Full Bayesian Significance Testing via Neural Networks: Appendix

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

Code

https://github.com/liuzh-buaa/nfbst.

Theoretical Derivation of Variational Inference for Bayesian Neural Networks

In this section, we provide a detailed theoretical derivation for training Bayesian neural networks through variational inference. Further, we provide the specific form of the objective function in the context of the variational family with diagonal Gaussian distributions.

Given a dataset $\mathcal{D} = \{(X^{(1)}, y^{(1)}), \ldots, (X^{(n)}, y^{(n)})\}$ of n i.i.d. pairs. For a Bayesian neural network, whose parameters are θ, we first assign a prior distribution as an initial belief $\pi(\theta)$ according to experience. This belief is gradually adjusted to fit data $\mathcal{D}$ by using the Bayesian rule, and the final belief is presented as the posterior distribution $P(\theta|\mathcal{D})$.

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

However, it is intractable to solve the integral in Eq (1). Through variational inference, we use a tractable distribution to approximate the real but intractable posterior distribution. 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

$$\vartheta^* = \arg\min_{\vartheta\in\Gamma} \mathrm{KL}(q_\vartheta(\theta)\|P(\theta|\mathcal{D})). \tag{2}$$

KL divergence describes the "distance" between two distributions. According to its definition, the optimization objective function can be formulated as

$$\arg\min \mathcal{L} = \mathrm{KL}(q_\vartheta(\theta)\|P(\theta|\mathcal{D})) = \int q_\vartheta(\theta) \log \frac{q_\vartheta(\theta)}{P(\theta|\mathcal{D})}\mathrm{d}\theta. \tag{3}$$

From the Bayesian formula Eq (1), Eq (3) can be simplified to

$$\begin{aligned}
\mathcal{L} &= \int q_\vartheta(\theta) \log \left(q_\vartheta(\theta)\frac{1}{P(\theta|\mathcal{D})}\right)\mathrm{d}\theta \\
&= \int q_\vartheta(\theta) \log \left(q_\vartheta(\theta)\frac{P(\mathcal{D})}{P(\mathcal{D}|\theta)\pi(\theta)}\right)\mathrm{d}\theta \\
&= -\int q_\vartheta(\theta) \log P(\mathcal{D}|\theta)\mathrm{d}\theta + \int q_\vartheta(\theta)\log\frac{q_\vartheta(\theta)}{\pi(\theta)}\mathrm{d}\theta \\
&\quad + \int q_\vartheta(\theta)\log p(\mathcal{D})\mathrm{d}\theta \\
&= -\mathbb{E}_{\theta\sim q_\vartheta(\theta)} \log P(\mathcal{D}|\theta) + \mathrm{KL}(q_\vartheta(\theta)\|\pi(\theta)) + \log P(\mathcal{D})\mathrm{d}\theta.
\end{aligned} \tag{4}$$

For the first term, if we assume

$$P(y|X, \theta) = N(\hat{y}(X, \theta), \sigma_0^2), \tag{5}$$

we can get

$$\begin{aligned}
\log P(\mathcal{D}|\theta) &= \log \prod_{i=1}^{n} P(y^{(i)}|X^{(i)}, \theta) \\
&= \sum_{i=1}^{n} \log P(y^{(i)}|X^{(i)}, \theta) \\
&= \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi}\sigma_0}\exp\left(-\frac{(y^{(i)} - \hat{y}(X^{(i)}, \theta))^2}{2\sigma_0^2}\right) \\
&= -\sum \frac{1}{2\sigma_0^2}(y^{(i)} - \hat{y}(X^{(i)}, \theta))^2 - n\log\sqrt{2\pi}\sigma_0.
\end{aligned} \tag{6}$$

This is equivalent to Mean Squared Error (MSE) in the regression task within a scaling factor. Then, using Monte Carlo, $\mathbb{E}_{\theta\sim q_\vartheta(\theta)} \log P(\mathcal{D}|\theta)$ means the average of MSE under $\theta \sim q_\vartheta(\theta)$.

For the second term, we adopt popular diagonal Gaussian distributions as the variational family Q in our experiment. We first consider θ follows a one-dimensional distribution as follows:

$$q_\vartheta(\theta) \sim \mathcal{N}(\mu, \sigma^2). \tag{7}$$

Correspondingly, we assume

$$\pi(\theta) \sim \mathcal{N}(\mu_\pi, \sigma_\pi^2). \tag{8}$$

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

Then we get

$$
\begin{aligned}
\mathrm{KL}\left(q_\vartheta(\theta)\|\pi(\theta)\right) &= \int q_\vartheta(\theta)\log\frac{q_\vartheta(\theta)}{\pi(\theta)}\mathrm{d}\theta \\
&= \int q_\vartheta(\theta)\log q_\vartheta(\theta)\mathrm{d}\theta - \int q_\vartheta(\theta)\log\pi(\theta)\mathrm{d}\theta \\
&= \int q_\vartheta(\theta)\log\Big(\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(\theta-\mu)^2}{2\sigma^2})\Big)\mathrm{d}\theta - \\
&\quad \int q_\vartheta(\theta)\log\Big(\frac{1}{\sqrt{2\pi}\sigma_\pi}\exp(-\frac{(\theta-\mu_\pi)^2}{2\sigma_\pi^2})\Big)\mathrm{d}\theta \\
&= -\frac{1}{2}\Big(1+\log(2\pi\sigma^2)\Big) - \\
&\quad \Big(-\frac{1}{2}\log(2\pi\sigma_\pi^2) - \frac{\sigma^2+(\mu-\mu_\pi)^2}{2\sigma_\pi^2}\Big) \\
&= \log\frac{\sigma_\pi}{\sigma} + \frac{\sigma^2+(\mu-\mu_\pi)^2}{2\sigma_\pi^2} - \frac{1}{2}.
\end{aligned}
\tag{9}
$$

This can also be extended to n-dimensional diagonal Gaussian distributions. That is,

$$
\pi(\theta) \sim \mathcal{N}(\mu_\pi, \Sigma_\pi), q_\vartheta(\theta) \sim \mathcal{N}(\mu, \Sigma) \tag{10}
$$

and

$$
\mu_\pi = \begin{bmatrix} \mu_{\pi 1} \\ \mu_{\pi 2} \\ \dots \\ \mu_{\pi n} \end{bmatrix}, \Sigma_\pi = \begin{bmatrix} \sigma_{\pi 1}^2 & 0 & \dots & 0 \\ 0 & \sigma_{\pi 2}^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma_{\pi n}^2 \end{bmatrix}, \tag{11}
$$

$$
\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \dots \\ \mu_n \end{bmatrix}, \Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma_n^2 \end{bmatrix}. \tag{12}
$$

Further, $q_\vartheta(\theta)$ can be denotes as

$$
q_\vartheta(\theta) = \frac{1}{\sqrt{(2\pi)^n|\Sigma|}}\exp\{-\frac{1}{2}(\theta-\mu)^T\Sigma^{-1}(\theta-\mu)\}, \tag{13}
$$

and $\pi(\theta)$ can be denotes as

$$
\pi(\theta) = \frac{1}{\sqrt{(2\pi)^n|\Sigma_\pi|}}\exp\{-\frac{1}{2}(\theta-\mu_\pi)^T\Sigma_\pi^{-1}(\theta-\mu_\pi)\}. \tag{14}
$$

Then, we can get

$$
\begin{aligned}
\mathrm{KL}\left(q_\vartheta(\theta)\|\pi(\theta)\right) &= \int q_\vartheta(\theta)\log\frac{q_\vartheta(\theta)}{\pi(\theta)}\mathrm{d}\theta \\
&= \int q_\vartheta(\theta)\log q_\vartheta(\theta)\mathrm{d}\theta - \int q_\vartheta(\theta)\log\pi(\theta)\mathrm{d}\theta \\
&= \log\frac{1}{\sqrt{(2\pi)^n|\Sigma|}} - \log\frac{1}{\sqrt{(2\pi)^n|\Sigma_\pi|}} \\
&\quad -\frac{1}{2}\int q_\vartheta(\theta)(\theta-\mu)^T\Sigma^{-1}(\theta-\mu)\mathrm{d}\theta \\
&\quad +\frac{1}{2}\int q_\vartheta(\theta)(\theta-\mu_\pi)^T\Sigma_\pi^{-1}(\theta-\mu_\pi)\mathrm{d}\theta \\
&= \log\frac{1}{\sqrt{(2\pi)^n|\Sigma|}} - \log\frac{1}{\sqrt{(2\pi)^n|\Sigma_\pi|}} \\
&\quad -\frac{1}{2}\mathbb{E}_{q_\vartheta(\theta)}(\theta-\mu)^T\Sigma^{-1}(\theta-\mu) \\
&\quad +\frac{1}{2}\mathbb{E}_{q_\vartheta(\theta)}(\theta-\mu_\pi)^T\Sigma_\pi^{-1}(\theta-\mu_\pi)
\end{aligned}
\tag{15}
$$

Proposition 1. Given

$$
\mathbf{x} \sim \mathcal{N}(\mu, \Sigma) \tag{16}
$$

where $\mathbf{x}$ is a n dimensional vector, that is $\mathbf{x} \in \mathbb{R}^{n\times 1}$. If we assume A is a $n \times n$ matrix, we have

$$
\mathbb{E}(\mathbf{x}^T A\mathbf{x}) = \mathrm{tr}(A\Sigma) + \mu^T A\mu. \tag{17}
$$

Proof. The trace of a matrix (denoted as "tr") is defined as the sum of the principle diagonal elements of a matrix, which has the following properties:

$$
\begin{aligned}
&\mathrm{tr}(M_1M_2) = \mathrm{tr}(M_2M_1) \\
&\Rightarrow \mathrm{tr}(M_1M_2M_3) = \mathrm{tr}(M_2M_3M_1) = \mathrm{tr}(M_3M_1M_2),
\end{aligned}
\tag{18}
$$

where M_1, M_2, M_3 are three compatible matrices. Therefore, we have

$$
\mathbf{x}^T A\mathbf{x} = \mathrm{tr}(\mathbf{x}^T A\mathbf{x}) = \mathrm{tr}(A\mathbf{x}\mathbf{x}^T). \tag{19}
$$

According to the properties of n-dimensional diagonal Gaussian distribution, we have

$$
\begin{aligned}
\Sigma &= \mathbb{E}\Big((\mathbf{x}-\mu)(\mathbf{x}-\mu)^T\Big) \\
&= \mathbb{E}\Big((\mathbf{x}-\mu)(\mathbf{x}^T-\mu^T)\Big) \\
&= \mathbb{E}\Big(\mathbf{x}\mathbf{x}^T - \mathbf{x}\mu^T - \mu\mathbf{x}^T + \mu\mu^T\Big) \\
&= \mathbb{E}(\mathbf{x}\mathbf{x}^T) - \mathbb{E}(\mathbf{x})\mu^T - \mu\mathbb{E}(\mathbf{x}^T) + \mu\mu^T \\
&= \mathbb{E}(\mathbf{x}\mathbf{x}^T) - \mu\mu^T.
\end{aligned}
\tag{20}
$$

Based on Eq (19),

$$
\begin{aligned}
\mathbb{E}(\mathbf{x}^T A\mathbf{x}) &= \mathbb{E}\Big(\mathrm{tr}(A\mathbf{x}\mathbf{x}^T)\Big) \\
&= \mathrm{tr}\Big(\mathbb{E}(A\mathbf{x}\mathbf{x}^T)\Big) \\
&= \mathrm{tr}\Big(A\mathbb{E}(\mathbf{x}\mathbf{x}^T)\Big) \\
&= \mathrm{tr}\Big(A(\Sigma+\mu\mu^T)\Big).
\end{aligned}
\tag{21}
$$

Then, based on Eq (20),

$$
\mathbb{E}(\mathbf{x}^T A\mathbf{x}) = \mathrm{tr}(A\Sigma) + \mathrm{tr}(A\mu\mu^T) = \mathrm{tr}(A\Sigma) + \mu^T A\mu. \tag{22}
$$

According to Proposition 1, we can simplify Eq (15) as

$$
\begin{aligned}
\mathrm{KL}\left(q_\vartheta(\theta)\|\pi(\theta)\right) &= \log\frac{1}{\sqrt{(2\pi)^n|\Sigma|}} - \log\frac{1}{\sqrt{(2\pi)^n|\Sigma_\pi|}} \\
&\quad -\frac{1}{2}\mathrm{tr}(\Sigma^{-1}\Sigma) - (\mu-\mu)^T\Sigma^{-1}(\mu-\mu) \\
&\quad +\frac{1}{2}\mathrm{tr}(\Sigma_\pi^{-1}\Sigma) - (\mu-\mu_\pi)^T\Sigma_\pi^{-1}(\mu-\mu_\pi) \\
&= \frac{1}{2}\log\frac{|\Sigma_\pi|}{|\Sigma|} + \frac{1}{2}\mathrm{tr}\left(\Sigma_\pi^{-1}\Sigma\right) \\
&\quad +\frac{1}{2}(\mu-\mu_\pi)^T\Sigma_\pi^{-1}(\mu-\mu_\pi) - \frac{1}{2}n \\
&= \sum_{i=1}^{n}\Big(\log\frac{\sigma_{\pi i}}{\sigma_i} + \frac{\sigma_i^2+(\mu_i-\mu_{\pi i})^2}{2\sigma_{\pi i}^2} - \frac{1}{2}\Big)
\end{aligned}
\tag{23}
$$

The result is consistent with Eq (9).

In conclusion, we have analyzed the meanings of different terms in the objective function Eq (4). The first term Eq (6) is related to data (such as MSE for regression task). The second term Eq (23) is only related to $\vartheta = (\mu, \sigma)$ further θ like a regularization term. The third term $\log P(\mathcal{D})$ is a constant once the data $\mathcal{D}$ is determined.

Theoretical Proof of the Convergence of Bayesian Evidence

In this section, we provide a detailed proof for the convergence of Bayesian Evidence in linear regression context. It still holds true when extended to nonlinear scenarios and we will provide a rigorous proof in the future work.

Problem Definition

We denote $f_0 : \mathcal{X} \to \mathbb{R}$ as the underlying relationship between features and target. Then, we assume that the target variable y is given by

$$y = f_0(\mathbf{x}) + \epsilon, \tag{24}$$

where $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^d$ is a vector of feature variables, $y \in \mathbb{R}$ is the dependent variable and ϵ is the noise variable.

In linear regression context, we assume

$$f_0(\mathbf{x}) = \mathbf{x}\beta = \mathbf{x}\begin{bmatrix}\beta_1\\ \beta_2\\ \cdots\\ \beta_d\end{bmatrix}, \tag{25}$$

and ϵ is a zero mean Gaussian random variable with variance (inverse precision) σ_0^2

$$p(\epsilon) \sim \mathcal{N}(\epsilon|0, \sigma_0^2). \tag{26}$$

Thus we can write

$$p(y|\mathbf{x}, \beta) = \mathcal{N}(y|\mathbf{x}\beta, \sigma_0^2). \tag{27}$$

Note that β_i is the coefficient corresponding to the $i-$th feature. Now consider a dataset of inputs $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ with corresponding target values $\{y_1, \ldots, y_n\}$. To simplify, we group them into the form of matrix, that is $\mathbf{X} \in \mathbb{R}^{n\times d}, \mathbf{y} \in \mathbb{R}^{n\times 1}$. By employing the technique of "block matrices" (or "partitioned matrices"), we denote $\mathbf{X}$ as

$$\mathbf{X} = [\mathbf{X}_1 \quad \mathbf{X}_2 \quad \cdots \quad \mathbf{X}_d], \tag{28}$$

where $\mathbf{X}_i$ is the $i-$th column of $\mathbf{X}$, representing the values of $i-$th feature of all data. Making the assumption that these data points are drawn independently from the distribution Eq(27), we obtain the following expression for the likelihood function, which is a function of the adjustable parameters β and σ_0^2, in the form

$$p(\mathbf{y}|\mathbf{X}, \beta) = \prod_{i=1}^{n} \mathcal{N}(y_i|\mathbf{x}_i\beta, \sigma_0^2) = \mathcal{N}(\mathbf{y}|\mathbf{X}\beta, \sigma_0^2\mathbf{I}). \tag{29}$$

To test the significance of a feature $\mathbf{X}_j$, the problem is formulated as:

$$H_0 : \beta_j = 0, \quad H_1 : \beta_j \neq 0. \tag{30}$$

Prior Distribution over Model Parameters β

We begin our discussion of the Bayesian treatment of linear regression by introducing a prior probability distribution over the model parameters β. First note that the likelihood function $p(\mathbf{x}|\beta)$ defined by Eq(29) is the exponential of a quadratic function of β. The corresponding conjugate prior is therefore given by a Gaussian distribution of the form

$$\pi(\beta) = \mathcal{N}(\beta|\mu_p, \Sigma_p) \tag{31}$$

having mean μ_p and covariance Σ_p. In our approach, we consider a zero-mean diagonal Gaussian so that

$$\mu_p = \mathbf{0}, \Sigma_p = \begin{bmatrix}\alpha_0^2 & 0 & \cdots & 0\\ 0 & \alpha_1^2 & \cdots & 0\\ \cdots & \cdots & \cdots & \cdots\\ 0 & 0 & \cdots & \alpha_d^2\end{bmatrix}. \tag{32}$$

Posterior Distribution over Model Parameters β

Next we compute the posterior distribution. From the Bayes' theorem, we have

$$p(\beta|\mathbf{X}, \mathbf{y}) \propto p(\mathbf{y}|\beta, \mathbf{X})\pi(\beta). \tag{33}$$

The posterior distribution is proportional to the product of the likelihood function and the prior. Due to the choice of a conjugate Gaussian prior distribution, the posterior will also be Gaussian. The derivation is as follows:

$$\begin{aligned}
p(\beta|\mathbf{X}, \mathbf{y}) &\propto p(\mathbf{y}|\mathbf{X}, \beta)\pi(\beta)\\
&= \mathcal{N}(\mathbf{y}|\mathbf{X}\beta, \sigma_0^2\mathbf{I})\mathcal{N}(\beta|\mathbf{0}, \Sigma_p)\\
&= \frac{1}{\sqrt{(2\pi)^n}\sigma_0^n}\exp\{-\frac{1}{2}(\mathbf{y}-\mathbf{X}\beta)^\top\sigma_0^{-2}\mathbf{I}(\mathbf{y}-\mathbf{X}\beta)\}\\
&\quad \frac{1}{\sqrt{(2\pi)^d|\Sigma_p|}}\exp\{-\frac{1}{2}\beta^\top\Sigma_p^{-1}\beta\}\\
&= c\exp\{-\frac{1}{2\sigma_0^2}(\mathbf{y}^\top - \beta^\top\mathbf{X}^\top)(\mathbf{y}-\mathbf{X}\beta) - \frac{1}{2}\beta^\top\Sigma_p^{-1}\beta\}\\
&= c\exp\{-\frac{1}{2}\beta^\top(\sigma_0^{-2}\mathbf{X}^\top\mathbf{X} + \Sigma_p^{-1})\beta - \frac{1}{2\sigma_0^2}\mathbf{y}^\top\mathbf{y}\\
&\qquad + \frac{1}{2\sigma_0^2}(\mathbf{y}^\top\mathbf{X}\beta + \beta^\top\mathbf{X}^\top\mathbf{y})\},
\end{aligned} \tag{34}$$

where $c = (\sqrt{(2\pi)^{n+d}|\Sigma_p|}\sigma_0^n)^{-1}$ is a constant. The aim is to simplify Eq(34) in the form of Gaussian distribution. In the end, we obtain

$$p(\beta|\mathbf{X}, \mathbf{y}) = \mathcal{N}(\beta|\mu, \Sigma), \tag{35}$$

where

$$\mu = \sigma_0^{-2}\Sigma\mathbf{X}^\top\mathbf{y}, \tag{36}$$

$$\Sigma = (\sigma_0^{-2}\mathbf{X}^\top\mathbf{X} + \Sigma_p^{-1})^{-1}. \tag{37}$$

Property of μ, Σ of β (Optional)

Due to

$$(\mathbf{X}^\top\mathbf{X})^\top = \mathbf{X}^\top\mathbf{X}, \tag{38}$$

$\mathbf{X}^\top\mathbf{X}$ is a symmetric matrix. By using the technique of "Eigenvalue Decomposition", we have

$$\mathbf{X}^\top\mathbf{X} = \mathbf{Q}\mathbf{D}\mathbf{Q}^{-1} = \mathbf{Q}\mathbf{D}\mathbf{Q}^\top, \tag{39}$$

and

$$\mathbf{Q} = [\mathbf{v}_1 \quad \mathbf{v}_2 \quad \cdots \quad \mathbf{v}_d], \mathbf{D} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \lambda_d \end{bmatrix}, \tag{40}$$

where $\lambda_1, \lambda_2, \cdots, \lambda_d$ are eigenvalues and $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_d$ are their corresponding eigenvectors satisfying

$$\lambda_i \mathbf{v}_i = \mathbf{X}^\top \mathbf{X} \mathbf{v}_i \quad i = 1, 2, \cdots, d. \tag{41}$$

If we assume a zero-mean isotropic Gaussian governed by a single parameter σ_p so that

$$\Sigma_p = \sigma_p^2 \mathbf{I}. \tag{42}$$

Then, we obtain

$$\begin{aligned} \Sigma &= (\sigma_0^{-2}\mathbf{Q}\mathbf{D}\mathbf{Q}^{-1} + \sigma_p^{-2}\mathbf{I})^{-1} \\ &= (\sigma_0^{-2}\mathbf{Q}\mathbf{D}\mathbf{Q}^{-1} + \sigma_p^{-2}\mathbf{Q}\mathbf{I}\mathbf{Q}^{-1})^{-1} \\ &= \left(\mathbf{Q}(\sigma_0^{-2}\mathbf{D} + \sigma_p^{-2}\mathbf{I})\mathbf{Q}^{-1}\right)^{-1} \\ &= \mathbf{Q}(\sigma_0^{-2}\mathbf{D} + \sigma_p^{-2}\mathbf{I})^{-1}\mathbf{Q}^{-1} \\ &= \mathbf{Q}(\sigma_0^{-2}\mathbf{D} + \sigma_p^{-2}\mathbf{I})^{-1}\mathbf{Q}^\top. \end{aligned} \tag{43}$$

We notate

$$\mathbf{M} = (\sigma_0^{-2}\mathbf{D} + \sigma_p^{-2}\mathbf{I})^{-1} = \begin{bmatrix} m_1 & 0 & \cdots & 0 \\ 0 & m_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & m_d \end{bmatrix} \tag{44}$$

where $m_i = 1/(\sigma_0^{-2}\lambda_i + \sigma_p^{-2})$, then we obtain

$$\Sigma = \mathbf{Q}\mathbf{M}\mathbf{Q}^\top = \sum_{i=1}^{d} m_i \mathbf{v}_i \mathbf{v}_i^\top. \tag{45}$$

$$\mu = \sigma_0^2 \Sigma \mathbf{X}^\top \mathbf{y} = \sigma_0^{-2} \sum_{i=1}^{d} m_i \mathbf{v}_i \mathbf{v}_i^\top \mathbf{X}^\top \mathbf{y}. \tag{46}$$

Posterior Distribution over Model Parameter β_j

From a geometric interpretation of the Gaussian distribution, we can also denote Σ as

$$\Sigma = \begin{bmatrix} \sigma(\beta_1, \beta_1) & \sigma(\beta_1, \beta_2) & \cdots & \sigma(\beta_1, \beta_d) \\ \sigma(\beta_2, \beta_1) & \sigma(\beta_2, \beta_2) & \cdots & \sigma(\beta_2, \beta_d) \\ \cdots & \cdots & \cdots & \cdots \\ \sigma(\beta_d, \beta_1) & \sigma(\beta_d, \beta_2) & \cdots & \sigma(\beta_d, \beta_d) \end{bmatrix}, \tag{47}$$

where $\sigma(\beta_i, \beta_j)$ is the covariance (or variance) of β_i and β_j. And we can also denote μ as

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \cdots \\ \mu_d \end{bmatrix}, \tag{48}$$

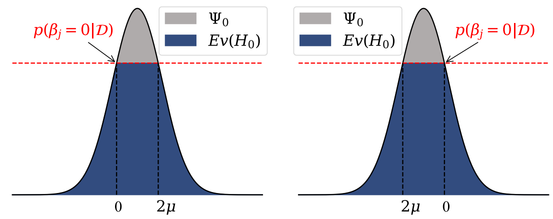
where μ_j is the mean of β_j. Therefore, the posterior distribution of β_j is

$$p(\beta_j | \mathbf{X}, \mathbf{y}) = \mathcal{N}(\beta_j | \mu_j, \sigma_j^2), \tag{49}$$

where

$$\mu_j = \sigma_0^{-2} \sigma_j^2 \mathbf{X}_j^\top \mathbf{y}, \tag{50}$$

$$\sigma_j^2 = \frac{1}{\sigma_0^{-2}\mathbf{X}_j^\top \mathbf{X}_j + \alpha_j^{-2}}. \tag{51}$$



(a) $\mu_j < 0$ (b) $\mu_j \geq 0$

Figure 1: Bayesian evidence calculated based on the distribution of β_j.

Convergence of Bayesian Evidence

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

$$\Psi_0 = \{\beta_j : p(\beta_j | \mathbf{X}, \mathbf{y}) > p(\beta_j = 0 | \mathbf{X}, \mathbf{y}\}, \tag{52}$$

where $p(\beta_j = 0 | \mathbf{X}, \mathbf{y})$ should be the maximum of the posterior density under the null hypothesis H_0 (Eq(30)). In our method, we adopt a more flexible valid Bayesian evidence for the null hypothesis provided by (**?**):

$$Ev(H_0) = 1 - \int_{\Psi_0} p\Big(\beta_j | \mathbf{X}, \mathbf{y}\Big) \mathrm{d}\beta_j \tag{53}$$

The geometric interpretation of the electric field (E-field) is illustrated in Figure 1. According to Eq(49), the posterior distribution of β_j, $p(\beta_j | \mathbf{X}, \mathbf{y})$, is a Gaussian distribution. Due to symmetry,

$$p(\beta_j = 0 | \mathbf{X}, \mathbf{y}) = p(\beta_j = 2\mu_j | \mathbf{X}, \mathbf{y}). \tag{54}$$

Further, when the value of β_j is between 0 and $2\mu_j$, there is $p(\beta_j | \mathbf{X}, \mathbf{y}) > p(\beta_j = 0 | \mathbf{X}, \mathbf{y})$. In other words,

$$\Psi_0 = \begin{cases} [2\mu_j, 0] & \mu_j < 0 \\ [0, 2\mu_j] & \mu_j \geq 0 \end{cases}. \tag{55}$$

Eq(53) can be simplified

$$Ev(H_0) = \begin{cases} 1 - \int_{2\mu_j}^{0} p(\beta_j | \mathbf{X}, \mathbf{y}) \mathbf{d}\beta_j & \mu_j < 0 \\ 1 - \int_{0}^{2\mu_j} p(\beta_j | \mathbf{X}, \mathbf{y}) \mathbf{d}\beta_j & \mu_j \geq 0 \end{cases}. \tag{56}$$

If we assume the feature $\mathbf{X}_j$ is insignificant, that is $\delta \mathbf{y} / \delta \mathbf{X}_j = 0$. This implies that $\mathbf{y}$ does not change with the variation of $\mathbf{X}_j$ in the direction of $\mathbf{X}_j$. From the perspective of geometry, $\mathbf{X}_j$ and $\mathbf{y}$ are $n-$dimensional vectors. It means that $\mathbf{y}$ is orthogonal to $\mathbf{X}_j$, hence

$$\lim_{n \to \infty} \mathbf{X}_j^\top \mathbf{y} = 0. \tag{57}$$

Moreover, when n goes to ∞, $\mathbf{X}^\top \mathbf{X} \to \infty$, hence

$$\lim_{n \to \infty} \sigma_j^2 = \lim_{n \to \infty} \frac{1}{\sigma_0^{-2}\mathbf{X}_j^\top \mathbf{X}_j + \alpha_j^{-2}} = 0. \tag{58}$$

According to Eq(50),

$$\lim_{n \to \infty} \mu_j = \sigma_0^{-2} (\lim_{n \to \infty} \sigma_j^2)(\lim_{n \to \infty} \mathbf{X}_j^\top \mathbf{y}) = 0, \tag{59}$$

and μ_j converges faster than σ_j^2. Therefore,

$$\lim_{n \to \infty} Ev(H_0) = 1. \tag{60}$$

Experimental Setup

In this section, we introduce the detailed experimental setup, including hardware settings and algorithm parameters. All the experiments are conducted on an Ubuntu machine equipped with 40 Intel(R) Xeon(R) Silver 4210 CPUs @ 2.20GHz with 10 physical cores, and the GPU is NVIDIA GeForce RTX 2080 Ti, armed with 11GB of GDDR6 memory.

For toy example, Dataset 1, Dataset 2, and energy efficiency datasets, we fit fully-connected feed-forward neural networks and Bayesian neural networks using the PyTorch package (v1.8.0). Both neural networks and Bayesian neural networks have the same structure. Specifically, they have three hidden layers with twenty nodes and a relu activation function, the same structure as f_0 in Dataset 1 but different from others. This also shows our approach does not enforce the same structure as f_0. For MNIST dataset, we fit a convolutional neural network and a Bayesian neural network with the same structure to perform digit classification. That is, it first repeats the convolution layer and the max pooling layer twice, followed by two fully connected layers, with ReLU as the activation function. Table 2 shows the mean squared error (MSE) for regression tasks or test accuracy for classification tasks and all models fit the data well.

For neural networks and Bayesian neural networks, they are trained by the Adam optimizer. Moreover, we use "torch.optim.lr_scheduler.StepLR" to decay the learning rate of each parameter group. We set the maximum epochs of training to 20000 and adopt the early stopping strategy with patience equals 40. That is, we stop training when the monitor metric (MSE for regression or accuracy for classification) does not decrease (or increase) for 40 epochs. For neural networks, we set an L2 regularization term in the loss function to avoid overfitting, which is implemented by setting weight decay in the optimizer. Note for Bayesian neural networks, we don't need to set it. The regularization term is calculated exactly by Eq (23). The descriptions of related parameters are listed as follows:

- optimizer::lr. The learning rate of Adam optimizer.
- optimizer::step_size. The period of learning rate decay.
- optimizer::decay_gamma. The multiplicative factor of learning rate decay.
- optimizer::weight_decay. The weight decay of Adam optimizer, acting as L2 penalty.
- Bayesian neural network::σ_0. The variance of the likelihood distribution of y in Eq (6).
- Bayesian neural network::μ_π. The mean of the prior distribution of θ in Eq (8).
- Bayesian neural network::σ_π. The variance of the prior distribution of θ in Eq (8).

We use the grid search algorithm to choose the optimal ones as parameters. Their specific values are shown in Table 1. In practice, we adopt diagonal Gaussian distribution as our variational family of model parameters and ensemble three Bayesian neural networks to spread the range of model parameters further.

Besides, for our method *n*FBST, we need to sample through the Bayesian neural network and then approximate the posterior probability density by Kernel Density Estimation (KDE). The sample size we set is one hundred. We use the Gaussian kernel function and choose the best bandwidth from $\{0.01, 0.05, 0.1, 1.0\}$ by GridSearchCV for each sampling through 5-fold cross-validation. The bandwidth here acts as a smoothing parameter, controlling the trade-off between bias and variance in the result. A large bandwidth leads to a very smooth (that is, high-bias) density distribution while a small bandwidth leads to an unsmooth (that is, high-variance) density distribution.

Descriptions of Energy Efficiency Dataset

In the paper, we conduct experiments on the energy efficient dataset, and the descriptions of its features and target are shown in Table 3. When it comes to efficient building design, the computation of the heating load (HL) is required to determine the specifications of the heating equipment needed to maintain comfortable indoor air conditions. Those eight variables have been frequently used in the EPB (Energy Performance of Buildings) literature to study energy-related topics in buildings.

Limitations of Our Approach

In our implementation, we adopt the Kernel Density Estimation (KDE) algorithm as the primary method to estimate the posterior probability density. Unfortunately, the lack of an effective GPU acceleration method restricts the calculation process to the CPU. Consequently, the overall pipeline cannot be completed entirely on the GPU, leading to an increase in the overall computing time. To address this issue and enhance the computational efficiency of our model, we aim to develop a GPU-accelerated version of KDE or explore other feasible alternative methods in future research.

Toy Example to Compare Distribution-based with Point-based Approaches

In the paper, we use a toy example to verify that *n*FBST identifies the global significance more accurately than other testing methods and it can provide more insights on the local or instance-wise significance. Here, we would like to illustrate further why we need significance testing rather than a point estimate by comparing feature importance analysis.

We consider the following data generation process

$$y = 8+x_0^2+x_1x_2+\cos(x_3)+\exp(x_4x_5)+0.1x_6+0x_7+\epsilon, \quad (61)$$

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. We first trained a neural network f and a Bayesian neural network with the same structure, and both of them fit the data well (MSE is about 1.0, which equals the variance of ϵ). If we adopt partial derivative as the testing statistic, for a specific data instance, we get

$$\frac{\partial f}{\partial x_6} = 0.034 < 0.197 = \frac{\partial f}{\partial x_7}. \quad (62)$$

Parameters	Toy Dataset	Dataset 1	Dataset 2	Energy Effieciency	MNIST
optimizer::lr	0.01,0.01	0.01,0.01	0.01,0.01	0.01,0.01	0.001,0.001
optimizer::step_size	50,30	100,100	100,100	100,100	inf,inf
optimizer::decay_gamma	0.5,0.5	0.5,0.5	0.5,0.5	0.1,0.1	\,\
optimizer::weight_decay	\,0.01	\,0	\,0	\,0.001	\,0
Bayesian neural network::σ_0	0.1,\	0.01,\	0.01,\	0.0001,\	0.001,\
Bayesian neural network::μ_π	0,\	0,\	0,\	0,\	0,\
Bayesian neural network::σ_π	0.1,\	0.1,\	0.1,\	0.001,\	0.01,\

Table 1: The optimal parameters for Bayesian neural network(BNN) and neural network(NN) on different datasets. The former of the tuple is for BNN, and the latter is for NN.

Models	Dataset 1	Dataset 2	Dataset 3	Energy Efficiency	MNIST
Bayesian Neural Network	0.012 (MSE)	0.014 (MSE)	0.010 (MSE)	0.144 (MSE)	99.2% (acc)
Neural Network	0.012 (MSE)	0.014 (MSE)	0.009 (MSE)	0.149 (MSE)	99.3% (acc)

Table 2: Mean Squared Error (MSE) or accuracy (acc) on different datasets

Table 3: Descriptions of variables for residential energy efficiency evaluation.

Variable	Description	Number of possible values
x_1	Relative Compactness	12
x_2	Surface Area	12
x_3	Wall Area	7
x_4	Roof Area	4
x_5	Overall Height	2
x_6	Orientation	4
x_7	Glazing Area	4
x_8	Glazing Area Distribution	6
y	Heating Load	586

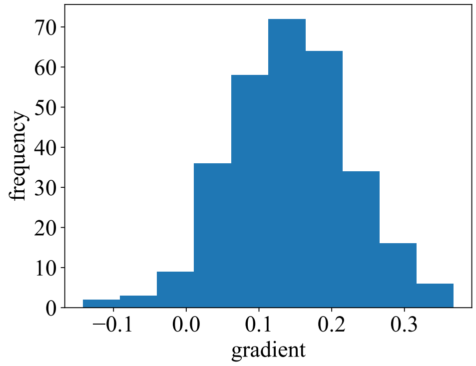
However, the feature x_6 is more significant than the insignificant feature x_7 actually. Then, we plot the gradient distribution obtained through the Bayesian neural network in the form of histograms (Figure 2). The Bayesian evidence obtained by Grad-*n*FBST is

$$Ev(H_0(x_6)) = 0.087 < 0.537 = Ev(H_0(x_7)). \quad (63)$$

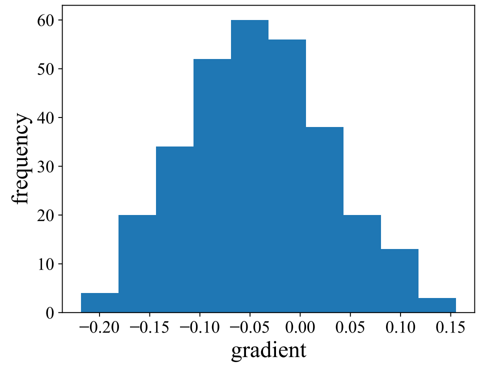
Note Bayesian evidence means the evidence in favor of H_0. In other words, the closer the Bayesian evidence to one, the more likely to accept H_0 that a feature is insignificant. It is consistent with the data generation process in Eq (61). In figure 2, although the peak where the probability density is greatest for both x_6 and x_7 is not exactly equal to 0, we can intuitively see the difference between them. For x_7, the second peak of the distribution is close to zero, the probability density of which is quite close to the probability density at the greatest peak. However, the probability density of zero for x_6 shrinks sharply near zero. Therefore, the final testing result of x_7 by Grad-*n*FBST is low, which means we are more likely to determine it as insignificant. However, the probability at zero for x_6 is obviously lower than the first peak, so we prefer to reject the proposition that x_6 is insignificant. In fact, this is probably not an uncommon situation. Although the performance of the neural network is strong, it still cannot be accurate with all data. That's one of the reasons why we need significance testing rather than a point-estimate method.

More Results of Simulation Experiments

In the paper, we show the comparison of AUC for each feature before and after *n*FBST under different eps. The average AUC of *n*FBST is higher than primary feature importance methods. Here we present more details in Table 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15 and Figure 3, 4, 5, 6, 7, 8 under different eps on two simulation datasets. The results show that *n*FBST outperforms feature importance methods on almost all features under different conditions.

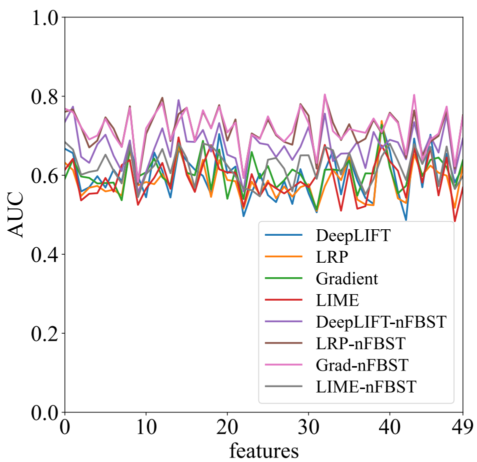


(a) Gradient distribution for x_6

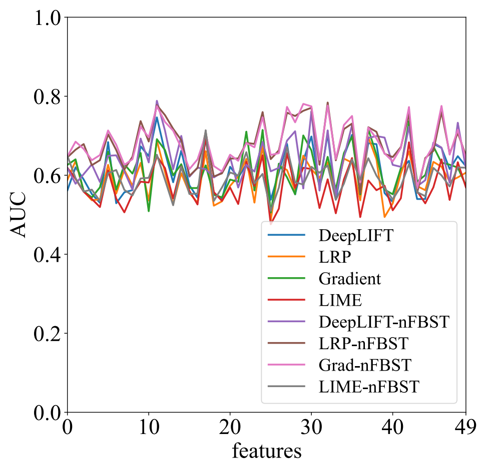


(b) Gradient distribution for x_7

Figure 2: Gradient distributions obtained from the *n*FBST.

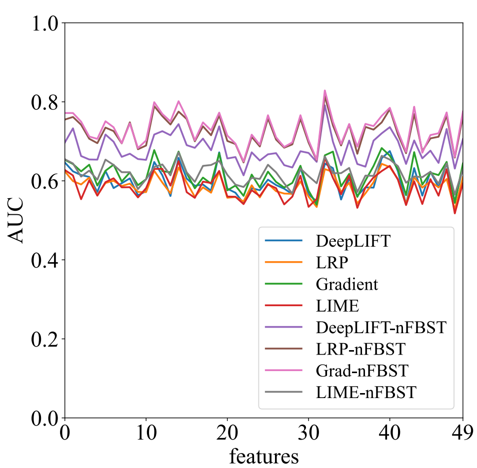


(a) Dataset 1

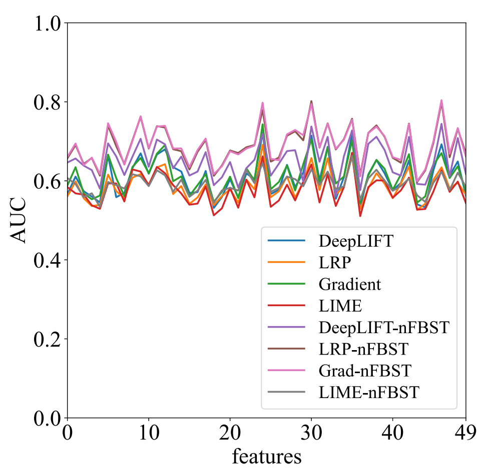


(b) Dataset 2

Figure 3: Comparison of AUC for each feature before and after *n*FBST under eps=0.001.



(a) Dataset 1



(b) Dataset 2

Figure 4: Comparison of AUC for each feature before and after *n*FBST under eps=0.01.

Feature	Gradient	Grad-*n*FBST	DeepLIFT	DeepLIFT-*n*FBST	LRP	LRP-*n*FBST	LIME	LIME-*n*FBST
x_0	0.678	**0.799**	0.661	**0.737**	0.636	**0.781**	0.648	**0.678**
x_1	0.668	**0.803**	0.639	**0.751**	0.629	**0.79**	0.646	**0.69**
x_2	0.653	**0.777**	0.649	**0.689**	0.616	**0.767**	0.6	**0.633**
x_3	0.643	**0.757**	0.627	**0.692**	0.618	**0.747**	0.62	**0.66**
x_4	0.607	**0.733**	0.581	**0.674**	0.58	**0.724**	0.576	**0.619**
x_5	0.658	**0.784**	0.65	**0.738**	0.617	**0.775**	0.62	**0.671**
x_6	0.663	**0.762**	0.609	**0.719**	0.631	**0.749**	0.626	**0.659**
x_7	0.609	**0.731**	0.598	**0.687**	0.592	**0.727**	0.593	**0.649**
x_8	0.606	**0.766**	0.61	**0.689**	0.578	**0.76**	0.602	**0.63**
x_9	0.595	**0.711**	0.567	**0.669**	0.572	**0.702**	0.564	**0.594**
x_{10}	0.618	**0.746**	0.599	**0.695**	0.586	**0.738**	0.595	**0.622**
x_{11}	0.71	**0.812**	0.679	**0.701**	0.658	**0.797**	0.653	**0.659**
x_{12}	0.686	**0.792**	0.651	**0.728**	0.645	**0.78**	0.667	**0.684**
x_{13}	0.635	**0.781**	0.595	**0.734**	0.589	**0.769**	0.596	**0.636**
x_{14}	0.7	**0.832**	0.661	**0.758**	0.652	**0.813**	0.688	**0.701**
x_{15}	0.662	**0.808**	0.651	**0.718**	0.619	**0.796**	0.596	**0.662**
x_{16}	0.604	**0.72**	0.593	**0.685**	0.58	**0.718**	0.578	**0.622**
x_{17}	0.641	**0.77**	0.592	**0.725**	0.605	**0.76**	0.609	**0.65**
x_{18}	0.619	**0.762**	0.579	**0.705**	0.59	**0.751**	0.606	**0.652**
x_{19}	0.725	**0.815**	0.715	**0.762**	0.678	**0.804**	0.665	**0.693**
x_{20}	0.61	**0.746**	0.595	**0.679**	0.579	**0.737**	0.568	**0.619**
x_{21}	0.628	**0.729**	0.597	**0.675**	0.599	**0.723**	0.569	**0.623**
x_{22}	0.559	**0.681**	0.528	**0.644**	0.541	**0.678**	0.539	**0.584**
x_{23}	0.639	**0.756**	0.598	**0.692**	0.602	**0.743**	0.594	**0.624**
x_{24}	0.632	**0.739**	0.6	**0.696**	0.6	**0.733**	0.607	**0.641**
x_{25}	0.654	**0.79**	0.614	**0.691**	0.613	**0.777**	0.612	**0.665**
x_{26}	0.626	**0.738**	0.607	**0.704**	0.598	**0.729**	0.577	**0.636**
x_{27}	0.597	**0.734**	0.588	**0.678**	0.573	**0.729**	0.556	**0.615**
x_{28}	0.606	**0.733**	0.581	**0.677**	0.578	**0.723**	0.567	**0.604**
x_{29}	0.672	**0.788**	0.657	**0.681**	0.628	**0.776**	0.626	**0.642**
x_{30}	0.614	**0.746**	0.592	**0.694**	0.586	**0.739**	0.563	**0.633**
x_{31}	0.562	**0.695**	0.552	**0.674**	0.552	**0.692**	0.563	**0.61**
x_{32}	0.693	**0.839**	0.653	**0.779**	0.642	**0.822**	0.665	**0.67**
x_{33}	0.705	**0.799**	0.653	**0.719**	0.656	**0.788**	0.655	**0.674**
x_{34}	0.606	**0.741**	0.583	**0.683**	0.587	**0.737**	0.589	**0.636**
x_{35}	0.625	**0.755**	0.614	**0.699**	0.602	**0.753**	0.61	**0.637**
x_{36}	0.575	**0.699**	0.574	**0.663**	0.558	**0.691**	0.563	**0.597**
x_{37}	0.625	**0.781**	0.606	**0.648**	0.586	**0.771**	0.601	**0.623**
x_{38}	0.666	**0.782**	0.617	**0.737**	0.625	**0.773**	0.637	**0.656**
x_{39}	0.674	**0.791**	0.64	**0.723**	0.638	**0.774**	0.624	**0.675**
x_{40}	0.7	**0.818**	0.692	**0.758**	0.652	**0.807**	0.648	**0.672**
x_{41}	0.673	**0.775**	0.655	**0.742**	0.634	**0.766**	0.625	**0.67**
x_{42}	0.578	**0.704**	0.563	**0.67**	0.568	**0.705**	0.564	**0.624**
x_{43}	0.679	**0.807**	0.652	**0.705**	0.615	**0.787**	0.607	**0.629**
x_{44}	0.593	**0.714**	0.562	**0.689**	0.577	**0.714**	0.563	**0.628**
x_{45}	0.658	**0.767**	0.645	**0.714**	0.616	**0.759**	0.624	**0.655**
x_{46}	0.649	**0.774**	0.606	**0.693**	0.604	**0.767**	0.573	**0.628**
x_{47}	0.695	**0.82**	0.684	**0.764**	0.646	**0.813**	0.652	**0.691**
x_{48}	0.569	**0.691**	0.55	**0.649**	0.545	**0.685**	0.534	**0.573**
x_{49}	0.664	**0.797**	0.641	**0.721**	0.621	**0.787**	0.621	**0.649**

Table 4: Comparison of AUC for each feature before and after *n*FBST under eps=0.001 on Dataset 1.

Feature	Gradient	Grad-*n*FBST	DeepLIFT	DeepLIFT-*n*FBST	LRP	LRP-*n*FBST	LIME	LIME-*n*FBST
x_0	0.626	**0.648**	0.562	**0.618**	0.588	**0.645**	0.619	**0.649**
x_1	0.64	**0.685**	**0.62**	0.578	0.636	**0.665**	0.591	**0.593**
x_2	0.56	**0.667**	0.589	**0.603**	0.559	**0.679**	0.558	**0.559**
x_3	0.544	**0.638**	0.552	**0.627**	0.537	**0.625**	0.54	**0.564**
x_4	0.569	**0.651**	0.53	**0.583**	0.539	**0.638**	0.519	**0.534**
x_5	0.66	**0.713**	**0.684**	0.649	0.626	**0.704**	**0.612**	0.605
x_6	0.563	**0.677**	0.529	**0.65**	0.554	**0.664**	0.541	**0.613**
x_7	**0.628**	0.626	0.557	**0.61**	0.614	**0.615**	0.506	**0.563**
x_8	0.604	**0.641**	0.562	**0.564**	0.574	**0.647**	**0.552**	0.549
x_9	0.656	**0.723**	0.673	**0.693**	0.631	**0.737**	0.584	**0.592**
x_{10}	0.509	**0.696**	**0.648**	0.632	0.536	**0.685**	0.582	**0.594**
x_{11}	0.691	**0.775**	0.747	**0.788**	0.688	**0.777**	0.648	**0.652**
x_{12}	0.664	**0.735**	0.658	**0.713**	0.608	**0.752**	**0.618**	0.588
x_{13}	0.6	**0.713**	0.582	**0.615**	0.527	**0.718**	**0.541**	0.524
x_{14}	0.628	**0.67**	0.662	**0.7**	0.605	**0.689**	**0.622**	0.621
x_{15}	0.568	**0.615**	0.574	**0.596**	0.555	**0.598**	**0.561**	0.552
x_{16}	0.568	**0.64**	0.543	**0.619**	0.53	**0.619**	0.526	**0.554**
x_{17}	0.625	**0.702**	**0.695**	0.615	0.661	**0.71**	0.698	**0.714**
x_{18}	0.551	**0.623**	0.555	**0.594**	0.523	**0.597**	**0.557**	0.535
x_{19}	0.544	**0.61**	0.539	**0.607**	0.534	**0.605**	0.535	**0.569**
x_{20}	0.589	**0.652**	0.622	**0.646**	0.573	**0.642**	0.569	**0.607**
x_{21}	0.583	**0.636**	**0.594**	0.569	0.597	**0.642**	0.527	**0.6**
x_{22}	**0.71**	0.682	0.627	**0.627**	0.641	**0.682**	0.634	**0.658**
x_{23}	0.561	**0.671**	0.569	**0.609**	0.531	**0.679**	0.573	**0.579**
x_{24}	0.715	**0.747**	0.671	**0.685**	0.677	**0.76**	**0.65**	0.603
x_{25}	0.506	**0.651**	0.538	**0.61**	0.486	**0.641**	0.476	**0.51**
x_{26}	0.629	**0.656**	0.608	**0.619**	0.625	**0.663**	0.515	**0.569**
x_{27}	0.595	**0.773**	0.679	**0.688**	0.613	**0.758**	0.656	**0.671**
x_{28}	0.552	**0.734**	0.552	**0.711**	0.562	**0.75**	0.568	**0.582**
x_{29}	0.701	**0.781**	**0.637**	0.566	0.65	**0.763**	**0.619**	0.576
x_{30}	0.663	**0.774**	0.698	**0.761**	0.615	**0.771**	0.616	**0.621**
x_{31}	0.571	**0.639**	**0.575**	0.561	0.593	**0.626**	0.517	**0.606**
x_{32}	0.647	**0.776**	0.705	**0.714**	0.633	**0.784**	0.59	**0.625**
x_{33}	0.564	**0.641**	0.539	**0.548**	0.552	**0.638**	0.504	**0.538**
x_{34}	0.656	**0.727**	**0.648**	0.645	0.642	**0.717**	**0.59**	0.579
x_{35}	0.698	**0.75**	0.702	**0.727**	0.632	**0.731**	**0.644**	0.64
x_{36}	0.551	**0.582**	0.558	**0.567**	0.536	**0.555**	0.494	**0.589**
x_{37}	0.716	**0.721**	0.681	**0.694**	0.696	**0.722**	0.587	**0.643**
x_{38}	0.652	**0.693**	0.679	**0.699**	0.641	**0.71**	0.562	**0.599**
x_{39}	0.564	**0.655**	0.555	**0.695**	0.494	**0.654**	**0.573**	0.568
x_{40}	0.552	**0.635**	0.532	**0.626**	0.529	**0.646**	0.511	**0.539**
x_{41}	0.618	**0.669**	0.614	**0.621**	0.604	**0.671**	0.542	**0.571**
x_{42}	0.737	**0.773**	0.636	**0.723**	0.66	**0.756**	**0.683**	0.628
x_{43}	**0.737**	0.567	0.54	**0.572**	**0.573**	0.555	**0.56**	0.557
x_{44}	0.6	**0.64**	0.54	**0.587**	0.562	**0.644**	0.529	**0.547**
x_{45}	0.673	**0.675**	0.68	**0.683**	0.634	**0.662**	0.569	**0.619**
x_{46}	0.632	**0.775**	0.669	**0.67**	0.622	**0.76**	**0.64**	0.6
x_{47}	0.627	**0.652**	**0.616**	0.581	0.584	**0.655**	0.537	**0.572**
x_{48}	0.619	**0.715**	0.648	**0.733**	0.595	**0.707**	**0.633**	0.62
x_{49}	0.574	**0.64**	**0.623**	0.617	0.606	**0.65**	0.569	**0.619**

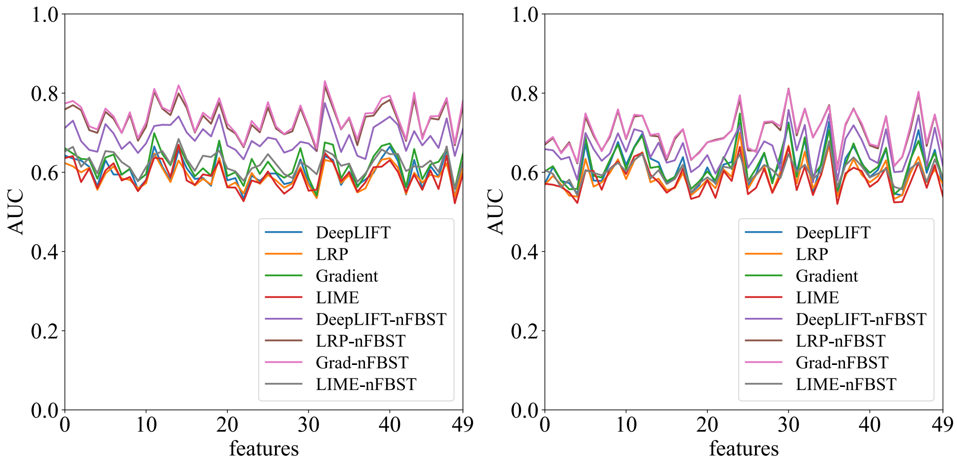
Table 5: Comparison of AUC for each feature before and after *n*FBST under eps=0.001 on Dataset 2.

Feature	Gradient	Grad-*n*FBST	DeepLIFT	DeepLIFT-*n*FBST	LRP	LRP-*n*FBST	LIME	LIME-*n*FBST
x_0	0.653	**0.772**	0.646	**0.697**	0.624	**0.755**	0.627	**0.655**
x_1	0.643	**0.772**	0.624	**0.733**	0.599	**0.762**	0.614	**0.644**
x_2	0.624	**0.75**	0.615	**0.663**	0.591	**0.742**	0.553	**0.611**
x_3	0.641	**0.712**	0.61	**0.654**	0.608	**0.707**	0.601	**0.626**
x_4	0.588	**0.706**	0.572	**0.654**	0.563	**0.695**	0.562	**0.603**
x_5	0.626	**0.751**	0.625	**0.717**	0.595	**0.735**	0.597	**0.654**
x_6	0.641	**0.736**	0.582	**0.699**	0.6	**0.726**	0.606	**0.64**
x_7	0.598	**0.695**	0.594	**0.661**	0.587	**0.695**	0.584	**0.622**
x_8	0.621	**0.745**	0.606	**0.669**	0.593	**0.748**	0.584	**0.622**
x_9	0.589	**0.683**	0.563	**0.655**	0.572	**0.68**	0.558	**0.579**
x_{10}	0.604	**0.702**	0.574	**0.653**	0.571	**0.689**	0.581	**0.604**
x_{11}	0.678	**0.799**	0.648	**0.717**	0.625	**0.789**	0.63	**0.635**
x_{12}	0.625	**0.769**	0.614	**0.726**	0.596	**0.764**	0.63	**0.642**
x_{13}	0.622	**0.75**	0.562	**0.715**	0.568	**0.743**	0.588	**0.614**
x_{14}	0.674	**0.802**	0.658	**0.744**	0.634	**0.775**	0.65	**0.672**
x_{15}	0.619	**0.76**	0.606	**0.69**	0.596	**0.756**	0.571	**0.621**
x_{16}	0.581	**0.703**	0.588	**0.684**	0.56	**0.7**	0.557	**0.597**
x_{17}	0.608	**0.748**	0.591	**0.707**	0.584	**0.738**	0.598	**0.638**
x_{18}	0.593	**0.728**	0.575	**0.677**	0.57	**0.715**	0.594	**0.64**
x_{19}	0.67	**0.773**	0.672	**0.738**	0.623	**0.766**	0.625	**0.65**
x_{20}	0.576	**0.715**	0.58	**0.657**	0.557	**0.701**	0.56	**0.615**
x_{21}	0.588	**0.694**	0.57	**0.66**	0.559	**0.693**	0.56	**0.59**
x_{22}	0.562	**0.649**	0.543	**0.614**	0.548	**0.647**	0.541	**0.584**
x_{23}	0.616	**0.717**	0.581	**0.672**	0.589	**0.711**	0.576	**0.608**
x_{24}	0.583	**0.691**	0.576	**0.651**	0.558	**0.687**	0.561	**0.605**
x_{25}	0.623	**0.766**	0.603	**0.667**	0.593	**0.758**	0.591	**0.641**
x_{26}	0.597	**0.71**	0.589	**0.67**	0.574	**0.704**	0.582	**0.621**
x_{27}	0.583	**0.686**	0.58	**0.64**	0.568	**0.684**	0.54	**0.594**
x_{28}	0.595	**0.697**	0.566	**0.634**	0.567	**0.692**	0.56	**0.568**
x_{29}	0.638	**0.766**	0.634	**0.675**	0.599	**0.757**	0.613	**0.631**
x_{30}	0.575	**0.703**	0.566	**0.671**	0.56	**0.697**	0.534	**0.612**
x_{31}	0.54	**0.653**	0.548	**0.648**	0.534	**0.652**	0.554	**0.594**
x_{32}	0.665	**0.829**	0.644	**0.791**	0.63	**0.814**	0.658	**0.662**
x_{33}	0.674	**0.75**	0.632	**0.699**	0.623	**0.74**	0.607	**0.619**
x_{34}	0.586	**0.693**	0.553	**0.64**	0.572	**0.689**	0.566	**0.619**
x_{35}	0.618	**0.744**	0.606	**0.702**	0.595	**0.743**	0.613	**0.633**
x_{36}	0.559	**0.686**	0.562	**0.643**	0.543	**0.668**	0.532	**0.572**
x_{37}	0.601	**0.744**	0.583	**0.636**	0.568	**0.737**	0.583	**0.613**
x_{38}	0.638	**0.738**	0.583	**0.702**	0.597	**0.73**	0.609	**0.613**
x_{39}	0.683	**0.764**	0.657	**0.72**	0.643	**0.748**	0.625	**0.663**
x_{40}	0.664	**0.785**	0.675	**0.736**	0.635	**0.781**	0.638	**0.654**
x_{41}	0.629	**0.726**	0.624	**0.701**	0.604	**0.717**	0.603	**0.637**
x_{42}	0.564	**0.673**	0.54	**0.643**	0.541	**0.67**	0.539	**0.595**
x_{43}	0.673	**0.788**	0.632	**0.702**	0.609	**0.769**	0.599	**0.611**
x_{44}	0.59	**0.671**	0.562	**0.657**	0.583	**0.675**	0.541	**0.609**
x_{45}	0.621	**0.716**	0.623	**0.676**	0.599	**0.707**	0.604	**0.623**
x_{46}	0.615	**0.72**	0.584	**0.645**	0.583	**0.712**	0.562	**0.593**
x_{47}	0.648	**0.773**	0.645	**0.731**	0.605	**0.765**	0.63	**0.644**
x_{48}	0.544	**0.661**	0.547	**0.631**	0.527	**0.659**	0.518	**0.565**
x_{49}	0.644	**0.776**	0.635	**0.706**	0.61	**0.763**	0.594	**0.629**

Table 6: Comparison of AUC for each feature before and after *n*FBST under eps=0.01 on Dataset 1.

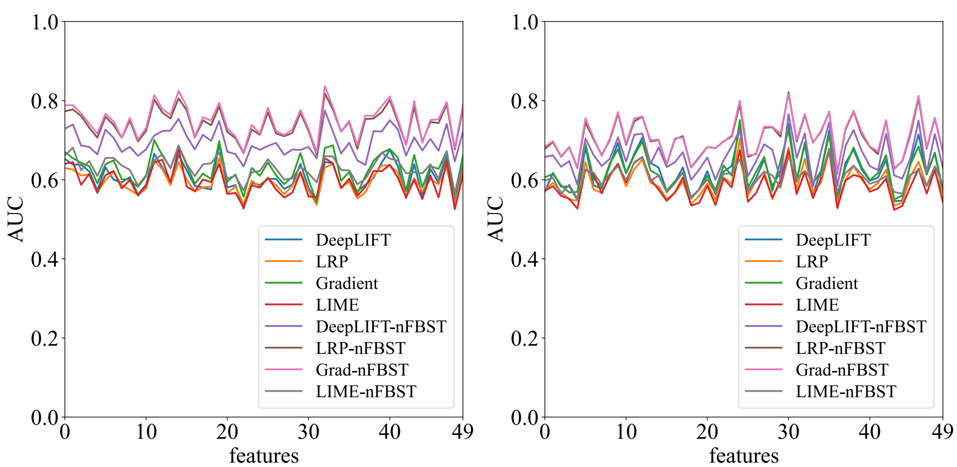
Feature	Gradient	Grad-*n*FBST	DeepLIFT	DeepLIFT-*n*FBST	LRP	LRP-*n*FBST	LIME	LIME-*n*FBST
x_0	0.588	**0.661**	0.567	**0.646**	0.561	**0.657**	0.584	**0.607**
x_1	0.635	**0.695**	0.611	**0.657**	0.599	**0.691**	0.568	**0.592**
x_2	0.568	**0.64**	0.574	**0.638**	0.556	**0.643**	**0.565**	0.559
x_3	0.553	**0.659**	0.56	**0.627**	0.536	**0.658**	0.538	**0.568**
x_4	0.563	**0.617**	0.545	**0.58**	0.539	**0.613**	0.529	**0.534**
x_5	0.666	**0.746**	0.665	**0.695**	0.616	**0.74**	0.594	**0.597**
x_6	0.607	**0.699**	0.559	**0.662**	0.574	**0.688**	**0.593**	0.589
x_7	0.561	**0.641**	0.57	**0.615**	0.556	**0.643**	0.548	**0.58**
x_8	0.635	**0.701**	0.633	**0.661**	0.608	**0.703**	**0.629**	0.617
x_9	0.658	**0.764**	0.67	**0.706**	0.617	**0.762**	**0.623**	0.611
x_{10}	0.619	**0.683**	0.618	**0.636**	0.591	**0.682**	**0.589**	0.587
x_{11}	0.669	**0.737**	0.667	**0.705**	0.634	**0.738**	**0.635**	0.627
x_{12}	0.691	**0.739**	0.68	**0.693**	0.642	**0.735**	**0.619**	0.614
x_{13}	0.599	**0.682**	**0.635**	0.632	0.567	**0.68**	**0.589**	0.572
x_{14}	0.611	**0.682**	0.624	**0.661**	0.587	**0.675**	0.568	**0.603**
x_{15}	0.562	**0.634**	0.569	**0.613**	0.542	**0.629**	0.54	**0.56**
x_{16}	0.589	**0.678**	0.571	**0.623**	0.558	**0.673**	0.542	**0.575**
x_{17}	0.617	**0.707**	0.625	**0.673**	0.591	**0.704**	0.586	**0.602**
x_{18}	0.547	**0.617**	0.531	**0.589**	0.538	**0.613**	0.513	**0.544**
x_{19}	0.572	**0.637**	0.559	**0.609**	0.562	**0.639**	0.53	**0.576**
x_{20}	0.611	**0.675**	0.604	**0.648**	0.579	**0.677**	**0.583**	0.582
x_{21}	0.557	**0.667**	0.558	**0.583**	0.545	**0.67**	0.532	**0.575**
x_{22}	0.632	**0.682**	0.619	**0.632**	0.603	**0.685**	0.603	**0.628**
x_{23}	0.601	**0.689**	0.604	**0.653**	0.579	**0.691**	0.558	**0.594**
x_{24}	0.743	**0.798**	**0.722**	0.719	0.691	**0.779**	**0.662**	0.645
x_{25}	0.579	**0.656**	0.567	**0.613**	0.558	**0.649**	0.534	**0.572**
x_{26}	0.597	**0.652**	0.578	**0.643**	0.574	**0.659**	0.55	**0.582**
x_{27}	0.639	**0.718**	0.64	**0.675**	0.611	**0.714**	0.59	**0.611**
x_{28}	0.575	**0.729**	0.584	**0.678**	0.557	**0.725**	0.55	**0.604**
x_{29}	0.642	**0.715**	**0.621**	0.598	0.597	**0.702**	**0.598**	0.586
x_{30}	0.705	**0.794**	0.715	**0.738**	0.658	**0.802**	**0.641**	0.629
x_{31}	0.591	**0.684**	0.602	**0.648**	0.577	**0.684**	0.545	**0.59**
x_{32}	0.685	**0.745**	0.686	**0.711**	0.657	**0.746**	0.617	**0.623**
x_{33}	0.593	**0.681**	0.554	**0.562**	0.565	**0.679**	0.536	**0.579**
x_{34}	0.611	**0.704**	0.609	**0.662**	0.586	**0.706**	0.584	**0.586**
x_{35}	0.703	**0.755**	0.719	**0.728**	0.662	**0.757**	**0.671**	0.663
x_{36}	0.544	**0.611**	0.533	**0.576**	0.529	**0.615**	0.511	**0.552**
x_{37}	0.614	**0.721**	0.616	**0.694**	0.581	**0.721**	0.584	**0.606**
x_{38}	0.653	**0.737**	0.653	**0.712**	0.619	**0.74**	0.601	**0.628**
x_{39}	0.631	**0.713**	0.619	**0.678**	0.593	**0.711**	0.601	**0.601**
x_{40}	0.577	**0.66**	0.577	**0.622**	0.556	**0.658**	0.557	**0.575**
x_{41}	0.616	**0.653**	0.594	**0.613**	0.593	**0.646**	0.576	**0.587**
x_{42}	0.667	**0.744**	0.651	**0.711**	0.636	**0.745**	**0.608**	0.606
x_{43}	0.545	**0.599**	0.54	**0.592**	0.527	**0.597**	0.527	**0.564**
x_{44}	0.561	**0.629**	0.533	**0.59**	0.542	**0.626**	0.529	**0.547**
x_{45}	0.64	**0.7**	0.626	**0.642**	0.603	**0.696**	0.573	**0.595**
x_{46}	0.67	**0.804**	0.693	**0.744**	0.634	**0.799**	**0.626**	0.623
x_{47}	0.607	**0.668**	**0.613**	0.611	0.581	**0.66**	0.572	**0.577**
x_{48}	0.637	**0.734**	0.649	**0.708**	0.595	**0.731**	0.598	**0.621**
x_{49}	0.573	**0.661**	0.563	**0.617**	0.565	**0.669**	0.544	**0.585**

Table 7: Comparison of AUC for each feature before and after *n*FBST under eps=0.01 on Dataset 2.



(a) Dataset 1 (b) Dataset 2

Figure 5: Comparison of AUC for each feature before and after *n*FBST under eps=0.02.



(a) Dataset 1 (b) Dataset 2

Figure 6: Comparison of AUC for each feature before and after *n*FBST under eps=0.03.

Feature	Gradient	Grad-*n*FBST	DeepLIFT	DeepLIFT-*n*FBST	LRP	LRP-*n*FBST	LIME	LIME-*n*FBST
x_0	0.661	**0.774**	0.643	**0.712**	0.623	**0.759**	0.636	**0.653**
x_1	0.647	**0.78**	0.635	**0.73**	0.615	**0.77**	0.641	**0.664**
x_2	0.634	**0.765**	0.63	**0.679**	0.6	**0.757**	0.575	**0.614**
x_3	0.632	**0.716**	0.614	**0.657**	0.612	**0.707**	0.609	**0.637**
x_4	0.584	**0.708**	0.565	**0.652**	0.556	**0.7**	0.56	**0.597**
x_5	0.637	**0.761**	0.63	**0.722**	0.597	**0.753**	0.605	**0.654**
x_6	0.645	**0.741**	0.594	**0.698**	0.615	**0.735**	0.623	**0.651**
x_7	0.594	**0.7**	0.595	**0.659**	0.584	**0.699**	0.579	**0.626**
x_8	0.609	**0.752**	0.608	**0.677**	0.581	**0.746**	0.588	**0.612**
x_9	0.578	**0.686**	0.557	**0.649**	0.559	**0.68**	0.552	**0.577**
x_{10}	0.595	**0.726**	0.58	**0.673**	0.572	**0.715**	0.58	**0.613**
x_{11}	0.699	**0.811**	0.666	**0.717**	0.648	**0.803**	0.637	**0.643**
x_{12}	0.646	**0.764**	0.615	**0.72**	0.611	**0.761**	0.635	**0.654**
x_{13}	0.618	**0.754**	0.58	**0.72**	0.575	**0.744**	0.588	**0.622**
x_{14}	0.669	**0.82**	0.656	**0.741**	0.63	**0.799**	0.67	**0.684**
x_{15}	0.626	**0.77**	0.61	**0.701**	0.597	**0.762**	0.579	**0.633**
x_{16}	0.59	**0.702**	0.583	**0.679**	0.567	**0.7**	0.568	**0.604**
x_{17}	0.61	**0.751**	0.586	**0.709**	0.584	**0.744**	0.595	**0.642**
x_{18}	0.591	**0.734**	0.566	**0.691**	0.569	**0.723**	0.591	**0.639**
x_{19}	0.68	**0.787**	0.678	**0.746**	0.637	**0.776**	0.629	**0.655**
x_{20}	0.591	**0.723**	0.58	**0.668**	0.564	**0.712**	0.562	**0.61**
x_{21}	0.6	**0.701**	0.585	**0.657**	0.574	**0.698**	0.562	**0.602**
x_{22}	0.566	**0.666**	0.533	**0.633**	0.546	**0.663**	0.527	**0.579**
x_{23}	0.634	**0.73**	0.592	**0.679**	0.598	**0.72**	0.58	**0.61**
x_{24}	0.596	**0.706**	0.578	**0.666**	0.571	**0.701**	0.573	**0.618**
x_{25}	0.627	**0.777**	0.597	**0.688**	0.593	**0.764**	0.602	**0.646**
x_{26}	0.6	**0.716**	0.597	**0.683**	0.58	**0.71**	0.57	**0.621**
x_{27}	0.586	**0.697**	0.57	**0.65**	0.561	**0.696**	0.546	**0.596**
x_{28}	0.599	**0.711**	0.575	**0.657**	0.572	**0.703**	0.562	**0.586**
x_{29}	0.661	**0.769**	0.632	**0.677**	0.613	**0.762**	0.608	**0.627**
x_{30}	0.586	**0.716**	0.573	**0.673**	0.568	**0.711**	0.553	**0.613**
x_{31}	0.541	**0.66**	0.545	**0.653**	0.535	**0.654**	0.556	**0.595**
x_{32}	0.676	**0.83**	0.641	**0.775**	0.631	**0.817**	0.649	**0.656**
x_{33}	0.67	**0.771**	0.631	**0.715**	0.626	**0.763**	0.626	**0.645**
x_{34}	0.585	**0.709**	0.568	**0.651**	0.573	**0.708**	0.578	**0.617**
x_{35}	0.619	**0.739**	0.602	**0.695**	0.593	**0.737**	0.6	**0.622**
x_{36}	0.563	**0.683**	0.567	**0.642**	0.549	**0.669**	0.551	**0.576**
x_{37}	0.594	**0.749**	0.582	**0.634**	0.559	**0.74**	0.575	**0.597**
x_{38}	0.64	**0.75**	0.598	**0.714**	0.602	**0.743**	0.613	**0.63**
x_{39}	0.667	**0.787**	0.636	**0.727**	0.632	**0.771**	0.614	**0.658**
x_{40}	0.673	**0.794**	0.665	**0.741**	0.636	**0.783**	0.631	**0.646**
x_{41}	0.639	**0.738**	0.629	**0.72**	0.612	**0.732**	0.601	**0.647**
x_{42}	0.563	**0.683**	0.544	**0.654**	0.546	**0.679**	0.55	**0.603**
x_{43}	0.659	**0.801**	0.631	**0.704**	0.599	**0.786**	0.594	**0.615**
x_{44}	0.583	**0.684**	0.559	**0.668**	0.574	**0.689**	0.556	**0.614**
x_{45}	0.62	**0.741**	0.616	**0.692**	0.594	**0.737**	0.605	**0.63**
x_{46}	0.627	**0.742**	0.589	**0.662**	0.588	**0.733**	0.558	**0.601**
x_{47}	0.658	**0.788**	0.658	**0.736**	0.623	**0.782**	0.643	**0.666**
x_{48}	0.559	**0.672**	0.538	**0.641**	0.538	**0.668**	0.522	**0.558**
x_{49}	0.647	**0.781**	0.625	**0.709**	0.61	**0.767**	0.595	**0.627**

Table 8: Comparison of AUC for each feature before and after *n*FBST under eps=0.02 on Dataset 1.

Feature	Gradient	Grad-*n*FBST	DeepLIFT	DeepLIFT-*n*FBST	LRP	LRP-*n*FBST	LIME	LIME-*n*FBST
x_0	0.603	**0.674**	0.572	**0.658**	0.569	**0.67**	0.571	**0.605**
x_1	0.615	**0.689**	0.613	**0.656**	0.591	**0.686**	0.569	**0.596**
x_2	0.58	**0.648**	0.578	**0.633**	0.569	**0.65**	**0.563**	0.562
x_3	0.557	**0.673**	0.573	**0.639**	0.541	**0.676**	0.55	**0.582**
x_4	0.558	**0.625**	0.546	**0.581**	0.539	**0.623**	0.523	**0.542**
x_5	0.689	**0.749**	0.672	**0.7**	0.634	**0.74**	0.604	**0.606**
x_6	0.591	**0.697**	0.579	**0.659**	0.564	**0.692**	**0.603**	0.599
x_7	0.587	**0.653**	0.578	**0.629**	0.575	**0.655**	0.556	**0.592**
x_8	0.631	**0.693**	0.623	**0.647**	0.598	**0.689**	**0.609**	0.608
x_9	0.67	**0.759**	0.677	**0.699**	0.633	**0.755**	**0.627**	0.624
x_{10}	0.608	**0.686**	0.619	**0.652**	0.583	**0.684**	0.596	**0.602**
x_{11}	0.663	**0.746**	0.664	**0.709**	0.629	**0.741**	**0.64**	0.631
x_{12}	0.693	**0.743**	0.698	**0.703**	0.651	**0.743**	**0.649**	0.646
x_{13}	0.611	**0.694**	0.635	**0.638**	0.575	**0.693**	0.594	**0.584**
x_{14}	0.615	**0.698**	0.625	**0.675**	0.584	**0.691**	0.575	**0.606**
x_{15}	0.578	**0.635**	0.572	**0.624**	0.555	**0.632**	0.549	**0.571**
x_{16}	0.588	**0.689**	0.589	**0.63**	0.56	**0.684**	0.563	**0.583**
x_{17}	0.619	**0.709**	0.638	**0.674**	0.598	**0.709**	0.61	**0.619**
x_{18}	0.558	**0.633**	0.548	**0.6**	0.542	**0.631**	0.533	**0.549**
x_{19}	0.577	**0.649**	0.567	**0.616**	0.56	**0.651**	0.539	**0.572**
x_{20}	0.602	**0.676**	0.613	**0.644**	0.581	**0.676**	0.581	**0.587**
x_{21}	0.579	**0.68**	0.571	**0.597**	0.56	**0.682**	0.536	**0.573**
x_{22}	0.639	**0.685**	0.62	**0.634**	0.607	**0.687**	0.604	**0.614**
x_{23}	0.608	**0.706**	0.626	**0.677**	0.588	**0.706**	0.578	**0.61**
x_{24}	0.749	**0.795**	**0.725**	0.718	0.701	**0.784**	**0.664**	0.646
x_{25}	0.577	**0.662**	0.566	**0.611**	0.558	**0.655**	0.545	**0.576**
x_{26}	0.61	**0.651**	0.596	**0.642**	0.593	**0.655**	0.561	**0.594**
x_{27}	0.65	**0.727**	0.649	**0.678**	0.616	**0.724**	0.609	**0.619**
x_{28}	0.572	**0.729**	0.581	**0.668**	0.55	**0.727**	0.548	**0.606**
x_{29}	0.646	**0.716**	**0.627**	0.605	0.602	**0.706**	**0.59**	0.585
x_{30}	0.723	**0.81**	0.726	**0.758**	0.667	**0.812**	**0.664**	0.646
x_{31}	0.588	**0.699**	0.601	**0.655**	0.572	**0.693**	0.555	**0.601**
x_{32}	0.689	**0.761**	0.684	**0.72**	0.652	**0.76**	0.615	**0.619**
x_{33}	0.589	**0.687**	0.548	**0.575**	0.56	**0.688**	0.543	**0.583**
x_{34}	0.618	**0.725**	0.622	**0.671**	0.593	**0.725**	0.594	**0.601**
x_{35}	0.706	**0.768**	0.728	**0.748**	0.668	**0.77**	0.678	**0.679**
x_{36}	0.554	**0.606**	0.54	**0.579**	0.539	**0.61**	0.52	**0.563**
x_{37}	0.62	**0.722**	0.635	**0.685**	0.59	**0.719**	0.598	**0.618**
x_{38}	0.673	**0.76**	0.666	**0.719**	0.633	**0.761**	0.612	**0.638**
x_{39}	0.621	**0.724**	0.625	**0.682**	0.594	**0.718**	0.603	**0.615**
x_{40}	0.599	**0.671**	0.588	**0.634**	0.575	**0.668**	0.563	**0.587**
x_{41}	0.614	**0.666**	0.606	**0.623**	0.59	**0.66**	0.578	**0.597**
x_{42}	0.662	**0.742**	0.659	**0.709**	0.631	**0.742**	**0.612**	0.609
x_{43}	0.544	**0.617**	0.546	**0.6**	0.533	**0.616**	0.524	**0.563**
x_{44}	0.562	**0.639**	0.542	**0.602**	0.543	**0.639**	0.525	**0.557**
x_{45}	0.633	**0.716**	0.629	**0.657**	0.598	**0.707**	0.574	**0.61**
x_{46}	0.68	**0.804**	0.707	**0.744**	0.639	**0.798**	0.625	**0.625**
x_{47}	0.609	**0.673**	0.599	**0.615**	0.574	**0.667**	0.561	**0.576**
x_{48}	0.649	**0.746**	0.657	**0.713**	0.607	**0.744**	0.618	**0.633**
x_{49}	0.584	**0.658**	0.566	**0.617**	0.57	**0.659**	0.54	**0.584**

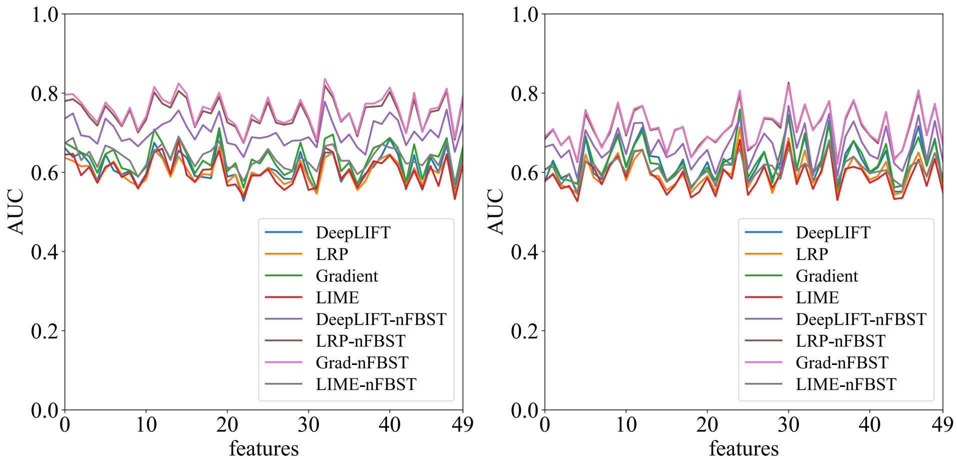
Table 9: Comparison of AUC for each feature before and after *n*FBST under eps=0.02 on Dataset 2.

Feature	Gradient	Grad-*n*FBST	DeepLIFT	DeepLIFT-*n*FBST	LRP	LRP-*n*FBST	LIME	LIME-*n*FBST
x_0	0.671	**0.788**	0.653	**0.729**	0.63	**0.773**	0.64	**0.661**
x_1	0.654	**0.789**	0.638	**0.74**	0.626	**0.778**	0.644	**0.681**
x_2	0.643	**0.768**	0.638	**0.686**	0.611	**0.761**	0.587	**0.624**
x_3	0.635	**0.742**	0.621	**0.683**	0.613	**0.732**	0.615	**0.647**
x_4	0.591	**0.717**	0.573	**0.664**	0.569	**0.706**	0.567	**0.609**
x_5	0.639	**0.767**	0.631	**0.727**	0.598	**0.759**	0.612	**0.655**
x_6	0.649	**0.748**	0.6	**0.706**	0.62	**0.739**	0.622	**0.655**
x_7	0.601	**0.706**	0.593	**0.67**	0.586	**0.706**	0.578	**0.636**
x_8	0.601	**0.756**	0.607	**0.683**	0.574	**0.75**	0.599	**0.625**
x_9	0.582	**0.702**	0.563	**0.66**	0.562	**0.697**	0.56	**0.585**
x_{10}	0.609	**0.731**	0.587	**0.677**	0.578	**0.724**	0.586	**0.612**
x_{11}	0.7	**0.814**	0.666	**0.712**	0.649	**0.802**	0.644	**0.649**
x_{12}	0.662	**0.779**	0.631	**0.723**	0.626	**0.77**	0.651	**0.664**
x_{13}	0.63	**0.764**	0.592	**0.724**	0.587	**0.755**	0.593	**0.629**
x_{14}	0.685	**0.825**	0.663	**0.754**	0.645	**0.806**	0.679	**0.687**
x_{15}	0.637	**0.785**	0.624	**0.71**	0.6	**0.776**	0.582	**0.64**
x_{16}	0.59	**0.708**	0.585	**0.677**	0.57	**0.707**	0.571	**0.609**
x_{17}	0.615	**0.758**	0.58	**0.714**	0.581	**0.75**	0.601	**0.639**
x_{18}	0.602	**0.75**	0.575	**0.698**	0.582	**0.738**	0.594	**0.642**
x_{19}	0.698	**0.795**	0.691	**0.748**	0.654	**0.785**	0.639	**0.669**
x_{20}	0.595	**0.73**	0.58	**0.671**	0.566	**0.721**	0.564	**0.603**
x_{21}	0.613	**0.71**	0.587	**0.665**	0.583	**0.705**	0.567	**0.61**
x_{22}	0.557	**0.669**	0.527	**0.634**	0.538	**0.667**	0.53	**0.573**
x_{23}	0.63	**0.739**	0.592	**0.686**	0.598	**0.727**	0.586	**0.619**
x_{24}	0.61	**0.717**	0.587	**0.676**	0.583	**0.714**	0.58	**0.622**
x_{25}	0.644	**0.782**	0.602	**0.686**	0.604	**0.77**	0.605	**0.652**
x_{26}	0.611	**0.723**	0.602	**0.693**	0.586	**0.717**	0.575	**0.627**
x_{27}	0.589	**0.713**	0.577	**0.663**	0.563	**0.71**	0.555	**0.6**
x_{28}	0.607	**0.727**	0.588	**0.676**	0.584	**0.718**	0.567	**0.603**
x_{29}	0.667	**0.776**	0.64	**0.675**	0.62	**0.769**	0.618	**0.633**
x_{30}	0.6	**0.728**	0.58	**0.683**	0.573	**0.723**	0.558	**0.616**
x_{31}	0.544	**0.67**	0.542	**0.654**	0.537	**0.664**	0.556	**0.593**
x_{32}	0.68	**0.837**	0.645	**0.775**	0.631	**0.818**	0.654	**0.66**
x_{33}	0.687	**0.784**	0.642	**0.722**	0.64	**0.777**	0.642	**0.659**
x_{34}	0.597	**0.722**	0.579	**0.666**	0.58	**0.721**	0.579	**0.625**
x_{35}	0.62	**0.749**	0.598	**0.694**	0.595	**0.746**	0.603	**0.624**
x_{36}	0.566	**0.687**	0.571	**0.649**	0.553	**0.678**	0.559	**0.589**
x_{37}	0.61	**0.763**	0.587	**0.638**	0.568	**0.754**	0.583	**0.602**
x_{38}	0.648	**0.761**	0.605	**0.723**	0.609	**0.755**	0.622	**0.637**
x_{39}	0.668	**0.786**	0.643	**0.722**	0.636	**0.771**	0.622	**0.668**
x_{40}	0.677	**0.811**	0.676	**0.751**	0.636	**0.802**	0.638	**0.654**
x_{41}	0.656	**0.753**	0.634	**0.727**	0.618	**0.746**	0.605	**0.648**
x_{42}	0.572	**0.688**	0.554	**0.661**	0.558	**0.688**	0.556	**0.616**
x_{43}	0.663	**0.799**	0.639	**0.707**	0.604	**0.784**	0.6	**0.619**
x_{44}	0.581	**0.694**	0.559	**0.674**	0.571	**0.696**	0.552	**0.616**
x_{45}	0.635	**0.751**	0.627	**0.703**	0.601	**0.746**	0.611	**0.639**
x_{46}	0.627	**0.753**	0.592	**0.673**	0.589	**0.743**	0.556	**0.603**
x_{47}	0.672	**0.797**	0.66	**0.741**	0.632	**0.79**	0.647	**0.672**
x_{48}	0.563	**0.68**	0.545	**0.646**	0.538	**0.677**	0.526	**0.566**
x_{49}	0.66	**0.788**	0.637	**0.721**	0.619	**0.779**	0.611	**0.638**

Table 10: Comparison of AUC for each feature before and after *n*FBST under eps=0.03 on Dataset 1.

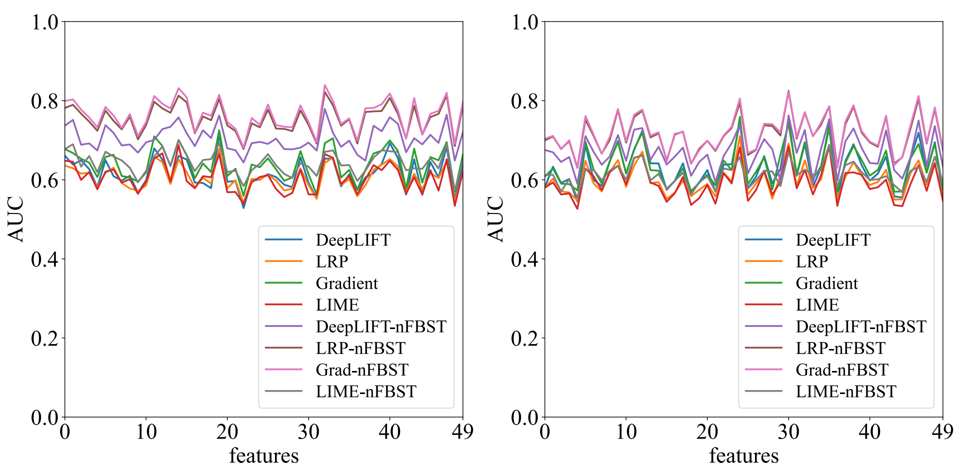
Feature	Gradient	Grad-nFBST	DeepLIFT	DeepLIFT-nFBST	LRP	LRP-nFBST	LIME	LIME-nFBST
x_0	0.606	**0.684**	0.57	**0.658**	0.573	**0.678**	0.572	**0.598**
x_1	0.615	**0.698**	0.617	**0.662**	0.592	**0.695**	0.584	**0.605**
x_2	0.586	**0.659**	0.582	**0.633**	0.569	**0.658**	0.559	**0.564**
x_3	0.568	**0.682**	0.584	**0.647**	0.55	**0.683**	0.552	**0.588**
x_4	0.569	**0.632**	0.549	**0.588**	0.548	**0.63**	0.527	**0.552**
x_5	0.691	**0.756**	0.683	**0.704**	0.645	**0.745**	**0.626**	0.624
x_6	0.603	**0.706**	0.586	**0.663**	0.575	**0.702**	0.614	**0.618**
x_7	0.583	**0.66**	0.578	**0.635**	0.567	**0.663**	0.566	**0.588**
x_8	0.642	**0.704**	0.637	**0.653**	0.609	**0.7**	0.613	**0.615**
x_9	0.677	**0.771**	0.693	**0.707**	0.639	**0.767**	**0.641**	0.634
x_{10}	0.616	**0.691**	0.617	**0.646**	0.584	**0.688**	0.593	**0.599**
x_{11}	0.661	**0.754**	0.665	**0.714**	0.626	**0.748**	**0.644**	0.642
x_{12}	0.697	**0.76**	0.707	**0.717**	0.651	**0.759**	**0.658**	0.657
x_{13}	0.624	**0.7**	**0.643**	0.64	0.592	**0.697**	**0.602**	0.597
x_{14}	0.612	**0.702**	0.631	**0.682**	0.583	**0.7**	0.575	**0.611**
x_{15}	0.572	**0.64**	0.572	**0.632**	0.551	**0.638**	0.547	**0.568**
x_{16}	0.596	**0.705**	0.594	**0.644**	0.568	**0.702**	0.567	**0.591**
x_{17}	0.617	**0.707**	0.631	**0.67**	0.596	**0.71**	0.606	**0.619**
x_{18}	0.556	**0.63**	0.552	**0.6**	0.538	**0.631**	0.534	**0.553**
x_{19}	0.576	**0.655**	0.576	**0.626**	0.56	**0.657**	0.542	**0.576**
x_{20}	0.61	**0.683**	0.622	**0.656**	0.591	**0.683**	0.585	**0.601**
x_{21}	0.574	**0.68**	0.563	**0.6**	0.549	**0.68**	0.536	**0.567**
x_{22}	0.636	**0.696**	0.621	**0.649**	0.606	**0.697**	0.603	**0.615**
x_{23}	0.61	**0.713**	0.635	**0.684**	0.59	**0.709**	0.581	**0.617**
x_{24}	0.751	**0.8**	**0.733**	0.727	0.704	**0.79**	**0.675**	0.652
x_{25}	0.58	**0.662**	0.57	**0.609**	0.559	**0.658**	0.544	**0.573**
x_{26}	0.62	**0.664**	0.604	**0.647**	0.599	**0.666**	0.566	**0.595**
x_{27}	0.658	**0.734**	0.651	**0.69**	0.621	**0.731**	0.612	**0.621**
x_{28}	0.572	**0.734**	0.583	**0.676**	0.551	**0.733**	0.555	**0.611**
x_{29}	0.647	**0.717**	**0.622**	0.6	0.602	**0.709**	**0.595**	0.58
x_{30}	0.726	**0.816**	0.742	**0.766**	0.681	**0.821**	**0.674**	0.66
x_{31}	0.596	**0.701**	0.6	**0.655**	0.576	**0.697**	0.564	**0.602**
x_{32}	0.695	**0.766**	0.69	**0.717**	0.653	**0.764**	0.62	**0.623**
x_{33}	0.591	**0.693**	0.564	**0.583**	0.568	**0.694**	0.553	**0.596**
x_{34}	0.622	**0.724**	0.622	**0.672**	0.593	**0.722**	0.603	**0.608**
x_{35}	0.714	**0.771**	0.724	**0.745**	0.672	**0.772**	0.676	**0.679**
x_{36}	0.563	**0.613**	0.544	**0.585**	0.543	**0.619**	0.528	**0.569**
x_{37}	0.632	**0.729**	0.643	**0.685**	0.599	**0.726**	0.598	**0.617**
x_{38}	0.682	**0.775**	0.676	**0.726**	0.636	**0.773**	0.612	**0.636**
x_{39}	0.632	**0.721**	0.626	**0.673**	0.6	**0.714**	0.606	**0.613**
x_{40}	0.599	**0.685**	0.598	**0.635**	0.576	**0.683**	0.57	**0.59**
x_{41}	0.617	**0.671**	0.605	**0.624**	0.592	**0.665**	0.578	**0.597**
x_{42}	0.662	**0.751**	0.651	**0.716**	0.625	**0.75**	0.603	**0.609**
x_{43}	0.55	**0.627**	0.546	**0.611**	0.534	**0.629**	0.524	**0.569**
x_{44}	0.56	**0.648**	0.548	**0.603**	0.542	**0.647**	0.534	**0.565**
x_{45}	0.639	**0.723**	0.642	**0.655**	0.609	**0.715**	0.582	**0.613**
x_{46}	0.684	**0.812**	0.714	**0.75**	0.646	**0.806**	**0.628**	0.624
x_{47}	0.613	**0.681**	0.613	**0.618**	0.584	**0.678**	0.565	**0.584**
x_{48}	0.669	**0.757**	0.668	**0.716**	0.621	**0.755**	0.625	**0.634**
x_{49}	0.576	**0.673**	0.567	**0.63**	0.566	**0.671**	0.544	**0.588**

Table 11: Comparison of AUC for each feature before and after nFBST under eps=0.03 on Dataset 2.



(a) Dataset 1 (b) Dataset 2

Figure 7: Comparison of AUC for each feature before and after *n*FBST under eps=0.04.



(a) Dataset 1 (b) Dataset 2

Figure 8: Comparison of AUC for each feature before and after *n*FBST under eps=0.05.

Feature	Gradient	Grad-nFBST	DeepLIFT	DeepLIFT-nFBST	LRP	LRP-nFBST	LIME	LIME-nFBST
x_0	0.675	**0.796**	0.661	**0.736**	0.637	**0.78**	0.645	**0.674**
x_1	0.663	**0.798**	0.639	**0.749**	0.629	**0.785**	0.647	**0.687**
x_2	0.65	**0.776**	0.649	**0.693**	0.615	**0.769**	0.592	**0.631**
x_3	0.641	**0.749**	0.625	**0.69**	0.617	**0.74**	0.614	**0.652**
x_4	0.599	**0.726**	0.579	**0.672**	0.577	**0.718**	0.573	**0.613**
x_5	0.646	**0.776**	0.645	**0.736**	0.609	**0.768**	0.614	**0.668**
x_6	0.657	**0.754**	0.605	**0.716**	0.628	**0.744**	0.624	**0.659**
x_7	0.611	**0.72**	0.594	**0.679**	0.593	**0.716**	0.588	**0.644**
x_8	0.605	**0.763**	0.604	**0.685**	0.577	**0.758**	0.599	**0.629**
x_9	0.59	**0.705**	0.562	**0.666**	0.566	**0.7**	0.559	**0.588**
x_{10}	0.614	**0.741**	0.596	**0.686**	0.582	**0.733**	0.592	**0.619**
x_{11}	0.709	**0.816**	0.675	**0.705**	0.657	**0.802**	0.651	**0.654**
x_{12}	0.674	**0.784**	0.64	**0.721**	0.635	**0.774**	0.66	**0.672**
x_{13}	0.631	**0.773**	0.595	**0.73**	0.589	**0.763**	0.595	**0.635**
x_{14}	0.688	**0.825**	0.655	**0.757**	0.64	**0.806**	0.68	**0.691**
x_{15}	0.647	**0.798**	0.637	**0.716**	0.609	**0.788**	0.594	**0.652**
x_{16}	0.597	**0.717**	0.591	**0.683**	0.576	**0.714**	0.575	**0.615**
x_{17}	0.63	**0.766**	0.588	**0.72**	0.596	**0.755**	0.607	**0.646**
x_{18}	0.617	**0.758**	0.585	**0.702**	0.594	**0.749**	0.608	**0.654**
x_{19}	0.712	**0.801**	0.704	**0.754**	0.664	**0.791**	0.654	**0.679**
x_{20}	0.603	**0.737**	0.589	**0.674**	0.575	**0.726**	0.566	**0.611**
x_{21}	0.623	**0.72**	0.595	**0.669**	0.595	**0.715**	0.57	**0.615**
x_{22}	0.561	**0.676**	0.528	**0.638**	0.541	**0.674**	0.538	**0.578**
x_{23}	0.633	**0.745**	0.595	**0.689**	0.6	**0.733**	0.592	**0.62**
x_{24}	0.622	**0.728**	0.592	**0.686**	0.592	**0.723**	0.592	**0.63**
x_{25}	0.653	**0.79**	0.611	**0.69**	0.612	**0.777**	0.608	**0.659**
x_{26}	0.617	**0.732**	0.604	**0.7**	0.592	**0.725**	0.574	**0.63**
x_{27}	0.593	**0.725**	0.584	**0.668**	0.57	**0.72**	0.555	**0.61**
x_{28}	0.603	**0.734**	0.583	**0.68**	0.578	**0.724**	0.57	**0.605**
x_{29}	0.675	**0.783**	0.652	**0.681**	0.63	**0.774**	0.621	**0.638**
x_{30}	0.61	**0.739**	0.587	**0.689**	0.581	**0.733**	0.556	**0.624**
x_{31}	0.557	**0.682**	0.548	**0.663**	0.546	**0.678**	0.561	**0.602**
x_{32}	0.685	**0.836**	0.65	**0.778**	0.638	**0.819**	0.663	**0.666**
x_{33}	0.695	**0.797**	0.649	**0.723**	0.65	**0.789**	0.651	**0.671**
x_{34}	0.603	**0.73**	0.576	**0.675**	0.583	**0.728**	0.584	**0.629**
x_{35}	0.625	**0.75**	0.608	**0.696**	0.603	**0.749**	0.607	**0.629**
x_{36}	0.569	**0.697**	0.572	**0.663**	0.556	**0.691**	0.561	**0.595**
x_{37}	0.618	**0.774**	0.597	**0.645**	0.578	**0.764**	0.594	**0.611**
x_{38}	0.662	**0.774**	0.614	**0.73**	0.622	**0.766**	0.628	**0.645**
x_{39}	0.67	**0.783**	0.642	**0.722**	0.636	**0.768**	0.623	**0.671**
x_{40}	0.687	**0.814**	0.685	**0.752**	0.646	**0.803**	0.643	**0.663**
x_{41}	0.661	**0.764**	0.646	**0.732**	0.624	**0.757**	0.615	**0.659**
x_{42}	0.575	**0.699**	0.557	**0.668**	0.565	**0.699**	0.561	**0.622**
x_{43}	0.671	**0.801**	0.644	**0.703**	0.611	**0.785**	0.605	**0.626**
x_{44}	0.582	**0.702**	0.56	**0.68**	0.571	**0.703**	0.556	**0.618**
x_{45}	0.645	**0.76**	0.638	**0.707**	0.61	**0.753**	0.616	**0.645**
x_{46}	0.64	**0.764**	0.6	**0.688**	0.598	**0.757**	0.566	**0.617**
x_{47}	0.687	**0.812**	0.67	**0.758**	0.639	**0.806**	0.646	**0.679**
x_{48}	0.569	**0.687**	0.552	**0.652**	0.546	**0.682**	0.532	**0.574**
x_{49}	0.664	**0.793**	0.643	**0.72**	0.623	**0.783**	0.618	**0.643**

Table 12: Comparison of AUC for each feature before and after nFBST under eps=0.04 on Dataset 1.

Feature	Gradient	Grad-*n*FBST	DeepLIFT	DeepLIFT-*n*FBST	LRP	LRP-*n*FBST	LIME	LIME-*n*FBST
x_0	0.606	**0.69**	0.577	**0.664**	0.577	**0.684**	0.578	**0.608**
x_1	0.622	**0.709**	0.63	**0.67**	0.597	**0.708**	0.595	**0.614**
x_2	0.583	**0.668**	0.586	**0.637**	0.567	**0.669**	0.559	**0.569**
x_3	0.58	**0.691**	0.595	**0.656**	0.561	**0.69**	0.566	**0.597**
x_4	0.571	**0.629**	0.545	**0.584**	0.546	**0.626**	0.527	**0.549**
x_5	0.692	**0.758**	0.688	**0.706**	0.645	**0.748**	**0.628**	0.626
x_6	0.621	**0.709**	0.598	**0.668**	0.587	**0.705**	0.613	**0.622**
x_7	0.593	**0.662**	0.582	**0.637**	0.57	**0.665**	0.57	**0.588**
x_8	0.643	**0.706**	0.644	**0.654**	0.616	**0.699**	0.617	**0.622**
x_9	0.692	**0.776**	0.697	**0.712**	0.649	**0.773**	**0.641**	0.636
x_{10}	0.613	**0.697**	0.615	**0.646**	0.58	**0.693**	0.588	**0.598**
x_{11}	0.668	**0.761**	0.671	**0.724**	0.633	**0.756**	0.652	**0.654**
x_{12}	0.704	**0.768**	0.714	**0.725**	0.658	**0.768**	**0.657**	0.656
x_{13}	0.623	**0.708**	0.641	**0.646**	0.593	**0.706**	0.598	**0.6**
x_{14}	0.621	**0.713**	0.638	**0.682**	0.591	**0.711**	0.582	**0.612**
x_{15}	0.578	**0.645**	0.578	**0.638**	0.555	**0.647**	0.543	**0.576**
x_{16}	0.598	**0.708**	0.594	**0.648**	0.57	**0.706**	0.567	**0.591**
x_{17}	0.624	**0.713**	0.632	**0.673**	0.595	**0.714**	0.602	**0.614**
x_{18}	0.564	**0.638**	0.558	**0.607**	0.549	**0.638**	0.536	**0.559**
x_{19}	0.586	**0.664**	0.588	**0.635**	0.572	**0.669**	0.55	**0.588**
x_{20}	0.612	**0.69**	0.626	**0.657**	0.591	**0.69**	0.586	**0.604**
x_{21}	0.58	**0.678**	0.565	**0.596**	0.551	**0.675**	0.54	**0.568**
x_{22}	0.634	**0.701**	0.625	**0.651**	0.605	**0.7**	0.608	**0.619**
x_{23}	0.618	**0.721**	0.636	**0.688**	0.594	**0.719**	0.584	**0.619**
x_{24}	0.759	**0.806**	**0.748**	0.739	0.714	**0.795**	**0.682**	0.659
x_{25}	0.587	**0.66**	0.575	**0.608**	0.563	**0.655**	0.543	**0.572**
x_{26}	0.622	**0.672**	0.609	**0.657**	0.598	**0.672**	0.573	**0.596**
x_{27}	0.652	**0.738**	0.652	**0.698**	0.617	**0.736**	0.615	**0.62**
x_{28}	0.573	**0.736**	0.585	**0.674**	0.548	**0.733**	0.56	**0.617**
x_{29}	0.652	**0.719**	**0.625**	0.602	0.605	**0.712**	**0.594**	0.583
x_{30}	0.739	**0.821**	0.745	**0.768**	0.687	**0.827**	**0.677**	0.663
x_{31}	0.607	**0.705**	0.609	**0.661**	0.581	**0.701**	0.57	**0.604**
x_{32}	0.699	**0.772**	0.692	**0.724**	0.655	**0.77**	0.622	**0.623**
x_{33}	0.595	**0.706**	0.576	**0.592**	0.574	**0.708**	0.558	**0.606**
x_{34}	0.629	**0.737**	0.626	**0.675**	0.597	**0.733**	0.602	**0.615**
x_{35}	0.721	**0.779**	0.731	**0.748**	0.677	**0.78**	0.679	**0.681**
x_{36}	0.564	**0.621**	0.544	**0.59**	0.545	**0.626**	0.53	**0.575**
x_{37}	0.633	**0.739**	0.647	**0.691**	0.602	**0.734**	0.609	**0.626**
x_{38}	0.679	**0.783**	0.68	**0.729**	0.639	**0.781**	0.614	**0.64**
x_{39}	0.649	**0.724**	0.633	**0.677**	0.61	**0.719**	0.608	**0.616**
x_{40}	0.601	**0.694**	0.599	**0.644**	0.581	**0.693**	0.574	**0.597**
x_{41}	0.616	**0.678**	0.613	**0.634**	0.591	**0.673**	0.584	**0.602**
x_{42}	0.672	**0.753**	0.655	**0.717**	0.627	**0.752**	0.599	**0.606**
x_{43}	0.562	**0.631**	0.551	**0.622**	0.544	**0.634**	0.533	**0.58**
x_{44}	0.566	**0.656**	0.551	**0.603**	0.549	**0.655**	0.535	**0.564**
x_{45}	0.642	**0.728**	0.648	**0.658**	0.609	**0.722**	0.587	**0.617**
x_{46}	0.689	**0.807**	0.717	**0.744**	0.65	**0.8**	**0.632**	0.627
x_{47}	0.617	**0.695**	0.617	**0.63**	0.588	**0.694**	0.566	**0.591**
x_{48}	0.685	**0.774**	0.682	**0.73**	0.633	**0.771**	0.634	**0.649**
x_{49}	0.577	**0.672**	0.564	**0.629**	0.562	**0.671**	0.55	**0.587**

Table 13: Comparison of AUC for each feature before and after *n*FBST under eps=0.04 on Dataset 2.

Feature	Gradient	Grad-nFBST	DeepLIFT	DeepLIFT-nFBST	LRP	LRP-nFBST	LIME	LIME-nFBST
x_0	0.678	**0.799**	0.661	**0.737**	0.636	**0.781**	0.648	**0.678**
x_1	0.668	**0.803**	0.639	**0.751**	0.629	**0.79**	0.646	**0.69**
x_2	0.653	**0.777**	0.649	**0.689**	0.616	**0.767**	0.6	**0.633**
x_3	0.643	**0.757**	0.627	**0.692**	0.618	**0.747**	0.62	**0.66**
x_4	0.607	**0.733**	0.581	**0.674**	0.58	**0.724**	0.576	**0.619**
x_5	0.658	**0.784**	0.65	**0.738**	0.617	**0.775**	0.62	**0.671**
x_6	0.663	**0.762**	0.609	**0.719**	0.631	**0.749**	0.626	**0.659**
x_7	0.609	**0.731**	0.598	**0.687**	0.592	**0.727**	0.593	**0.649**
x_8	0.606	**0.766**	0.61	**0.689**	0.578	**0.76**	0.602	**0.63**
x_9	0.595	**0.711**	0.567	**0.669**	0.572	**0.702**	0.564	**0.594**
x_{10}	0.618	**0.746**	0.599	**0.695**	0.586	**0.738**	0.595	**0.622**
x_{11}	0.71	**0.812**	0.679	**0.701**	0.658	**0.797**	0.653	**0.659**
x_{12}	0.686	**0.792**	0.651	**0.728**	0.645	**0.78**	0.667	**0.684**
x_{13}	0.635	**0.781**	0.595	**0.734**	0.589	**0.769**	0.596	**0.636**
x_{14}	0.7	**0.832**	0.661	**0.758**	0.652	**0.813**	0.688	**0.701**
x_{15}	0.662	**0.808**	0.651	**0.718**	0.619	**0.796**	0.596	**0.662**
x_{16}	0.604	**0.72**	0.593	**0.685**	0.58	**0.718**	0.578	**0.622**
x_{17}	0.641	**0.77**	0.592	**0.725**	0.605	**0.76**	0.609	**0.65**
x_{18}	0.619	**0.762**	0.579	**0.705**	0.59	**0.751**	0.606	**0.652**
x_{19}	0.725	**0.815**	0.715	**0.762**	0.678	**0.804**	0.665	**0.693**
x_{20}	0.61	**0.746**	0.595	**0.679**	0.579	**0.737**	0.568	**0.619**
x_{21}	0.628	**0.729**	0.597	**0.675**	0.599	**0.723**	0.569	**0.623**
x_{22}	0.559	**0.681**	0.528	**0.644**	0.541	**0.678**	0.539	**0.584**
x_{23}	0.639	**0.756**	0.598	**0.692**	0.602	**0.743**	0.594	**0.624**
x_{24}	0.632	**0.739**	0.6	**0.696**	0.6	**0.733**	0.607	**0.641**
x_{25}	0.654	**0.79**	0.614	**0.691**	0.613	**0.777**	0.612	**0.665**
x_{26}	0.626	**0.738**	0.607	**0.704**	0.598	**0.729**	0.577	**0.636**
x_{27}	0.597	**0.734**	0.588	**0.678**	0.573	**0.729**	0.556	**0.615**
x_{28}	0.606	**0.733**	0.581	**0.677**	0.578	**0.723**	0.567	**0.604**
x_{29}	0.672	**0.788**	0.657	**0.681**	0.628	**0.776**	0.626	**0.642**
x_{30}	0.614	**0.746**	0.592	**0.694**	0.586	**0.739**	0.563	**0.633**
x_{31}	0.562	**0.695**	0.552	**0.674**	0.552	**0.692**	0.563	**0.61**
x_{32}	0.693	**0.839**	0.653	**0.779**	0.642	**0.822**	0.665	**0.67**
x_{33}	0.705	**0.799**	0.653	**0.719**	0.656	**0.788**	0.655	**0.674**
x_{34}	0.606	**0.741**	0.583	**0.683**	0.587	**0.737**	0.589	**0.636**
x_{35}	0.625	**0.755**	0.614	**0.699**	0.602	**0.753**	0.61	**0.637**
x_{36}	0.575	**0.699**	0.574	**0.663**	0.558	**0.691**	0.563	**0.597**
x_{37}	0.625	**0.781**	0.606	**0.648**	0.586	**0.771**	0.601	**0.623**
x_{38}	0.666	**0.782**	0.617	**0.737**	0.625	**0.773**	0.637	**0.656**
x_{39}	0.674	**0.791**	0.64	**0.723**	0.638	**0.774**	0.624	**0.675**
x_{40}	0.7	**0.818**	0.692	**0.758**	0.652	**0.807**	0.648	**0.672**
x_{41}	0.673	**0.775**	0.655	**0.742**	0.634	**0.766**	0.625	**0.67**
x_{42}	0.578	**0.704**	0.563	**0.67**	0.568	**0.705**	0.564	**0.624**
x_{43}	0.679	**0.807**	0.652	**0.705**	0.615	**0.787**	0.607	**0.629**
x_{44}	0.593	**0.714**	0.562	**0.689**	0.577	**0.714**	0.563	**0.628**
x_{45}	0.658	**0.767**	0.645	**0.714**	0.616	**0.759**	0.624	**0.655**
x_{46}	0.649	**0.774**	0.606	**0.693**	0.604	**0.767**	0.573	**0.628**
x_{47}	0.695	**0.82**	0.684	**0.764**	0.646	**0.813**	0.652	**0.691**
x_{48}	0.569	**0.691**	0.55	**0.649**	0.545	**0.685**	0.534	**0.573**
x_{49}	0.664	**0.797**	0.641	**0.721**	0.621	**0.787**	0.621	**0.649**

Table 14: Comparison of AUC for each feature before and after nFBST under eps=0.05 on Dataset 1.

Feature	Gradient	Grad-*n*FBST	DeepLIFT	DeepLIFT-*n*FBST	LRP	LRP-*n*FBST	LIME	LIME-*n*FBST
x_0	0.608	**0.704**	0.581	**0.675**	0.58	**0.7**	0.58	**0.611**
x_1	0.628	**0.711**	0.633	**0.669**	0.603	**0.709**	0.593	**0.613**
x_2	0.59	**0.677**	0.592	**0.645**	0.571	**0.678**	0.563	**0.57**
x_3	0.589	**0.7**	0.602	**0.658**	0.568	**0.697**	0.565	**0.602**
x_4	0.573	**0.632**	0.545	**0.59**	0.548	**0.63**	0.526	**0.554**
x_5	0.696	**0.761**	0.688	**0.71**	0.648	**0.752**	**0.628**	0.625
x_6	0.626	**0.714**	0.599	**0.671**	0.591	**0.711**	0.609	**0.625**
x_7	0.591	**0.666**	0.584	**0.638**	0.571	**0.666**	0.574	**0.599**
x_8	0.648	**0.709**	0.646	**0.659**	0.62	**0.701**	0.619	**0.623**
x_9	0.696	**0.779**	0.698	**0.713**	0.65	**0.776**	**0.636**	0.635
x_{10}	0.616	**0.698**	0.616	**0.648**	0.582	**0.694**	0.588	**0.6**
x_{11}	0.678	**0.765**	0.676	**0.727**	0.636	**0.759**	0.655	**0.656**
x_{12}	0.72	**0.777**	0.724	**0.731**	0.671	**0.774**	**0.662**	0.661
x_{13}	0.624	**0.712**	0.642	**0.65**	0.591	**0.707**	0.593	**0.599**
x_{14}	0.623	**0.72**	0.641	**0.684**	0.594	**0.717**	0.585	**0.615**
x_{15}	0.577	**0.647**	0.574	**0.638**	0.551	**0.649**	0.544	**0.577**
x_{16}	0.597	**0.716**	0.598	**0.659**	0.568	**0.714**	0.566	**0.597**
x_{17}	0.629	**0.721**	0.642	**0.681**	0.597	**0.722**	0.607	**0.618**
x_{18}	0.571	**0.641**	0.565	**0.61**	0.558	**0.64**	0.536	**0.563**
x_{19}	0.592	**0.675**	0.593	**0.645**	0.575	**0.678**	0.553	**0.594**
x_{20}	0.613	**0.699**	0.625	**0.664**	0.589	**0.699**	0.587	**0.609**
x_{21}	0.586	**0.675**	0.57	**0.594**	0.558	**0.673**	0.539	**0.571**
x_{22}	0.649	**0.712**	0.638	**0.66**	0.618	**0.71**	0.617	**0.629**
x_{23}	0.624	**0.734**	0.639	**0.693**	0.596	**0.731**	0.591	**0.624**
x_{24}	0.759	**0.805**	**0.746**	0.739	0.71	**0.795**	**0.682**	0.657
x_{25}	0.589	**0.668**	0.578	**0.616**	0.563	**0.663**	0.547	**0.578**
x_{26}	0.621	**0.677**	0.608	**0.661**	0.59	**0.676**	0.574	**0.599**
x_{27}	0.656	**0.742**	0.66	**0.703**	0.619	**0.738**	0.623	**0.623**
x_{28}	0.576	**0.741**	0.591	**0.677**	0.552	**0.736**	0.562	**0.621**
x_{29}	0.663	**0.72**	**0.628**	0.598	0.611	**0.712**	**0.595**	0.584
x_{30}	0.741	**0.823**	0.753	**0.763**	0.692	**0.825**	**0.684**	0.67
x_{31}	0.614	**0.708**	0.61	**0.668**	0.585	**0.704**	0.579	**0.612**
x_{32}	0.695	**0.771**	0.69	**0.726**	0.649	**0.769**	0.625	**0.626**
x_{33}	0.601	**0.71**	0.582	**0.593**	0.579	**0.711**	0.563	**0.612**
x_{34}	0.634	**0.745**	0.631	**0.678**	0.601	**0.739**	0.611	**0.625**
x_{35}	0.73	**0.786**	0.74	**0.753**	0.687	**0.786**	0.685	**0.688**
x_{36}	0.57	**0.63**	0.55	**0.598**	0.552	**0.634**	0.534	**0.583**
x_{37}	0.643	**0.747**	0.648	**0.697**	0.607	**0.741**	0.617	**0.637**
x_{38}	0.686	**0.788**	0.69	**0.73**	0.645	**0.785**	0.619	**0.646**
x_{39}	0.65	**0.723**	0.637	**0.68**	0.61	**0.719**	0.613	**0.619**
x_{40}	0.61	**0.696**	0.598	**0.642**	0.587	**0.694**	0.576	**0.6**
x_{41}	0.625	**0.686**	0.616	**0.641**	0.594	**0.683**	0.581	**0.601**
x_{42}	0.674	**0.762**	0.659	**0.725**	0.625	**0.76**	0.6	**0.609**
x_{43}	0.57	**0.637**	0.558	**0.624**	0.55	**0.641**	0.536	**0.586**
x_{44}	0.571	**0.658**	0.555	**0.603**	0.551	**0.657**	0.533	**0.569**
x_{45}	0.654	**0.737**	0.654	**0.656**	0.616	**0.728**	0.588	**0.622**
x_{46}	0.69	**0.812**	0.719	**0.75**	0.649	**0.804**	**0.637**	0.633
x_{47}	0.618	**0.693**	0.619	**0.634**	0.587	**0.691**	0.572	**0.595**
x_{48}	0.692	**0.783**	0.695	**0.736**	0.641	**0.779**	0.642	**0.658**
x_{49}	0.579	**0.677**	0.56	**0.633**	0.558	**0.675**	0.547	**0.594**

Table 15: Comparison of AUC for each feature before and after *n*FBST under eps=0.05 on Dataset 2.