

## A APPENDIX

To display the generalised performance of the proposed methods, we validate the methods with the same experimental conditions as those shown in the draft, but with different parameter settings. Each setting contains 20 generated datasets, which are used to train RF, SVM and NN separately.

Table 1: Comparison of Kendall’s Tau and Spearman’s Coefficient for different models, correlation coefficients, sample sizes, and methods. All the results are shown based on models trained on linear synthetic data.

Model	CorrCoef	SampleSize	Method	KendallTau	KendallVar	SpearmanCoef	SpearmanVar
RF	0.2	200	quantile-based	0.7746	1.23E-32	0.866	1.23E-32
RF	0.5	200	quantile-based	0.7746	4.93E-32	0.866	1.23E-32
RF	0.8	200	quantile-based	0.7746	4.93E-32	0.866	1.23E-32
RF	0.2	2000	quantile-based	0.7746	4.93E-32	0.866	1.23E-32
RF	0.5	2000	quantile-based	0.7746	4.93E-32	0.866	1.23E-32
RF	0.8	2000	quantile-based	0.7746	0	0.866	1.23E-32
RF	0.2	20000	quantile-based	0.7746	1.23E-32	0.866	1.23E-32
RF	0.5	20000	quantile-based	0.7746	0	0.866	1.23E-32
RF	0.8	20000	quantile-based	0.7746	1.23E-32	0.866	1.23E-32
RF	0.2	200	kernel-based	0.7746	1.23E-32	0.866	1.23E-32
RF	0.5	200	kernel-based	0.7746	4.93E-32	0.866	1.23E-32
RF	0.8	200	kernel-based	0.7746	4.93E-32	0.866	1.23E-32
RF	0.2	2000	kernel-based	0.7746	4.93E-32	0.866	1.23E-32
RF	0.5	2000	kernel-based	0.7746	4.93E-32	0.866	1.23E-32
RF	0.8	2000	kernel-based	0.7746	0	0.866	1.23E-32
RF	0.2	20000	kernel-based	0.7746	1.23E-32	0.866	1.23E-32
RF	0.5	20000	kernel-based	0.7746	0	0.866	1.23E-32
RF	0.8	20000	kernel-based	0.7746	1.23E-32	0.866	1.23E-32
SVM	0.2	200	quantile-based	0.7746	4.93E-32	0.866	1.23E-32
SVM	0.5	200	quantile-based	0.7746	1.23E-32	0.866	1.23E-32
SVM	0.8	200	quantile-based	0.7746	1.23E-32	0.866	1.23E-32
SVM	0.2	2000	quantile-based	0.7746	1.23E-32	0.866	1.23E-32
SVM	0.5	2000	quantile-based	0.7746	4.93E-32	0.866	1.23E-32
SVM	0.8	2000	quantile-based	0.7746	1.23E-32	0.866	1.11E-31
SVM	0.2	20000	quantile-based	0.7746	0	0.866	0
SVM	0.5	20000	quantile-based	0.7746	4.93E-32	0.866	4.93E-32
SVM	0.8	20000	quantile-based	0.7746	0	0.866	0
SVM	0.2	200	kernel-based	0.7746	4.93E-32	0.866	1.23E-32
SVM	0.5	200	kernel-based	0.7746	1.23E-32	0.866	1.23E-32
SVM	0.8	200	kernel-based	0.7746	1.23E-32	0.866	1.23E-32
SVM	0.2	2000	kernel-based	0.7746	1.23E-32	0.866	1.23E-32
SVM	0.5	2000	kernel-based	0.7746	4.93E-32	0.866	1.23E-32
SVM	0.8	2000	kernel-based	0.7746	1.23E-32	0.866	1.11E-31
SVM	0.2	20000	kernel-based	0.7746	0	0.866	0
SVM	0.5	20000	kernel-based	0.7746	4.93E-32	0.866	4.93E-32
SVM	0.8	20000	kernel-based	0.7746	0	0.866	0
NN	0.2	200	quantile-based	0.7746	1.23E-32	0.866	1.23E-32
NN	0.5	200	quantile-based	0.7746	1.23E-32	0.866	1.23E-32
NN	0.8	200	quantile-based	0.7746	1.23E-32	0.866	1.23E-32
NN	0.2	2000	quantile-based	0.7746	1.23E-32	0.866	1.23E-32
NN	0.5	2000	quantile-based	0.7746	1.23E-32	0.866	1.23E-32
NN	0.8	2000	quantile-based	0.7746	1.23E-32	0.866	1.23E-32
NN	0.2	20000	quantile-based	0.7746	4.93E-32	0.866	1.23E-32
NN	0.5	20000	quantile-based	0.7746	4.93E-32	0.866	1.23E-32
NN	0.8	20000	quantile-based	0.7746	4.93E-32	0.866	1.23E-32
NN	0.2	200	kernel-based	0.7746	1.23E-32	0.866	1.23E-32
NN	0.5	200	kernel-based	0.7746	1.23E-32	0.866	1.23E-32
NN	0.8	200	kernel-based	0.7746	1.23E-32	0.866	1.23E-32
NN	0.2	2000	kernel-based	0.7746	1.23E-32	0.866	1.23E-32
NN	0.5	2000	kernel-based	0.7746	1.23E-32	0.866	1.23E-32
NN	0.8	2000	kernel-based	0.7746	1.23E-32	0.866	1.23E-32
NN	0.2	20000	kernel-based	0.7746	4.93E-32	0.866	1.23E-32
NN	0.5	20000	kernel-based	0.7746	4.93E-32	0.866	1.23E-32
NN	0.8	20000	kernel-based	0.7746	4.93E-32	0.866	1.23E-32

---

054 The CorrCoef in the table represents the correlation level. For linear cases, this is the correlation  
055 score between the  $x_1$  and  $x_2, x_3$ . For nonlinear cases, this coefficient represents the level of how  
056 much the distractor will influence feature  $x_1$ , i.e., the weights of  $x_1$  to generate the distractors. We  
057 use Kendall’s tau and Spearman correlation scores with associated variance to display the perfor-  
058 mance. The two correlation scores are calculated between the explanation results and the feature  
059 weights used for generating the class-related information.

060 Table 1 demonstrates the results of the proposed methods applied to models trained on linear syn-  
061 thetic data. The proposed methods are not influenced much under these simple scenarios.

062  
063 Table 2 displays the results of the proposed methods applied to models trained on nonlinear syn-  
064 thetic data. The quantile-based method is significantly influenced under the small sample amount  
065 condition, while the kernel-based method is slightly influenced. However, as the number of samples  
066 increased, the computation costs of the kernel-based method significantly increased. The computa-  
067 tion cost of the quantile-based method increases slightly.

068  
069  
070  
071  
072  
073  
074  
075  
076  
077  
078  
079  
080  
081  
082  
083  
084  
085  
086  
087  
088  
089  
090  
091  
092  
093  
094  
095  
096  
097  
098  
099  
100  
101  
102  
103  
104  
105  
106  
107

Table 2: Comparison of Kendall’s Tau and Spearman’s Coefficient for different models, correlation coefficients, sample sizes, and methods. All the results are shown based on models trained on nonlinear synthetic data.

Model	CorrCoef	SampleSize	Method	KendallTau	KendallVar	SpearmanCoef	SpearmanVar
RF	0.4	200	quantile-based	-0.1291	0.446	-0.1443	0.4986
RF	0.7	200	quantile-based	-0.0422	0.5199	-0.0471	0.5813
RF	1	200	quantile-based	-0.0086	0.4396	-0.0096	0.4915
RF	0.4	2000	quantile-based	0.7746	2.22E-16	0.866	1.11E-16
RF	0.7	2000	quantile-based	0.7746	2.22E-16	0.866	1.11E-16
RF	1	2000	quantile-based	0.7746	0	0.866	0
RF	0.4	20000	quantile-based	0.7746	1.11E-16	0.866	2.22E-16
RF	0.7	20000	quantile-based	0.7746	0	0.866	0
RF	1	20000	quantile-based	0.7746	0	0.866	0
RF	0.4	200	kernel-based	0.7746	2.22E-16	0.866	1.11E-16
RF	0.7	200	kernel-based	0.7641	0.0511	0.8542	0.0571
RF	1	200	kernel-based	0.6541	0.208	0.7313	0.2325
RF	0.4	2000	kernel-based	0.7746	2.22E-16	0.866	1.11E-16
RF	0.7	2000	kernel-based	0.7746	0	0.866	0
RF	1	2000	kernel-based	0.7746	0	0.866	0
RF	0.4	20000	kernel-based	0.7746	1.11E-16	0.866	2.22E-16
RF	0.7	20000	kernel-based	0.7746	0	0.866	0
RF	1	20000	kernel-based	0.7746	0	0.866	0
SVM	0.4	200	quantile-based	0.0258	0.4019	0.0289	0.4494
SVM	0.7	200	quantile-based	0.1291	0.446	0.1443	0.4986
SVM	1	200	quantile-based	0.0689	0.4663	0.077	0.5214
SVM	0.4	2000	quantile-based	0.7746	0	0.866	1.11E-16
SVM	0.7	2000	quantile-based	0.7746	1.11E-16	0.866	0
SVM	1	2000	quantile-based	0.7746	0	0.866	0
SVM	0.4	20000	quantile-based	0.7746	2.22E-16	0.866	1.11E-16
SVM	0.7	20000	quantile-based	0.7746	2.22E-16	0.866	1.11E-16
SVM	1	20000	quantile-based	0.7746	2.22E-16	0.866	1.11E-16
SVM	0.4	200	kernel-based	0.766	0.0463	0.8564	0.0518
SVM	0.7	200	kernel-based	0.7746	2.22E-16	0.866	1.11E-16
SVM	1	200	kernel-based	0.7488	0.0775	0.8372	0.0866
SVM	0.4	2000	kernel-based	0.7746	0	0.866	1.11E-16
SVM	0.7	2000	kernel-based	0.7746	1.11E-16	0.866	0
SVM	1	2000	kernel-based	0.7746	0	0.866	0
SVM	0.4	20000	kernel-based	0.7746	2.22E-16	0.866	1.11E-16
SVM	0.7	20000	kernel-based	0.766	0.0463	0.8564	0.0518
SVM	1	20000	kernel-based	0.7746	2.22E-16	0.866	1.11E-16
NN	0.4	200	quantile-based	-0.043	0.566	-0.0481	0.6328
NN	0.7	200	quantile-based	0.1033	0.4195	0.1155	0.469
NN	1	200	quantile-based	0.0258	0.4636	0.0289	0.5183
NN	0.4	2000	quantile-based	0.7746	2.22E-16	0.866	1.11E-16
NN	0.7	2000	quantile-based	0.7746	1.11E-16	0.866	1.11E-16
NN	1	2000	quantile-based	0.7746	1.11E-16	0.866	1.11E-16
NN	0.4	20000	quantile-based	0.7746	2.22E-16	0.866	1.11E-16
NN	0.7	20000	quantile-based	0.7746	2.22E-16	0.866	1.11E-16
NN	1	20000	quantile-based	0.7746	2.22E-16	0.866	1.11E-16
NN	0.4	200	kernel-based	0.6197	0.2171	0.6928	0.2427
NN	0.7	200	kernel-based	0.6197	0.2547	0.6928	0.2848
NN	1	200	kernel-based	0.5594	0.2988	0.6255	0.334
NN	0.4	2000	kernel-based	0.7746	2.22E-16	0.866	1.11E-16
NN	0.7	2000	kernel-based	0.7746	1.11E-16	0.866	1.11E-16
NN	1	2000	kernel-based	0.7316	0.1644	0.8179	0.1838
NN	0.4	20000	kernel-based	0.7746	2.22E-16	0.866	1.11E-16
NN	0.7	20000	kernel-based	0.7574	0.0927	0.8468	0.1036
NN	1	20000	kernel-based	0.7402	0.1288	0.8275	0.144