w/ Full:

Integrated Gradient

mean precision causal@5 0.17788461538461525
mean recall causal@5 0.25987293956043955
mean precision attrb@5 0.2038461538461539
mean recall attrb@5 0.2927712912087913
MAP@5 Causal 0.16939102564102562
MAP@5 Attrib 0.1971794871794872

Saliency

mean precision causal@5 0.17788461538461525
mean recall causal@5 0.25987293956043955
mean precision attrb@5 0.18846153846153843
mean recall attrb@5 0.26969436813186815
MAP@5 Causal 0.16939102564102562
MAP@5 Attrib 0.18072115384615384

DeepLift

mean precision causal@5 0.17788461538461525
mean recall causal@5 0.25987293956043955
mean precision attrb@5 0.20865384615384633
mean recall attrb@5 0.300062957875458
MAP@5 Causal 0.16939102564102562
MAP@5 Attrib 0.20935897435897435

DeepLiftSHAP

mean precision causal@5 0.17788461538461525
mean recall causal@5 0.25987293956043955
mean precision attrb@5 0.21442307692307708
mean recall attrb@5 0.3118189102564104
MAP@5 Causal 0.16939102564102562
MAP@5 Attrib 0.22076923076923072

GradSHAP

mean precision causal@5 0.17788461538461525

mean recall causal@5 0.25987293956043955
mean precision attrb@5 0.20480769230769236
mean recall attrb@5 0.30004578754578753
MAP@5 Causal 0.16939102564102562
MAP@5 Attrib 0.1956891025641025

Input x Grad

mean precision causal@5 0.17788461538461525
mean recall causal@5 0.25987293956043955
mean precision attrb@5 0.21057692307692327
mean recall attrb@5 0.3032680860805862
MAP@5 Causal 0.16939102564102562
MAP@5 Attrib 0.2075

w/o DR:

Integrated Grad.

mean precision causal@5 0.13942307692307662
mean recall causal@5 0.20763507326007324
mean precision attrb@5 0.17403846153846145
mean recall attrb@5 0.2590430402930403
MAP@5 Causal 0.16249999999999995
MAP@5 Attrib 0.18873397435897438

0it [00:00, ?it/s]
<captum.attr._core.saliency.Saliency object at 0x790550c02140>
660it [08:29, 1.29it/s]
mean precision causal@5 0.13942307692307662
mean recall causal@5 0.20763507326007324
mean precision attrb@5 0.12499999999999975
mean recall attrb@5 0.189320054945055
MAP@5 Causal 0.16249999999999995
MAP@5 Attrib 0.16442307692307687
0it [00:00, ?it/s]
<captum.attr._core.deep_lift.DeepLift object at 0x79055165ad70>
660it [08:26, 1.30it/s]
mean precision causal@5 0.13942307692307662
mean recall causal@5 0.20763507326007324
mean precision attrb@5 0.17788461538461528

mean recall attrb@5 0.26864697802197807
MAP@5 Causal 0.16249999999999995
MAP@5 Attrib 0.17187500000000003
0it [00:00, ?it/s]
<captum.attr._core.deep_lift.DeepLiftShap object at 0x790551edae30>
660it [08:48, 1.25it/s]
mean precision causal@5 0.13942307692307662
mean recall causal@5 0.20763507326007324
mean precision attrb@5 0.21730769230769245
mean recall attrb@5 0.31462339743589757
MAP@5 Causal 0.16249999999999995
MAP@5 Attrib 0.22110576923076922
0it [00:00, ?it/s]
<captum.attr._core.gradient_shap.GradientShap object at 0x790551642bc0>
660it [08:16, 1.33it/s]
mean precision causal@5 0.13942307692307662
mean recall causal@5 0.20763507326007324
mean precision attrb@5 0.20673076923076936
mean recall attrb@5 0.3025469322344326
MAP@5 Causal 0.16249999999999995
MAP@5 Attrib 0.20929487179487177
0it [00:00, ?it/s]
<captum.attr._core.input_x_gradient.InputXGradient object at 0x79055156b520>
660it [08:19, 1.32it/s]
mean precision causal@5 0.13942307692307662
mean recall causal@5 0.20763507326007324
mean precision attrb@5 0.17403846153846142
mean recall attrb@5 0.26464056776556777
MAP@5 Causal 0.16249999999999995
MAP@5 Attrib 0.1685416666666667

w/o inlp

0it [00:00, ?it/s]
<captum.attr._core.integrated_gradients.IntegratedGradients object at 0x7c24f42dcca0>

660it [02:46, 3.96it/s]

mean precision causal@5 0.20480769230769236
mean recall causal@5 0.30384043040293046
mean precision attrb@5 0.18461538461538451
mean recall attrb@5 0.2655219780219781
MAP@5 Causal 0.18879807692307685
MAP@5 Attrib 0.1676121794871795

0it [00:00, ?it/s]
<captum.attr._core.saliency.Saliency object at 0x7c2506da5750>

660it [01:49, 6.04it/s]

mean precision causal@5 0.20480769230769236
mean recall causal@5 0.30384043040293046
mean precision attrb@5 0.20961538461538465
mean recall attrb@5 0.31785714285714295
MAP@5 Causal 0.18879807692307685
MAP@5 Attrib 0.2485737179487179

0it [00:00, ?it/s]
<captum.attr._core.deep_lift.DeepLift object at 0x7c24f42dff40>

660it [01:51, 5.91it/s]

mean precision causal@5 0.20480769230769236
mean recall causal@5 0.30384043040293046
mean precision attrb@5 0.1846153846153846
mean recall attrb@5 0.2638564560439562
MAP@5 Causal 0.18879807692307685
MAP@5 Attrib 0.17826923076923062

0it [00:00, ?it/s]
<captum.attr._core.deep_lift.DeepLiftShap object at 0x7c24b966b4f0>

660it [02:33, 4.30it/s]

mean precision causal@5 0.20480769230769236
mean recall causal@5 0.30384043040293046
mean precision attrb@5 0.19423076923076918
mean recall attrb@5 0.2869333791208793
MAP@5 Causal 0.18879807692307685
MAP@5 Attrib 0.19822115384615382

0it [00:00, ?it/s]
<captum.attr._core.gradient_shap.GradientShap object at 0x7c24b9696d70>

660it [01:53, 5.82it/s]

mean precision causal@5 0.20480769230769236
mean recall causal@5 0.30384043040293046
mean precision attrb@5 0.20096153846153852
mean recall attrb@5 0.2949061355311357
MAP@5 Causal 0.18879807692307685
MAP@5 Attrib 0.20149038461538457

0it [00:00, ?it/s]
<captum.attr._core.input_x_gradient.InputXGradient object at 0x7c2506da5750>

660it [01:51, 5.91it/s]
mean precision causal@5 0.20480769230769236
mean recall causal@5 0.30384043040293046
mean precision attrb@5 0.1807692307692307
mean recall attrb@5 0.25773237179487196
MAP@5 Causal 0.18879807692307685
MAP@5 Attrib 0.16708333333333322

w/o adversarial

0it [00:00, ?it/s]
<captum.attr._core.integrated_gradients.IntegratedGradients object at 0x7c24b8d9eb90>

660it [09:31, 1.15it/s]

mean precision causal@5 0.14423076923076897
mean recall causal@5 0.20395489926739932
mean precision attrb@5 0.19230769230769232
mean recall attrb@5 0.27670558608058615
MAP@5 Causal 0.13333333333333325
MAP@5 Attrib 0.174198717948718

0it [00:00, ?it/s]
<captum.attr._core.saliency.Saliency object at 0x7c24b8dc8be0>

660it [08:15, 1.33it/s]

mean precision causal@5 0.14423076923076897
mean recall causal@5 0.20395489926739932
mean precision attrb@5 0.1913461538461538
mean recall attrb@5 0.292651098901099
MAP@5 Causal 0.13333333333333325
MAP@5 Attrib 0.1554967948717947

0it [00:00, ?it/s]
<captum.attr._core.deep_lift.DeepLift object at 0x7c24b8d8eec0>

660it [08:18, 1.32it/s]

mean precision causal@5 0.14423076923076897
mean recall causal@5 0.20395489926739932
mean precision attrb@5 0.1788461538461537
mean recall attrb@5 0.2596382783882785
MAP@5 Causal 0.13333333333333325
MAP@5 Attrib 0.1565224358974358

0it [00:00, ?it/s]
<captum.attr._core.deep_lift.DeepLiftShap object at 0x7c24b8dcc7f0>

660it [08:55, 1.23it/s]

mean precision causal@5 0.14423076923076897
mean recall causal@5 0.20395489926739932
mean precision attrb@5 0.2028846153846156
mean recall attrb@5 0.296537316849817
MAP@5 Causal 0.13333333333333325
MAP@5 Attrib 0.20371794871794868

0it [00:00, ?it/s]
<captum.attr._core.gradient_shap.GradientShap object at 0x7c24b8d9ebc0>

660it [08:22, 1.31it/s]

mean precision causal@5 0.14423076923076897
mean recall causal@5 0.20395489926739932
mean precision attrb@5 0.19807692307692326
mean recall attrb@5 0.2942708333333336
MAP@5 Causal 0.13333333333333325
MAP@5 Attrib 0.2006730769230768

0it [00:00, ?it/s]
<captum.attr._core.input_x_gradient.InputXGradient object at 0x7c24b8dcc7f0>

660it [08:08, 1.35it/s]
mean precision causal@5 0.14423076923076897
mean recall causal@5 0.20395489926739932
mean precision attrb@5 0.19230769230769237
mean recall attrb@5 0.2730425824175826
MAP@5 Causal 0.13333333333333325
MAP@5 Attrib 0.17339743589743592

Integrated Gradient
Causal keywords

LAS mean -0.007740368789507959 LAS std 0.040730586945622295
LUS mean -0.007575757575757574 LUS std 0.03664662612862977
Comprehensiveness mean 0.0021570572373232405 Comprehensiveness std
0.01911704466344151
Sufficiency mean 0.11260833004296253 Sufficiency std 0.014726657356743307

Avg. f1 between prediction of proposed model and SVM w/ both inp and exp 0.619055991552799
Avg. f1 between prediction of proposed model and SVM w/ only inp 0.6179316654395448
Avg. f1 between prediction of proposed model and SVM w/ only exp 0.46503907203604095

Gradient attribution keywords

LAS mean 0.01838545996470668 LAS std 0.052393658931548646
LUS mean 0.022727272727272728 LUS std 0.0333677508265837
Comprehensiveness mean 0.030601415597236316 Comprehensiveness std 0.014318765420616105
Sufficiency mean 0.08881835726537254 Sufficiency std 0.012162940909293752
Avg. f1 between prediction of proposed model and SVM w/ both inp and exp 0.6595688017858778
Avg. f1 between prediction of proposed model and SVM w/ only inp 0.6179316654395448
Avg. f1 between prediction of proposed model and SVM w/ only exp 0.5291117445050798

harmonic mean keywords

LAS mean 0.013935767307983637 LAS std 0.04035713182273682
LUS mean 0.016666666666666663 LUS std 0.03281273913289048
Comprehensiveness mean 0.0195572432474035 Comprehensiveness std 0.015063164844456918
Sufficiency mean 0.10948298013929877 Sufficiency std 0.01141138863155644
Avg. f1 between prediction of proposed model and SVM w/ both inp and exp 0.6490150557630194
Avg. f1 between prediction of proposed model and SVM w/ only inp 0.6179316654395448
Avg. f1 between prediction of proposed model and SVM w/ only exp 0.5036636743925325

Kendal's tau, spearman's rho causal vs attribution (0.016815210932857993, 0.02454264002251618)

Kendal's tau, spearman's rho causal vs harmonic

(0.06444840562487622, 0.08952372017387498)

Rank Average ICaCE (Causal)
Rank Average ICaCE (Attribution)

(0.00030524816103865256, 0.00030499018212240845)

Saliency

Causal Keywords

LAS mean -0.007740368789507959 LAS std 0.040730586945622295
LUS mean -0.007575757575757574 LUS std 0.03664662612862977
Comprehensiveness mean 0.0021570572373232405 Comprehensiveness std 0.01911704466344151
Sufficiency mean 0.11260833004296253 Sufficiency std 0.014726657356743307
Avg. f1 between prediction of proposed model and SVM w/ both inp and exp 0.619055991552799
Avg. f1 between prediction of proposed model and SVM w/ only inp 0.6179316654395448
Avg. f1 between prediction of proposed model and SVM w/ only exp 0.46503907203604095

Gradient Attribution Keywords

LAS mean 0.000309038155378092 LAS std 0.03564313340188133
LUS mean 0.0 LUS std 0.03455077045754964
Comprehensiveness mean 0.005284808354239437 Comprehensiveness std 0.017499305566834655
Sufficiency mean 0.11622874918318746 Sufficiency std 0.012123685145103901
Avg. f1 between prediction of proposed model and SVM w/ both inp and exp 0.628038037260145
Avg. f1 between prediction of proposed model and SVM w/ only inp 0.6179316654395448
Avg. f1 between prediction of proposed model and SVM w/ only exp 0.4204876550528551

Harmonic Mean

LAS mean -0.010973165010713146 LAS std 0.044431742659039014
LUS mean -0.012121212121212123 LUS std 0.04380858271151806
Comprehensiveness mean 0.0031456734591899967 Comprehensiveness std 0.026506934552299993
Sufficiency mean 0.11368759974502451 Sufficiency std 0.011237557697291322
Avg. f1 between prediction of proposed model and SVM w/ both inp and exp 0.6129739958312254
Avg. f1 between prediction of proposed model and SVM w/ only inp 0.6179316654395448
Avg. f1 between prediction of proposed model and SVM w/ only exp 0.4383801005844089

Kendal's tau, spearman's rho causal vs attribution

(0.8599722717369777, 0.9166088125840447)

Kendal's tau, spearman's rho causal vs harmonic

(0.9122994652406419, 0.9615035806986271)

ICaCE Causal, Attribution

(0.00030524816103865256, 0.00030524667655512893)

DeepLift

Causal

LAS mean -0.007740368789507959 LAS std 0.040730586945622295
LUS mean -0.007575757575757574 LUS std 0.03664662612862977
Comprehensiveness mean 0.0021570572373232405 Comprehensiveness std 0.01911704466344151
Sufficiency mean 0.11260833004296253 Sufficiency std 0.014726657356743307
Avg. f1 between prediction of proposed model and SVM w/ both inp and exp 0.619055991552799
Avg. f1 between prediction of proposed model and SVM w/ only inp 0.6179316654395448
Avg. f1 between prediction of proposed model and SVM w/ only exp 0.46503907203604095

Attribution

LAS mean -0.020551875787443157 LAS std 0.05323616604848854
LUS mean 0.019696969696969695 LUS std 0.04068400479423501
Comprehensiveness mean 0.029394959953815136 Comprehensiveness std 0.01247535682177633
Sufficiency mean 0.02148757631344868 Sufficiency std 0.018994372247116042
Avg. f1 between prediction of proposed model and SVM w/ both inp and exp 0.6728950618947075
Avg. f1 between prediction of proposed model and SVM w/ only inp 0.6179316654395448
Avg. f1 between prediction of proposed model and SVM w/ only exp 0.6519413974687287

Harmonic

LAS mean -0.0054609720699874175 LAS std 0.07263055038339786
LUS mean 0.018181818181818 LUS std 0.046847347981031226
Comprehensiveness mean 0.04348279362205159 Comprehensiveness std 0.015910781955207308
Sufficiency mean 0.048366403066181646 Sufficiency std 0.009917973248886942
Avg. f1 between prediction of proposed model and SVM w/ both inp and exp 0.6777453071303629
Avg. f1 between prediction of proposed model and SVM w/ only inp 0.6179316654395448
Avg. f1 between prediction of proposed model and SVM w/ only exp 0.6408922006787731

Kendal's tau, spearman's rho causal vs attribution

(-0.0006932065755595178, -0.002395471745316948)

Kendal's tau, spearman's rho causal vs harmonic

(0.035333729451376514, 0.04721205866716703)

ICaCE Causal, Attribution

(0.00030524816103865256, 0.0003049828811120403)

GradientSHAP

Causal

LAS mean -0.007740368789507959 LAS std 0.040730586945622295
LUS mean -0.007575757575757574 LUS std 0.03664662612862977
Comprehensiveness mean 0.0021570572373232405 Comprehensiveness std 0.01911704466344151
Sufficiency mean 0.11260833004296253 Sufficiency std 0.014726657356743307
Avg. f1 between prediction of proposed model and SVM w/ both inp and exp 0.619055991552799
Avg. f1 between prediction of proposed model and SVM w/ only inp 0.6179316654395448
Avg. f1 between prediction of proposed model and SVM w/ only exp 0.46503907203604095

Attribution

LAS mean 0.005957313946926161 LAS std 0.02964593417516629
LUS mean 0.0030303030303030303 LUS std 0.03015113445777636
Comprehensiveness mean 0.008467676587089263 Comprehensiveness std 0.01494089460755256
Sufficiency mean 0.11673798407101268 Sufficiency std 0.014282244265603563
Avg. f1 between prediction of proposed model and SVM w/ both inp and exp 0.6312806393054475
Avg. f1 between prediction of proposed model and SVM w/ only inp 0.6179316654395448
Avg. f1 between prediction of proposed model and SVM w/ only exp 0.44823907730849033

Harmonic

LAS mean -0.010743168614197314 LAS std 0.017920414345334413
LUS mean 0.010606060606060607 LUS std 0.0077257871418072496
Comprehensiveness mean 0.022981157307047622 Comprehensiveness std 0.007459058049434636
Sufficiency mean 0.054176177783388034 Sufficiency std 0.010506453579822694
Avg. f1 between prediction of proposed model and SVM w/ both inp and exp 0.6553934705977492

Avg. f1 between prediction of proposed model and SVM w/ only inp 0.639205895752934
Avg. f1 between prediction of proposed model and SVM w/ only exp 0.5996812426206338

Kendal's tau, spearman's rho causal vs attribution

(0.009982174688057044, 0.014141414141414142)

Kendal's tau, spearman's rho causal vs harmonic

(0.015329768270944746, 0.01989867717421897)

ICaCE Causal, Attribution
(0.0003052481610386525 6, 0.000304985446867528)

DeepLiftSHAP

Causal
LAS mean -0.007740368789507959 LAS std 0.040730586945622295
LUS mean -0.007575757575757574 LUS std 0.03664662612862977
Comprehensiveness mean 0.0021570572373232405 Comprehensiveness std 0.01911704466344151
Sufficiency mean 0.11260833004296253 Sufficiency std 0.014726657356743307
Avg. f1 between prediction of proposed model and SVM w/ both inp and exp 0.619055991552799
Avg. f1 between prediction of proposed model and SVM w/ only inp 0.6179316654395448
Avg. f1 between prediction of proposed model and SVM w/ only exp 0.46503907203604095

Attribution

LAS mean -0.021698693936427295 LAS std 0.04481026286959445
LUS mean 0.022727272727272728 LUS std 0.036014740375771855
Comprehensiveness mean 0.034802940246079 Comprehensiveness std 0.014216986896139625
Sufficiency mean 0.016643930052204197 Sufficiency std 0.015455203183896777
Avg. f1 between prediction of proposed model and SVM w/ both inp and exp 0.670279266497048
Avg. f1 between prediction of proposed model and SVM w/ only inp 0.6179316654395448
Avg. f1 between prediction of proposed model and SVM w/ only exp 0.6798093154448972

Harmonic
LAS mean -0.021946553528572527 LAS std 0.06071853530588233
LUS mean 0.004545454545454545 LUS std 0.04234299579540037
Comprehensiveness mean 0.03624434717821472 Comprehensiveness std 0.01688049458603115
Sufficiency mean 0.06480191344206476 Sufficiency std 0.018305069012365664
Avg. f1 between prediction of proposed model and SVM w/ both inp and exp 0.6539491931817936
Avg. f1 between prediction of proposed model and SVM w/ only inp 0.6179316654395448
Avg. f1 between prediction of proposed model and SVM w/ only exp 0.621235363051446

Causal vs attribution
(-0.003664091899386016, -0.0053601025737248665)

Causal vs harmonic
(0.012576747870865521, 0.017162335428589293)

ICaCE Causal vs attribution
(0.00030524816103865256, 0.00030497682513980245)