deep neural networks into the significance testing to capture the nonlinear correlations. To overcome the barrier of computing statistics under the complex fitting functions, we solve the significance testing problem from the Bayesian perspective (Kass and Raftery 1995), and propose a novel approach that conducts the Full Bayesian Significance Testing for neural networks, abbreviated as *n*FBST (neural FBST). Given the testing statistics, *n*FBST can test the correctness of a proposition for both population-level or sub-population-level problems, by comparing the posterior probabilities of it and its opposite. In addition, *n*FBST is a general framework that can be extended based on different testing statistics, such as Grad-*n*FBST, LRP-*n*FBST, DeepLIFT-*n*FBST, LIME-*n*FBST, and so on. A range of experiments on both simulated and real data are conducted to show the advantages of our method. The main contributions can be summarized as follows:

- We are the first to introduce deep neural networks into significance testing. Our approach replaces complicated theoretical derivation by fitting distributions in a Bayesian way, and the neural network serves as a good estimator of f_0 without assuming specific forms.
- We design a complete procedure using Full Bayesian Significance Testing for Neural Networks, namely *n*FBST. It is a general framework that can be extended

*Corresponding author (jywang@buaa.edu.cn)

[1]https://online.stat.psu.edu/stat462/node/249/

based on different implementations and different testing statistics, such as Grad-*n*FBST, LRP-*n*FBST, DeepLIFT-*n*FBST, LIME-*n*FBST, and so on.

- Our proposed *n*FBST can solve both local and global significance testing problems while previous methods only focus on the latter. Under non-linear assumptions, global significance may be inconsistent with local or instance-wise significance.
- We conduct extensive experiments to verify the advantage of our method on better testing results.

Theoretical Method

Classical Frequentist Significance Test

We denote $f_0 : \mathcal{X} \subset \mathbb{R}^d \to \mathbb{R}$ as the underlying and unknown conditional mean function, namely $\mathrm{E}[\mathrm{Y}|\mathrm{X} = \mathrm{x}]$, of the population (X, Y). Then, we consider the data generation process of (X, Y) as follows:

$$y = f_0(X) + \epsilon, \tag{1}$$

where ϵ is a random error such that $\mathrm{E}[\epsilon|\mathrm{X}] = \mathrm{E}[\epsilon] = 0$. Significance testing first defines a testing statistic η, then proposes two contradictory propositions (or hypotheses) H_0 and H_1, which represent the null hypothesis and the alternative hypothesis respectively. Classical significance testing is regarded as a procedure for measuring the consistency of data with the null hypothesis by the calculation of a *p-value* (tail area under the null hypothesis) (De Bragança Pereira and Stern 1999). The process is as follows:

- First, we make assumptions about the population distribution of f_0 and denote it as $f_0(\beta)$, whose parameters are β. Then we derive the theoretical distribution of $\eta(\beta)$ under the assumptions.
- Second, based on the observed data $\mathcal{D}$, we fit an optimal estimator $\hat{f}(\hat{\beta})$ as an approximation function of f_0, whose parameters are $\hat{\beta}$.
- Third, we calculate $\eta(\hat{\beta})$ and *p-value* further to determine whether the distribution of $\eta(\beta)$ is reasonable under the null hypothesis using sample information $\eta(\hat{\beta})$.

The problem of testing the significance of a feature is:

$$H_0 : \eta(\beta) = 0 \quad H_1 : \eta(\beta) \neq 0, \tag{2}$$

where $\eta(\beta)$ is a measure of feature importance. For example, if we assume f_0 satisfies linear relationships as follows:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_d x_d + \epsilon. \tag{3}$$

Whether the coefficient of a feature x_j is equal to zero determines its significance, that is $\eta(\beta) = \beta_j$. However, there are two main defects in classical significance testing.

- First, the effectiveness of classical significance testing is based on reasonable assumptions about f_0. However, it is difficult to find such precise assumptions when the data distribution is actually complicated.
- Second, some models, such as deep learning, excel in accurately fitting complex data distributions. However, the more complex assumption of f_0, the more computational theoretical distribution of $\eta(\beta)$, even intractable.

Full Bayesian Significance Test

In order to solve the problems, we adopt the Full Bayesian significance Testing (FBST) (De Bragança Pereira and Stern 1999; de B. Pereira, Stern, and Wechsler 2008). FBST is a statistical methodology that allows for the testing of precise hypotheses in a Bayesian framework. Here, "full" means that one only needs to use the posterior distribution to test without the specific assumptions for f_0. In contrast to classical significance testing which uses *p-value* to reject or fail to reject the null hypothesis, FBST provides a measure of evidence in favor of or against the null hypothesis, taking into account prior information and the strength of observations.

Let $P(H)$ be the prior of the hypothesis H and $P(\mathcal{D}|H)$ be the likelihood function of H given observations $\mathcal{D}$. The posterior probability distributions for the null and alternative hypotheses are then calculated using Bayes' theorem as

$$P(H|\mathcal{D}) = \frac{P(\mathcal{D}|H) \times P(H)}{P(\mathcal{D})} \propto P(\mathcal{D}|H) \times P(H). \tag{4}$$

It is consistent with the process by which people adjust their assessments in response to observed data. The evidence in favor of the null hypothesis is quantified by the Bayes Factor. Its value reflects which proposition is more likely under the observed data. If it is greater than 1, we believe it provides evidence in favor of H_0. The evidence is moderate if it is greater than 3 and strong if it is greater than 10 (Jeffreys 1998). On the contrary, it provides evidence against H_0 if it is smaller than 1, moderate evidence for less than $\frac{1}{3}$, and strong evidence for less than 0.1.

From the above analysis, we can conclude that FBST doesn't need to assume a specific distribution form of f_0, but calculate $P(\mathcal{D}|H_0)$ and $P(\mathcal{D}|H_1)$ instead. In other words, the current goal is to obtain a good estimator to fit $P(\mathcal{D}|H)$.

Approximate the Distribution of Testing Statistics

According to the universal approximation theorem, neural networks with appropriate size can approximate an extensive class of functions to a desired degree of accuracy (Hornik, Stinchcombe, and White 1989). In this paper, we propose to use Bayesian neural networks to fit the likelihood $P(\mathcal{D}|H)$. As a technique that combines Bayesian Theory and neural networks, Bayesian neural networks can fit complex relationships and produce a probability distribution over model parameters θ that expresses our beliefs regarding how likely the different parameter values are.

Given a dataset $\mathcal{D} = \{(X^{(1)}, y^{(1)}), \ldots, (X^{(n)}, y^{(n)})\}$, *n*FBST first uses Bayesian neural networks, whose parameters are θ, to fit $\mathcal{D}$. Before training, a prior distribution is assigned to model parameters θ as an initial belief $\pi(\theta)$ according to experience. This belief is gradually adjusted to fit data $\mathcal{D}$ by using the Bayesian rule. The final belief is presented as the posterior distribution

$$P(\theta|\mathcal{D}) = \frac{P(\mathcal{D}|\theta)\pi(\theta)}{P(\mathcal{D})} = \frac{\pi(\theta)\prod_{i=1}^{n} P(y^{(i)}|X^{(i)}, \theta)}{\int_{\Theta}\prod_{i=1}^{n} P(y^{(i)}|X^{(i)}, \theta)\mathrm{d}\theta}, \tag{5}$$

where Θ is the parameter space. Given a new case X, the prediction made by the Bayesian neural network is the

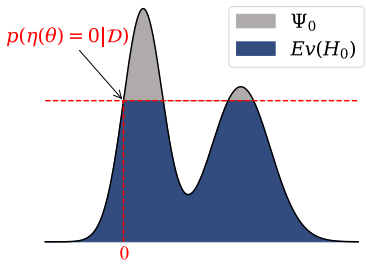


Figure 1: Bayesian evidence calculated based on the distribution of $\eta(\theta)$.

weighted average of an ensemble

$$P(y|X,\mathcal{D}) = \int_{\Theta} P(y|X,\theta)P(\theta|\mathcal{D})\mathrm{d}\theta. \tag{6}$$

Then, based on the posterior distribution of θ, we obtain the posterior distribution of the testing statistic $\eta(\theta)$ and denote $p(\eta(\theta)|\mathcal{D})$ as its probability density. The testing problem is formulated as:

$$H_0 : \eta(\theta) = 0 \quad H_1 : \eta(\theta) \neq 0 \tag{7}$$

We denote the whole space of $\eta(\theta)$ as Ψ such that $\eta(\theta) \in \Psi$. Then, we define the region whose probability greater than $p(\eta(\theta) = 0|\mathcal{D})$ according to the following formula:

$$\Psi_0 = \{\eta(\theta) : p(\eta(\theta)|\mathcal{D}) > p(\eta(\theta) = 0|\mathcal{D})\}, \tag{8}$$

where $p(\eta(\theta) = 0|\mathcal{D})$ is the maximum of the posterior density under the null hypothesis H_0 (Figure 1). The Bayes Factor is not the only way to calculate evidence and its result is influenced by prior distribution. In our method, we adopt a more flexible valid Bayesian evidence for the null hypothesis provided by (De Bragança Pereira and Stern 1999):

$$\begin{aligned} Ev(H_0) &= 1 - \int_{\Psi_0} p\Big(\eta(\theta)|\mathcal{D}\Big)\mathrm{d}\eta(\theta) \\ &= 1 - \int_{\Psi} \mathbb{1}\Big(\eta(\theta) \in \Psi_0\Big) p\Big(\eta(\theta)|\mathcal{D}\Big)\mathrm{d}\eta(\theta). \end{aligned} \tag{9}$$

Using the Monte Carlo method, the above formula can be further simplified to

$$\begin{aligned} Ev(H_0) &\approx 1 - \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\Big(\eta_i(\theta) \in \Psi_0\Big) \\ &= 1 - \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\Big(p(\eta_i(\theta)|\mathcal{D}) > p(0|\mathcal{D})\Big), \end{aligned} \tag{10}$$

where $\eta_i(\theta)$ is sampled m times based on the posterior probability density of $\eta(\theta)$. The result of Eq (10) is called Bayesian evidence, whose value is between 0 and 1. The closer the Bayesian evidence to 1, the more likely to accept H_0. The closer the Bayesian evidence to 0, the more likely to reject H_0. Moreover, we have mathematically proven that under certain constraints, as the sample size approaches infinity, $Ev(H_0)$ for insignificant features converges to 1. The detailed process of proof is provided in the Appendix.

Implementation Approach

Calculate the Distribution of Testing Statistics

So far, we have clarified the entire process of FBST, but there still remain two implementation details that need to be elaborated on. First, to perform *n*FBST to deal with the testing problem Eq (7), we need to calculate the posterior distribution of θ. Second, after obtaining the distribution of θ, we need to calculate the distribution of the testing statistic $\eta(\theta)$.

In practice, it is intractable to solve the integral in Eq (5). A popular way, known as Variational Inference (VI), entails approximating the real but intractable posterior distribution with a tractable distribution called variational distribution (Blei, Kucukelbir, and McAuliffe 2017; Jaakkola and Jordan 2000). Therefore, Eq (5) could be efficiently approximated. Formally, variational family $Q = \{q_\vartheta : \vartheta \in \Gamma\}$ is a predefined family of tractable distributions on model parameter space Θ, where ϑ is the parameter of variational distribution and Γ is the range of ϑ. The optimal variational distribution q_{ϑ^*} is chosen from Q such that

$$\vartheta^* = \arg\min_{\vartheta\in\Gamma} \mathrm{KL}(q_\vartheta(\theta)\|P(\theta|\mathcal{D})). \tag{11}$$

KL divergence describes the "distance" between two distributions. We set diagonal Gaussian distributions as the prior and variational families of parameter θ. This assumption is common in many works (Blundell et al. 2015; Kendall and Gal 2017). Under this assumption, applying Bayesian rule Eq (5), Eq (11) can be further simplified as

$$\vartheta^* = \arg\min_{\vartheta\in\Gamma} -\mathbb{E}[\log P(\mathcal{D}|\theta)] + \mathrm{KL}(q_\vartheta(\theta)\|\pi(\theta)) + \log P(\mathcal{D}). \tag{12}$$

The derivation is shown in Appendix. The first term is related to data (such as MSE for regression task); The second term is only related to parameters θ like regularization term, and the third term is a constant. In the end, we finish approximating the posterior distribution of parameters $P(\theta|\mathcal{D})$ with variational distribution $q_{\vartheta^*}(\theta)$.

We adopt Kernel Density Estimation (KDE) (Scott 1979; Parzen 1962) to estimate the posterior probability density of $\eta(\theta)$, which is a common non-parametric method. The process is as follows:

- First, draw samples of parameters θ with size m from the approximate posterior distribution $q_{\vartheta^*}(\theta)$ randomly, that is $\{\theta_1, \ldots, \theta_m\}$.
- Second, calculate $\{\eta(\theta_1), \ldots, \eta(\theta_m)\}$ to obtain sample from the posterior of $\eta(\theta)$.
- Third, estimate $p(\eta(\theta)|\mathcal{D})$ using KDE

$$p\Big(\eta(\theta)|\mathcal{D}\Big) = \frac{1}{mh}\sum_{i=1}^{m} K\left(\frac{\eta(\theta) - \eta(\theta_i)}{h}\right), \tag{13}$$

  where K is the kernel function, and h is the window width (also known as bandwidth). The commonly used kernel function is the Gaussian kernel function.

Finally, by calculating Bayesian evidence in Eq (10), we finish the entire process of *n*FBST. *n*FBST is a general and flexible framework that can be easily extended based on different implementations, including VI and KDE. In the case

of VI, we have derived the detailed training procedure in the Appendix. The approximation between variational and posterior distributions can be gauged through the prediction error. As for KDE, its robust theoretical underpinning guarantees convergence and consistency, as evidenced by the work of (Parzen 1962). Consequently, the margin of error in our approach remains within a reasonable range.

Design of Testing Statistics

To test the significance of a feature x_j, we first need a reasonable measure as the testing statistic to represent the relationship between x_j and y. *n*FBST is a flexible framework that can be applied to global significance, local significance, and instance-wise significance testing problems. The weighted average of the partial derivative, with weights defined by a positive measure μ, is adopted as the testing statistic in (Horel and Giesecke 2020):

$$\eta(\theta) = \int_{\mathcal{X}} \left(\frac{\partial f_0(x)}{\partial x_j} \right)^2 d\mu(x). \tag{14}$$

This reflects the global significance of the overall data, whose value will not change when the data distribution is fixed. In non-linear contexts, the significance of a feature in sub-population distribution is dynamic and varies with its range. We consider a simple case $f_0(X) = \mathrm{ReLU}(x_0)$ where $x_0 \sim \mathcal{N}(0,1)$. It can be calculated that $\eta(\theta) = \frac{1}{2}$, which is a constant without considering the specific values of x_0. If we define $\mathcal{X}_i \subseteq \mathcal{X}$, the local significance testing statistic value is dynamic as $\mathcal{X}_i$ varies, that is

$$\eta(\theta, \mathcal{X}_i) = \int_{\mathcal{X}_i} \left(\frac{\partial f_0(X)}{\partial x_j} \right)^2 d\mu(X). \tag{15}$$

In this example, if we define $\mathcal{X}_1 = \{X : x_0 < 0\}, \mathcal{X}_2 = \{X : x_0 \geq 0\}$, we obtain $\eta(\theta, \mathcal{X}_1) = 0, \eta(\theta, \mathcal{X}_2) = \frac{1}{2}$. It is clear that under partial derivative settings, x_0 is insignificant when its value is less than zero, but significant when its value is greater than zero. Further, $\mathcal{X}_i$ can contain only one data, that is the instance-wise significance testing statistic

$$\eta(\theta, x_j) = \frac{\partial f_0(X)}{\partial x_j}. \tag{16}$$

*n*FBST is a general framework and supports significance testing based on various feature importance measures. In our implementation, we select LRP(Binder et al. 2016), LIME(Ribeiro, Singh, and Guestrin 2016), and DeepLIFT(Shrikumar, Greenside, and Kundaje 2017) as testing statistics and the corresponding methods are called LRP-*n*FBST, LIME-*n*FBST, and DeepLIFT-*n*FBST respectively. Eq (16) uses gradient as the testing statistic and we name it Grad-*n*FBST.

For global significance, Eq (14) is difficult to capture enough information when the distribution of the testing statistic is complex. Therefore, we propose a Quantile-based Global Significance, namely Q-GS. First, we sort all Bayesian evidence of instance-wise significance in descending order. Then, we set a threshold λ and select the quantile of the sorted evidence. It satisfies that the percentage of evidence over it reaches the threshold λ.

Experiments

Toy Example

We consider the following data generation process

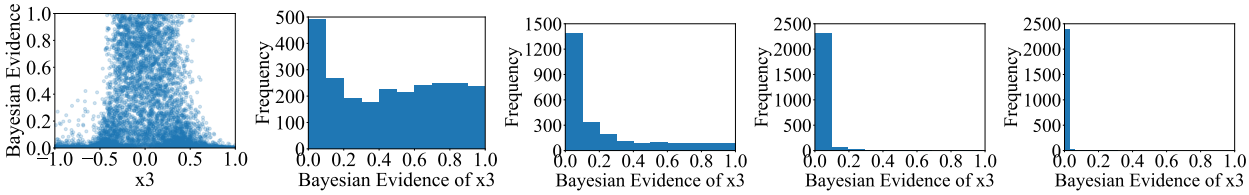
$$y = 8 + x_0^2 + x_1 x_2 + \cos(x_3) + \exp(x_4 x_5) + 0.1 x_6 + 0 x_7 + \epsilon, \tag{17}$$

where $X = [x_0, x_1, \ldots, x_7] \sim \mathcal{U}(-1,1)^8, \epsilon \sim \mathcal{N}(0,1)$. The variable x_7 has no influence on y. Our goal is to differentiate x_7 from other features, that is to determine x_7 as insignificance but others as significance. We compare the following three classical testing methods as the baselines:

- **Bootstrap (Efron 1979).** It gets the distribution of the testing statistic from samples repeatedly drawn from the original data and simulates the mean and variance of the population to perform *Z*-test.
- **Likelihood ratio test (Fisher 1922).** By training an unconstrained model incorporating all variables and a nested model with restricted variables, and comparing their likelihoods, we will obtain the standard asymptotic chi-square distribution if the unconstrained model is assumed correctly.
- ***t*-test for linear models (Student 1908).** It calculates the estimated coefficient divided by its standard error and tests the result whether to follow the *t*-distribution.

From table 1 we have the following observations:

- First, for the three classical testing methods, only Bootstrap accurately identifies the global significance of all features. Usually, we set the significance level α and compare *p-value* with it. If the *p-value* is smaller than the significance level, we reject the null hypothesis and accept the alternative hypothesis. That is, the smaller the *p-value*, the more we can determine a feature is significant. If we set $\alpha = 0.05$, the *p-value* of Bootstrap satisfies that only x_7 is greater than α but others are not. However, the likelihood ratio test and *t*-test can hardly distinguish correctly. This is probably because their assumptions about f_0 are too strong and lead to errors.
- Second, all *n*FBST methods based on different testing statistics perform well. Here, we set $\lambda = 0.5$ to obtain Q-GS. It means only when more than half of the instance-wise testing results reflect insignificance, we determine the feature as insignificant globally. The smaller the Q-GS, the less the evidence for H_0, and we tend to reject that the feature is insignificant. The results show all *n*FBST methods provide strong evidence about x_7 for H_0 but little evidence for other features.
- Third, instance-wise significance can provide us with more insights than global significance. For the likelihood ratio test and *t*-test methods, *p-value* for x_3 is high even the highest. We plot scatters of the evidence for x_3 obtained by *n*FBST and plot histograms under different x_3 intervals. As shown in Figure 2, its evidence of Grad-*n*FBST is more concentrated on one when the value of x_3 is close to zero. This is consistent with Eq (17) as $\partial f_0(X)/\partial x_3 = -\sin(x_3)$. We conclude that global significance is more coarse-grained than instance-wise significance due to averaging different situations together.



(a) Scatter of Evidence (b) $0 \leq |x_3| < 0.25$ (c) $0.25 \leq |x_3| < 0.5$ (d) $0.5 \leq |x_3| < 0.75$ (e) $0.75 \leq |x_3| < 1$

Figure 2: Bayesian evidence of x_3 obtained by Grad-*n*FBST under different intervals on the toy example.

Feature	*p-value*			Q-GS($\lambda = 0.5$) by *n*FBST		
	Bootstrap	likelihood ratio test	*t*-test	Grad-*n*FBST	DeepLIFT-*n*FBST	LRP-*n*FBST
x_0	< 0.001	0.696	0.442	<0.001	<0.001	<0.001
x_1	<0.001	0.063	0.592	0.021	<0.001	0.02
x_2	<0.001	0.087	0.598	0.029	<0.001	0.027
x_3	0.027	**0.813**	0.838	0.011	0.043	0.01
x_4	<0.001	0.209	0.604	0.005	0.007	0.003
x_5	<0.001	0.361	0.559	0.006	0.01	0.007
x_6	<0.001	0.318	<0.001	0.278	0.203	0.273
x_7	**0.383**	0.049	**0.922**	**0.637**	**0.637**	**0.617**

Table 1: Global significance testing results of different algorithms on the toy example, the maximum values are bolded.

Simulation Experiments

In this section, we conduct experiments on three simulation datasets for analysis. Compared to three classical testing methods, we found *n*FBST can perfectly distinguish the global significance while others cannot. Compared to five feature importance methods, we found *n*FBST can improve the ability to test instance-wise significance.

Data Generation Process We consider the input features $X = [x_0, x_1, \ldots, x_{99}] \sim \mathcal{U}(-1, 1)^{100}$ and the data generation process

$$y = f_0(X) + \epsilon, \tag{18}$$

where $\epsilon \sim \mathcal{N}(0, 0.01)$ and f_0 is a neural network function whose weights and biases are initialized randomly. Then only the last fifty features are insignificant by setting the corresponding weights to zero. Our goal is to select these fifty insignificant features accurately.

According to the above generation process, we generate two sets of 10,000 independent samples, namely Dataset 1 and Dataset 2. The difference between them is that the structure of f_0 for Dataset 1 is three hidden layers of 20 nodes but three hidden layers of 16 nodes for Dataset 2. In our experiments, we only adopt the structure with three hidden layers of 20 nodes as the trained model to simulate conditions where it has the same or different structure from f_0. Then, we reduce the data size to one-tenth of the Dataset 2, that is 1,000 independent samples, namely Dataset 3, to simulate the scenario of small data.

Test the Global Significance For the global significance of each feature, there are two possible testing results, insignificant or significant. Therefore, we can consider the task of testing as a binary classification problem. Specifically, significant represents the positive class, and insignificant represents the negative class. TP means identifying significant features correctly. TN means identifying insignificant features correctly. FN means misidentifying a feature that should be significant as an insignificant feature. FP means misidentifying a feature that should be insignificant as a significant feature. Precision=TP/(TP+FP), reflects the accuracy of correctly testing significant features. Recall=TP/(TP+FN), reflects the completeness of correctly identifying significant features. F1-score combines precision and recall to calculate the harmonic mean.

Table 2 shows these metrics on three simulation datasets compared to three classical testing methods. As with the settings of "Toy Example", we set the significance level for classical testing methods as 0.05 and $\lambda = 0.5$ to obtain Q-GS. First, Bootstrap tends to determine a feature as significant thus resulting in high recall but poor precision. On the contrary, the likelihood ratio test tends to determine a feature as insignificant thus resulting in poor recall but fine precision. If we compare comprehensively, *t*-test method outperforms the two methods with the highest F1-score. Second, all *n*FBST methods based on different testing statistics perform perfectly, with the F1-score of 1. Compared to the classical testing methods, the improvement is largely due to the flexible hypothesis of *n*FBST, as neural networks can fit more complex cases of f_0.

Test the Instance-wise Significance For the instance-wise significance of each feature, there are also two possible testing results, insignificant or significant. Specifically, there are 10,000 instances of 100 features for Dataset 1 and

Dataset	Metric	Bootstrap	likelihood ratio test	t-test	Grad-nFBST	DeepLIFT-nFBST	LRP-nFBST	LIME-nFBST
Dataset 1	Precision	0.54	0.91	0.96	1	1	1	1
	Recall	0.98	0.80	0.86	1	1	1	1
	F1-score	0.70	0.85	0.91	1	1	1	1
Dataset 2	Precision	0.55	0.94	0.91	1	1	1	1
	Recall	0.98	0.66	0.86	1	1	1	1
	F1-score	0.71	0.78	0.89	1	1	1	1
Dataset 3	Precision	0.51	0.87	0.90	0.82	0.80	0.83	0.85
	Recall	1	0.54	0.54	0.84	0.90	0.86	0.80
	F1-score	0.68	0.67	0.68	0.83	0.85	0.84	0.82

Table 2: Precision, Recall and F1-score for global significance on different Datasets.

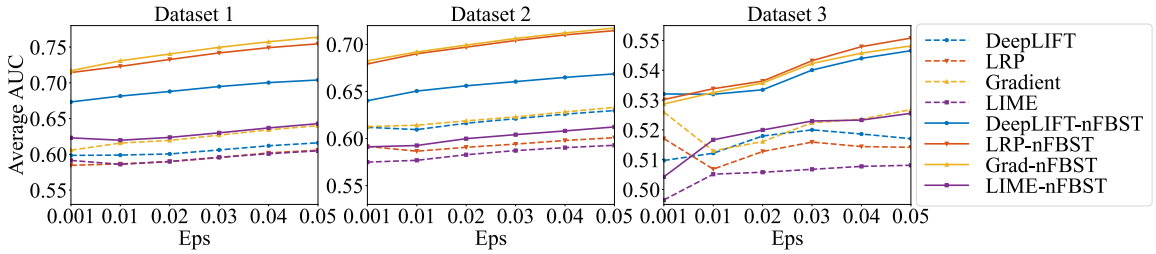


Figure 3: Average AUC of all features for instance-wise significance before and after nFBST under different eps.

Dataset 2. As the last fifty features are insignificant by setting the corresponding weights to zero, our focus is mainly on the first fifty features. Because the instance-wise significance of these features varies with their values. First, we calculate the gradients of f_0 on each instance and adjust different precision thresholds, namely eps, to label the instance-wise significance. If the gradient is less than eps, we label it insignificant, otherwise significant. Then, we evaluate the performance by ROC and AUC. Most existing testing methods don't distinguish global significance and instance-wise significance and only focus on the former. Therefore, we select feature importance analysis methods as baselines. They assign a feature importance score for a prediction, which reflects the significance of the feature learned from the model.

From Figure 3, we have the following observations:

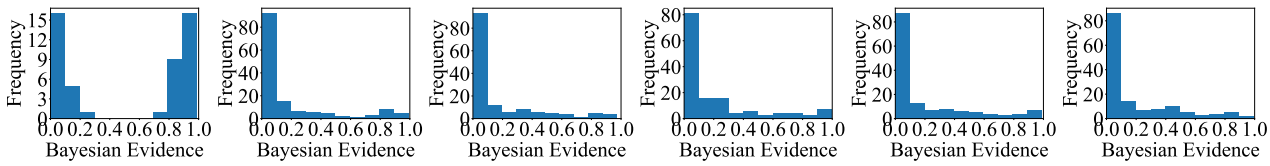
- First, diverse measure-based nFBST consistently surpasses primary feature importance methods across various epsilon settings. Through the comparison of Grad and Grad-nFBST, DeepLIFT and DeepLIFT-nFBST, LRP and LRP-nFBST, LIME and LIME-nFBST, we infer that the integration of nFBST enhances the capacity to discern instance-wise significant and insignificant features. Moreover, it can be concluded that our approach remains unaffected by the precision of the ground truth. The distinct AUC for each feature under varying epsilon values is presented in the Appendix.
- Second, LIME and LIME-nFBST perform worse than other methods. That's because LIME is a perturbation-based method that constructs a local linear model based on the data collected by perturbing near sample points. Its performance is limited by sampling efficiency.

Real World Experiments

In this section, we analyzed the performance of nFBST in real world scenarios through UCI and image datasets. On energy efficiency, we focused on analyzing feature x_8 and found the instance-wise testing results are consistent with its physical truth. On MNIST, by comparing different feature importance methods, we found nFBST recognized the object information in the image more prominently.

The energy efficiency dataset comprises 768 samples and 8 features. It aims to predict the dependent target y (HL, heating load), which determines the specifications of the heating equipment needed to maintain comfortable indoor air conditions. The descriptions of the features and target are shown in the Appendix. From figure 4, we find that testing results are more concentrated around one when x_8 equals zero, while others are not. It indicates that the instance-wise significance of x_8 is different under different values, insignificant if its value is zero. The research in (Tsanas and Xifara 2012) confirms our findings. There are six possible values for x_8 (0, 1, 2, 3, 4, 5) in total. When x_8 equals zero, it means no glazing areas and that's why x_8 doesn't make sense in this situation. In conclusion, nFBST can effectively discover instance-wise significance in real world data.

The testing problem for MNIST is defined as testing each pixel of a digit image and distinguishing significant pixels from insignificant pixels for the target. It involves a weakly supervised semantic segmentation task in computer vision. Ideally, the pixels related to the target class should be assigned higher scores than the background pixels. For feature importance analysis methods, they generate a saliency map based on feature importance scores. For nFBST, Bayesian



(a) $x_8 = 0$ (b) $x_8 = 1$ (c) $x_8 = 2$ (d) $x_8 = 3$ (e) $x_8 = 4$ (f) $x_8 = 5$

Figure 4: Histograms of gradient distributions for different values of x_8 on energy efficiency dataset.

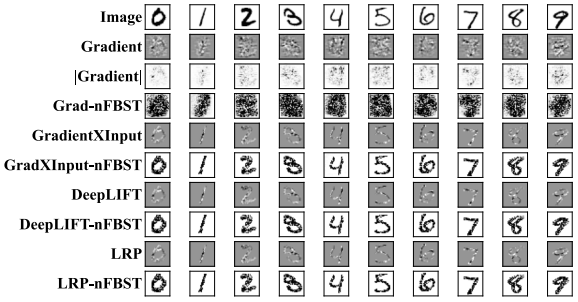


Figure 5: Visualization of scores calculated by different methods for the target class.

evidence represents the evidence supporting H_0, which is pixel-wise insignificance in this task. Figure 5 shows that the object pixels are more prominent after using *n*FBST. Besides, Grad-*n*FBST identifies a messy area because the primary gradient method recognizes poorly. When we multiply it with the input, the performance is improved and this problem doesn't exist in other methods (LRP and DeepLIFT).

Related Works

In recent years, there has been an increasing amount of literature on the interpretability of deep learning (Wang et al. 2019a,b, 2020b, 2021, 2022; Wang, Wu, and Zhao 2021; Wang, Feng, and Wu 2019; Cong et al. 2021; Ji et al. 2020, 2022), one of which is feature importance analysis. The first group propagates an importance score from the output neuron backward to the input. Most of them are gradient-based, including saliency map(Simonyan, Vedaldi, and Zisserman 2013), deconvolution(Zeiler and Fergus 2014), guided back-propagation(Springenberg et al. 2014) and integrated gradients(Sundararajan, Taly, and Yan 2017). The first three methods have different strategies to calculate the gradient when passing through the ReLU layer, but cannot show negative contributions and face discontinuity. The integrated gradients method increases the computational cost by computing the integral. Another approach is the Layer-wise Relevance Propagation proposed by (Binder et al. 2016). (Kindermans et al. 2016) shows that the LRP rules for ReLU networks are equivalent within a scaling factor to gradient $\times$ input in some conditions. Moreover, DeepLIFT (Shrikumar, Greenside, and Kundaje 2017) and SHAP (Lundberg and Lee 2017) do not compute gradients but are also based on back-propagation. In contrast, the second group of methods makes perturbations to individual inputs or neurons(Zeiler and Fergus 2014). A typical approach is LIME (Ribeiro, Singh, and Guestrin 2016), where data are collected by perturbing near sample points to construct a local linear model. However, it is computationally expensive and requires a large number of samples to obtain reliable results.

The above methods aim to explain the predictions of a model locally at a specific instance, while others aim to understand how the model works globally. The partial dependence plot (PDP) shows the marginal impact of one or two features on the model prediction (Friedman 2001)(Greenwell, Boehmke, and McCarthy 2018). (Datta, Sen, and Zick 2016) measures the impact by calculating the difference in the quantity of interest when the data is generated according to the true distribution and the hypothetical distribution designed deliberately. SP-LIME extends LIME to global by selecting typical points(Ribeiro, Singh, and Guestrin 2016).

There is also prior work treating the significance of variables. One is to regard neural networks as parametric formulations (Olden and Jackson 2002; White 1989a,b; Vuong 1989) but restricts the model structure. The testing statistic is not necessarily identifiable due to the non-identifiability of neural networks. The other is to regard neural networks as nonparametric models (Gozalo 1993; Lavergne and Vuong 1996; Yatchew 1992; Fan and Li 1996; Lavergne and Vuong 2000; Racine 1997). However, most of them study in the context of kernel regressions and can be computationally challenging because of Bootstrap. The latest related research restricts the model structure to a single hidden layer and only tests the global significance (Horel and Giesecke 2020).

Conclusion

In this paper, we propose to conduct the Full Bayesian Significance Testing for neural networks, called *n*FBST. It is a general framework that can be extended based on different measures. To the best of our knowledge, we are the first to introduce significance testing into deep neural networks. What's more, it offers a new perspective of exploring knowledge hidden behind the underlying relationship between features and targets in a rigorous way, rather than explaining the estimated relationship which contains estimation errors due to the randomness of the data generation process. Extensive experiments on simulation and real-world datasets confirm the advantages of our proposed approach.

Acknowledgments

Prof. Wang's work was supported by the National Natural Science Foundation of China (No. 72171013, 72222022, 72242101), the Fundamental Research Funds for the Central Universities (YWF-23-L-829) and the DiDi Gaia Collaborative Research Funds. Dr. He's work was supported by the China National Postdoctoral Program for Innovative Talents (BX20230195).

References

Binder, A.; Montavon, G.; Lapuschkin, S.; Müller, K.-R.; and Samek, W. 2016. Layer-wise relevance propagation for neural networks with local renormalization layers. In *International Conference on Artificial Neural Networks*, 63–71. Springer.

Blei, D. M.; Kucukelbir, A.; and McAuliffe, J. D. 2017. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518): 859–877.

Blundell, C.; Cornebise, J.; Kavukcuoglu, K.; and Wierstra, D. 2015. Weight uncertainty in neural networks. *arXiv preprint arXiv:1505.05424*.

Cong, L. W.; Tang, K.; Wang, J.; and Zhang, Y. 2021. AlphaPortfolio: Direct construction through deep reinforcement learning and interpretable AI. *Available at SSRN 3554486*.

Datta, A.; Sen, S.; and Zick, Y. 2016. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE symposium on security and privacy (SP)*, 598–617. IEEE.

de B. Pereira, C. A.; Stern, J. M.; and Wechsler, S. 2008. Can a significance test be genuinely Bayesian? *Bayesian Analysis*, 3(1): 79 – 100.

De Bragança Pereira, C. A.; and Stern, J. M. 1999. Evidence and credibility: Full Bayesian significance test for precise hypotheses. *Entropy*, 1(4): 99–110.

Efron, B. 1979. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1): 1 – 26.

Fan, Y.; and Li, Q. 1996. Consistent model specification tests: omitted variables and semiparametric functional forms. *Econometrica: Journal of the econometric society*, 865–890.

Fisher, R. A. 1922. On the interpretation of χ 2 from contingency tables, and the calculation of P. *Journal of the royal statistical society*, 85(1): 87–94.

Friedman, J. H. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

Gozalo, P. L. 1993. A consistent model specification test for nonparametric estimation of regression function models. *Econometric Theory*, 9(3): 451–477.

Greenwell, B. M.; Boehmke, B. C.; and McCarthy, A. J. 2018. A simple and effective model-based variable importance measure. *arXiv preprint arXiv:1805.04755*.

He, Y.; Cui, P.; Ma, J.; Zou, H.; Wang, X.; Yang, H.; and Yu, P. S. 2020. Learning stable graphs from multiple environments with selection bias. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2194–2202.

Horel, E.; and Giesecke, K. 2020. Significance tests for neural networks. *Journal of Machine Learning Research*, 21(227): 1–29.

Hornik, K.; Stinchcombe, M.; and White, H. 1989. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5): 359–366.

Jaakkola, T. S.; and Jordan, M. I. 2000. Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10(1): 25–37.

Jeffreys, H. 1998. *The theory of probability*. OuP Oxford.

Ji, J.; Wang, J.; Jiang, Z.; Jiang, J.; and Zhang, H. 2022. STDEN: Towards physics-guided neural networks for traffic flow prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 4048–4056.

Ji, J.; Wang, J.; Jiang, Z.; Ma, J.; and Zhang, H. 2020. Interpretable spatiotemporal deep learning model for traffic flow prediction based on potential energy fields. In *2020 IEEE International Conference on Data Mining (ICDM)*, 1076–1081. IEEE.

Kass, R. E.; and Raftery, A. E. 1995. Bayes factors. *Journal of the american statistical association*, 90(430): 773–795.

Kendall, A.; and Gal, Y. 2017. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30.

Kindermans, P.-J.; Schütt, K.; Müller, K.-R.; and Dähne, S. 2016. Investigating the influence of noise and distractors on the interpretation of neural networks. *arXiv preprint arXiv:1611.07270*.

Lavergne, P.; and Vuong, Q. 2000. Nonparametric significance testing. *Econometric Theory*, 16(4): 576–601.

Lavergne, P.; and Vuong, Q. H. 1996. Nonparametric selection of regressors: The nonnested case. *Econometrica: Journal of the Econometric Society*, 207–219.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Matthews, J.; Altman, D. G.; Campbell, M.; and Royston, P. 1990. Analysis of serial measurements in medical research. *British Medical Journal*, 300(6719): 230–235.

Olden, J. D.; and Jackson, D. A. 2002. Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks. *Ecological modelling*, 154(1-2): 135–150.

Orlitzky, M. 2012. How can significance tests be deinstitutionalized? *Organizational Research Methods*, 15(2): 199–228.

Ortega, A.; and Navarrete, G. 2017. Bayesian hypothesis testing: an alternative to null hypothesis significance testing (NHST) in psychology and social sciences. In *Bayesian inference*. IntechOpen.

Parzen, E. 1962. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3): 1065–1076.

Racine, J. 1997. Consistent significance testing for nonparametric regression. *Journal of Business & Economic Statistics*, 15(3): 369–378.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.

Rutledge, T.; and Loh, C. 2004. Effect sizes and statistical testing in the determination of clinical significance in behavioral medicine research. *Annals of Behavioral Medicine*, 27(2): 138–145.

Scott, D. W. 1979. On optimal and data-based histograms. *Biometrika*, 66(3): 605–610.

Shrikumar, A.; Greenside, P.; and Kundaje, A. 2017. Learning important features through propagating activation differences. In *International conference on machine learning*, 3145–3153. PMLR.

Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Springenberg, J. T.; Dosovitskiy, A.; Brox, T.; and Riedmiller, M. 2014. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*.

Student. 1908. The probable error of a mean. *Biometrika*, 6(1): 1–25.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, 3319–3328. PMLR.

Tsanas, A.; and Xifara, A. 2012. Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. *Energy and buildings*, 49: 560–567.

Vuong, Q. H. 1989. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society*, 307–333.

Wang, J.; Feng, K.; and Wu, J. 2019. SVM-based deep stacking networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 5273–5280.

Wang, J.; Ji, J.; Jiang, Z.; and Sun, L. 2022. Traffic flow prediction based on spatiotemporal potential energy fields. *IEEE Transactions on Knowledge and Data Engineering*.

Wang, J.; Peng, Z.; Wang, X.; Li, C.; and Wu, J. 2020a. Deep fuzzy cognitive maps for interpretable multivariate time series prediction. *IEEE transactions on fuzzy systems*, 29(9): 2647–2660.

Wang, J.; Wang, Z.; Li, J.; and Wu, J. 2018. Multilevel wavelet decomposition network for interpretable time series analysis. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2437–2446.

Wang, J.; Wu, N.; Lu, X.; Zhao, W. X.; and Feng, K. 2021. Deep Trajectory Recovery with Fine-Grained Calibration using Kalman Filter. *IEEE Transactions on Knowledge and Data Engineering*, 33(3): 921–934.

Wang, J.; Wu, N.; and Zhao, W. X. 2021. Personalized route recommendation with neural network enhanced search algorithm. *IEEE Transactions on Knowledge and Data Engineering*, 34(12): 5910–5924.

Wang, J.; Wu, N.; Zhao, W. X.; Peng, F.; and Lin, X. 2019a. Empowering A* search algorithms with neural networks for personalized route recommendation. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 539–547.

Wang, J.; Wu, Y.; Li, M.; Lin, X.; Wu, J.; and Li, C. 2020b. Interpretability is a kind of safety: An interpreter-based ensemble for adversary defense. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 15–24.

Wang, J.; Yang, C.; Jiang, X.; and Wu, J. 2023. WHEN: A Wavelet-DTW Hybrid Attention Network for Heterogeneous Time Series Analysis. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2361–2373.

Wang, J.; Zhang, Y.; Tang, K.; Wu, J.; and Xiong, Z. 2019b. AlphaStock: A Buying-Winners-and-Selling-Losers Investment Strategy Using Interpretable Deep Reinforcement Attention Networks. In *Proceedings of the 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '19, 1900–1908. New York, NY, USA: Association for Computing Machinery. ISBN 9781450362016.

White, H. 1989a. Learning in artificial neural networks: A statistical perspective. *Neural computation*, 1(4): 425–464.

White, H. 1989b. Some asymptotic results for learning in single hidden-layer feedforward network models. *Journal of the American Statistical Association*, 84(408): 1003–1013.

Yatchew, A. J. 1992. Nonparametric regression tests based on least squares. *Econometric Theory*, 8(4): 435–451.

Zeiler, M. D.; and Fergus, R. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*, 818–833. Springer.