The effect of whitening on explanation performance

Anonymous Author
Anonymous Institution

Abstract

Explainable artificial intelligence (XAI) promises to provide information about models, their training data, and given test inputs to users of machine learning systems. As many XAI method are algorithmically defined, the ability of these method to provide correct answers to relevant questions needs to be theoretically verified and/or empirically validated. Prior work (Haufe et al., 2014; Wilming et al., 2023) has pointed out that popular feature attribution methods tend to assign significant importance to input features lacking a statistical association with the prediction target, leading to misinterpretations. This phenomenon is caused by the presence of dependent noises and is absent when all features are mutually independent. This motivates the question whether whitening, a common preprocessing effectively decorrelating the data before training, can avoid such misinterpretations. Using an established benchmark (Clark et al., 2024b) comprising ground truth-based definitions of explanation correctness and quantitative metrics of explanation performance, we evaluate 16 popular feature attribution methods in combination with 5 different whitening transforms, and compare their performance to baselines. The results show that whitening's impact on XAI performance is multifaceted, with some whitening techniques showing marked improvement in performance, though the degree of this improvement varies by XAI method and model architecture. The variability revealed in the experiments can be explained by the complexity of the relationship between the quality of pre-processing and the subsequent effectiveness of XAI methods, which underlines the significance of pre-processing techniques for model interpretability.

Preliminary work. Under review by AISTATS 2025. Do not distribute.

1 Introduction

In recent years, there has been a growing focus on empirically validating the performance of so-called explainable artificial intelligence (XAI) methods by examining the accuracy of their explanations (such as, Tjoa & Guan, 2020; Li et al., 2021; Zhou et al., 2022; Arras et al., 2022; Gevaert et al., 2022; Agarwal et al., 2022; Oliveira et al., 2024; Wilming et al., 2024). While some such studies use ground truth explanations, they often face limitations in their objective assessment of explanation correctness, the variety of XAI methods analyzed, and the complexity of the explanation tasks. Many existing ground truth problems are designed in a way that avoids realistic correlations between class-related and class-unrelated features (such as image foreground versus background). In real-world scenarios, however, such dependencies can introduce suppressor variables, noisy features that are not directly associated with the prediction target but can be utilized by the model (for example, for denoising, e.g., Haufe et al., 2014). For instance, in image data, background elements representing lighting conditions could act as suppressor variables. A model may leverage this information to adjust for lighting variations, thereby enhancing object detection. More comprehensive discussions on suppressor variables are available in Conger (1974); Friedman & Wall (2005); Haufe et al. (2014); Wilming et al. (2023).

A common XAI paradigm is to assign an 'importance' score to each feature of a given input. It has been shown, though, empirically and theoretically, that various popular feature attribution methods tend to systematically assign importance to suppressor variables in linear settings (Wilming et al., 2022, 2023). Extending this result, the work of Clark et al. (2024b) introduces the XAI-TRIS datasets, composed of four binary image classification problems, one linear and three non-linear. In each dataset, different types and combinations of tetrominoes (Golomb, 1996), geometric shapes consist-

ing of four blocks, need to be distinguished from one another. These tetromino images are overlaid on different types of noisy backgrounds: white noise (WHITE) and correlated (CORR) background; the latter induces a suppression effect through Gaussian smoothing.

The tetrominoes then represent discriminative features serving as ground truth explanations. Clark et al. (2024b) show that contemporary XAI methods fail to highlight tetrominoes consistently and, in some cases, are outperformed by model-ignorant edge detectors.

It is assumed that the suppression effect degrades explanation performance, and one potential approach to reduce this impact is to use data whitening techniques. Whitenings are multivariate linear transformations that transform the original features into a new space in which all features are uncorrelated and have unit variance, thus reducing feature redundancy. Notably, some whitening transformation maintain a 1:1 correspondence between original and transformed features, making it possible to visualize importance attributions in input space and assessing their efficacy as explanations.

In this paper, we take the XAI-TRIS datasets and the associated experimental pipeline proposed by Clark et al. (2024b) to assess whether the use of whitening techniques can improve the performance of XAI methods with respect to correctness of the explanations produced. The data scenario where the WHITE background type gets utilized serves as a baseline due to having no correlations between features of the background, hence we are not applying whitening methods. Then we test if applying whitening methods to the CORR background type can reduce the impact of suppressor variables. Here, we expect explanations to be more aligned with discriminative features, and hence to see improved explanation performance.

2 Methods

Our general workflow of applying and benchmarking post-hoc XAI methods follows previous work (Wilming et al., 2022; Clark et al., 2024b,a). We take a dataset generated with explicitly known class-related features defining the classification task and the ground truth for explanations, and train a machine learning model. The trained model is then applied to test inputs, for which output explanations are computed by XAI methods. We exclusively consider feature attribution methods, which assign an 'importance' score to each feature of the input. We then apply two performance metrics to compare produced explanations and the ground truth explanation for the given sample, giving us measures of the explanation performance of each method. Below, we highlight each of these steps, with more depth and the exact parameterizations given in the supplementary materials.

All code to generate data and perform the resulting analysis is publicly available and provided (anonymized) for this submission[1], under a GNU General Public License. Consent from the authors of XAI-TRIS (Clark et al., 2024b) to use their code in this manner was obtained prior to experimentation. The full analysis for this paper was done using the free infrastructure provided on Google Colaboratory[2], including the use of an 8-core TPU for model training.

2.1 Data Generation

We utilize the datasets supplied by the XAI-TRIS suite (Clark et al., 2024b), providing four binary image classification problems. We make use of the 8×8-px variant with two background types – the uncorrelated (WHITE) and correlated (CORR) backgrounds. The CORR background type takes the WHITE background and smoothes it with a Gaussian filter, inducing correlations between features. This also induces a suppression effect where background pixels overlapping with the placed tetromino are correlated to nearby background pixels. The four classification scenarios are defined as:

1. **Linear (LIN)** In the linear case, the classification problem is between a T-shaped tetromino versus an L-shaped tetromino pattern placed at the same fixed positions of the image throughout the entire dataset.

2. **Multiplicative (MULT)** The multiplicative scenario is similar to the LIN scenario with classifying T- versus L-shaped tetrominoes, however here each tetromino is multiplied with the background to induce non-linearity.

3. **Translations and rotations (RIGID)** Here, the T- and L-shaped tetromino patterns are not in a fixed position and are randomly translated and rotated anywhere in the sample, and added together with the underlying background.

4. **XOR** Both of the T- and L- shaped tetrominoes are present in each sample, but the classification problem is defined as the XOR configurations of adding or subtracting both tetrominoes from the background versus adding/subtracting one of each tetromino in the other class.

[1] `https://anonymous.4open.science/r/xai-whitening-aistats`

[2] `https://colab.research.google.com/`

2.1.1 Data whitening

Whitening represents a linear transformation applied to a d-dimensional random vector $\mathbf{x} = (x_1, \ldots, x_d)^\top$, which has a mean $\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)^\top$ and a positive definite $d \times d$ covariance matrix $\mathrm{var}[\mathbf{x}] = \boldsymbol{\Sigma}$. This transformation maps $\mathbf{x}$ to a new random vector:

$$\mathbf{z} = (z_1, \ldots, z_d)^\top = \mathbf{W}\mathbf{x} \tag{1}$$

where $\mathbf{z}$ maintains the same dimension d and has a "white" covariance with unit diagonal, $\mathrm{var}[\mathbf{z}] = \mathbf{I}$. The $d \times d$ matrix $\mathbf{W}$ is termed the whitening matrix. Whitening is especially critical in multivariate data analysis for both computational and statistical simplification and is frequently utilized in preprocessing and as part of modeling (Zuber & Strimmer, 2009; Hao et al., 2015). Whitening extends beyond merely standardizing a random variable, which is performed through:

$$\mathbf{z} = \mathbf{V}^{-\frac{1}{2}}\mathbf{x} \tag{2}$$

with $\mathbf{V} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$ containing the variances $\mathrm{var}[x_i] = \sigma_i^2$. This leads to $\mathrm{var}[z_i] = 1$, although it does not address correlations. Standardization and whitening transformations are often coupled with mean-centering of $\mathbf{x}$ or $\mathbf{z}$ to ensure $\mathbb{E}[\mathbf{z}] = 0$, though this is not mandatory for ensuring unit variances or white covariance. The whitening transformation as defined requires selecting a suitable whitening matrix $\mathbf{W}$. Since $\mathrm{var}[\mathbf{z}] = \mathbf{I}$, it follows that $\mathbf{W}\boldsymbol{\Sigma}\mathbf{W}^\top = \mathbf{I}$, thus $\mathbf{W}(\boldsymbol{\Sigma}\mathbf{W}^\top\mathbf{W}) = \mathbf{W}$, under the condition that

$$\mathbf{W}^\top\mathbf{W} = \boldsymbol{\Sigma}^{-1}. \tag{3}$$

Nevertheless, this condition does not uniquely specify the whitening matrix $\mathbf{W}$. In fact, given $\boldsymbol{\Sigma}$, there are infinitely many matrices $\mathbf{W}$ that fulfill this condition, each leading to a distinct whitening transformation producing orthogonal yet differently sphered random variables (Kessy et al., 2018).

For all whitening techniques, we regularize the covariance matrices before further calculation. This is done by checking if the smallest eigenvalue is negative or close to zero (under a threshold of 1×10^{-16}), and then adding a regularizing value slightly larger than the absolute minimal eigenvalue to the diagonal values of the covariance matrix. We study the following five techniques for which the Supplementary Materials B contains more details.

Symmetric Orthogonalization The method of Symmetric Orthogonalization (Annavarapu, 2013) transforms data into mutually uncorrelated variables such that the difference between original and transformed features is least-squares minimized. A one-to-one correspondence between original and transformed features is therefore available.

Cholesky Whitening Cholesky whitening (Kessy et al., 2018) applies the Cholesky decomposition to the covariance matrix. This decomposition leads to a lower triangular transformation matrix that leads to uncorrelated uniform-variance transformed features. Notably, the triangular structure induces an ordering, whereby the first feature remains unchanged, the second feature gets orthogonalized w.r.t. the first, the third feature gets orthogonalized w.r.t to the first two, and so on. Hence, this whitening depends on the order of pixels, which in our case is (N, H, W) for N the number of samples, H, the height of the image and W, the width.

Sphering This standard whitening technique multiplies the data with the inverse square root of the covariance matrix. Geometrically, this means first rotating the data onto the principal axes, then scaling the data to have unit variance across all principal axes, and then finally to rotate the data back into input space. As such, there is a one-to-one correspondence between original and transformed features (Kessy et al., 2018).

Partial Regression In contrast to the global approaches of the previous methods, partial regression (Velleman & Welsch, 1981) focuses on removing the linear dependence of each feature on the others, one at a time. This approach involves regressing each feature against all others and replacing it with the residuals of this regression. While this method does not directly ensure uncorrelated features with unit variance, it aims to remove some of the shared information between them.

Optimal Signal Preservation Whitening This technique aims to preserve the signal of the data while reducing redundancy among features (Kessy et al., 2018). It combines the correlation matrix-based whitening similar to sphering but adjusts the transformation based on the variances of the original features to maintain the signal strength. Such transformations are particularly designed to balance between decorrelation and preserving the original signal information of the features.

For each of these whitening techniques, the XAI-TRIS data is initially centered, and the resulting covariance matrix is regularized to ensure numerical stability.

Our aim is that, by applying these whitening methods to the CORR background scenarios, the correlated background features (inducing suppressors) will be decorrelated, thus reducing the impact of suppressors on the resulting explanations, and providing explanations with a more clear and direct attribution to the ground truth tetrominoes.

Figure 1 shows an example for each classification sce-

nario and background type (WHITE and CORR), as well as, for each whitening method, the whitened equivalent of the corresponding CORR sample. Higher SNRs are used here for demonstrative purpose.

2.2 Classifiers

Following the approach of Clark et al. (2024b), three different architectures are employed: (1) a Linear Logistic Regression (LLR) model, a single-layer fully-connected neural network; (2) a Multi-Layer Perceptron (MLP) with four fully-connected layers, using ReLU activations; and (3) a Convolutional Neural Network (CNN) with four ReLU-activated convolutional layers followed by max-pooling. All models lead to a two-neuron softmax-activated output layer. Specific implementation details are described in Supplementary Materials D. We train a model for each CORR scenario and also for each whitening method applied to the respective CORR scenario, making use of the appropriate whitened data for each model. We ensure that each trained model achieves at least 80% test accuracy, so that the resulting trained models have comparative performance.

2.3 XAI Methods

We analyze sixteen widely recognized methods within the domain of XAI. The core discussion is centered around the evaluation of four distinct XAI methods: Local Interpretable Model-Agnostic Explanations (LIME) (Ribeiro et al., 2016), Layer-wise Relevance Propagation (LRP) (Bach et al., 2015), Gradient SHAP (Lundberg & Lee, 2017), and Integrated Gradients (Sundararajan et al., 2017). The full list of methods studied and associated results can be seen in Supplementary Material Section E. Predominantly, default parameters are adhered to, with exceptions noted where a baseline $b = 0$ is explicitly defined, reflecting a widely recognized convention in the field Mamalakis et al. (2022).

The input for an XAI method is a trained ML model, the given test sample or batch of multiple samples designated for explanation, and (where relevant) the baseline test reference $b = 0$. The full results in the supplementary materials (Figures 7 and 8) make use of four model-ignorant techniques to establish baselines of explanation performance. This enables the assessment of whether the often intricate XAI methods genuinely offer superior explanations compared to approaches devoid of model-specific insights. The first method considered is the Sobel filter, employing both horizontal and vertical filter kernels to estimate the first-order derivatives of data. The second method utilized is the Laplace filter, which approximates the second-order derivatives of data using a single symmetrical kernel. Both methodologies serve as edge detection operators and are applied to each test sample. Additionally, random samples from a uniform distribution and the rectified test data sample itself are employed as 'explanations' for comparison purposes.

2.4 Explanation Performance Metrics

We define a 'correct' explanation as one which highlights truly important features for the classification task (i.e. features of the tetrominoes) and does not place false-positive importance on features outside of said ground truth.

We adopt the quantitative metrics used by Clark et al. (2024b), namely precision and earth mover's distance (EMD). These metrics serve as an objective and empirical foundation for analyzing how well a model's explanations align with a set of class-dependent features identified as ground truth.

The precision metric is calculated as the ratio of the correctly identified features within the top-k features ranked by their absolute importance scores to the total number of truly important features identified in the sample. The focus on the highest-ranking features reflects the real-world scenario where only the most influential factors are typically considered in decision-making processes (e.g., a doctor using a subset of symptoms to form a diagnosis).

The EMD quantifies the minimal expenditure required to transform one distribution into another. It is also known as the optimal transport distance. Applied in our context, this involves the cost needed to transform a continuous-valued explanation into the ground truth, with both distributions normalized to have equal 'mass'. The calculation of EMD utilizes the Euclidean distance between pixels as the ground metric. To calculated the EMD, we use the algorithm introduced by Bonneel et al. (2011) as implemented in the Python Optimal Transport library by Flamary et al. (2021). A normalized EMD performance score is defined by taking the optimal transport from an explanation to ground truth and dividing by the maximum euclidean distance possible. In practice we take one minus this score such that a score of 1.0 is the 'perfect' explanation.

As discussed by Clark et al. (2024b), both metrics assess the model's ability to highlight features that are truly relevant, as per the ground truth, while minimizing the inclusion of less significant (false-positive) features.

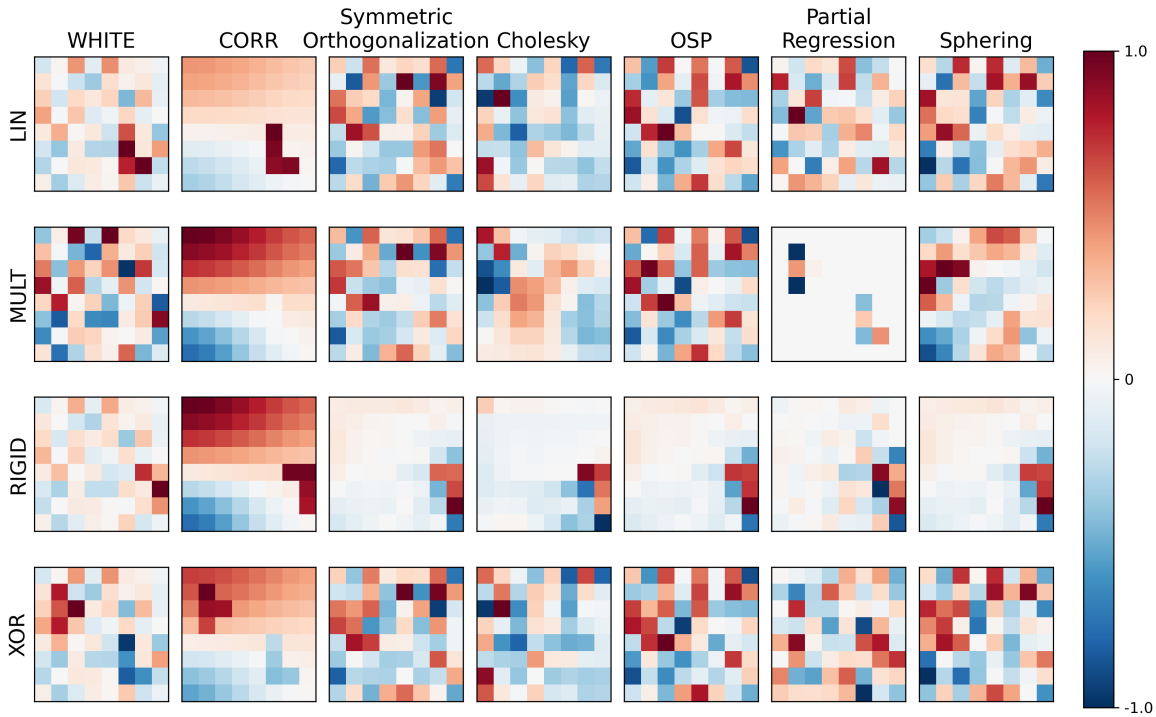

Figure 1: Examples of data for each scenario with the white noise (WHITE) and smoothed (CORR) backgrounds, and where various whitening methods are applied across the whole respective CORR dataset. While individual samples may look noisy and devoid of the tetromino features of the original, class-conditional information is still preserved globally in these features. The signal-to-noise ratios (SNRs) used here are increased for demonstrative purposes. The final analyzed parameterizations follow the SNRs used by Clark et al. (2024b).

3 Results

3.1 Qualitative Analysis

Supplementary Materials Figure 4 depicts the absolute-valued global importance heatmaps, the mean of all explanations for every correctly-predicted sample, for the LIN, MULT, and XOR scenarios. As the RIGID scenario has no static ground truth pattern, calculating a global importance map is not possible. Shown are results for four XAI methods (Gradient SHAP, LIME, LRP, and Integrated Gradients respectively) for each of the three models (LLR, MLP, CNN respectively) followed by the model-ignorant Laplace filter. This is shown for a random correctly-predicted sample, including the RIGID scenario, in Supplementary Materials Figure 2.

Interestingly, not all whitening techniques impact XAI interpretations uniformly. The Cholesky whitening and partial regression techniques demonstrate slightly more focused attributions but still show notable activity in regions outside the foreground signal. This indicates that while some varied amount of suppression of background noise is achieved, overall the techniques seem more prone to the negative influence of suppressor variables on explanation performance. Contrasting that, the optimal signal preservation and sphering methods, designed to preserve more of the data structure, only subtly modify explanations and can be observed to yield importance maps that more closely aligned with the true signal, indicating a stronger reduction of potential suppressor variable influence. Symmetric orthogonalization presents the most concentrated patterns of importance, closely mirroring the ground truth and demonstrating the highest resilience to the potentially misleading effects of suppressor variables among the examined whitening techniques. Figure 2 presents the importance maps obtained for a correctly-predicted data sample, for data with no whitening applied and data for which the symmetric orthogonalization whitening method was applied. Within the variety of XAI methods, gradient-based methods like Gradient SHAP and Integrated Gradients, illustrate an observable evolution from more dispersed attribution patterns in the non-whitened case, to more concentrated patterns as the data undergoes the various types of whitening.

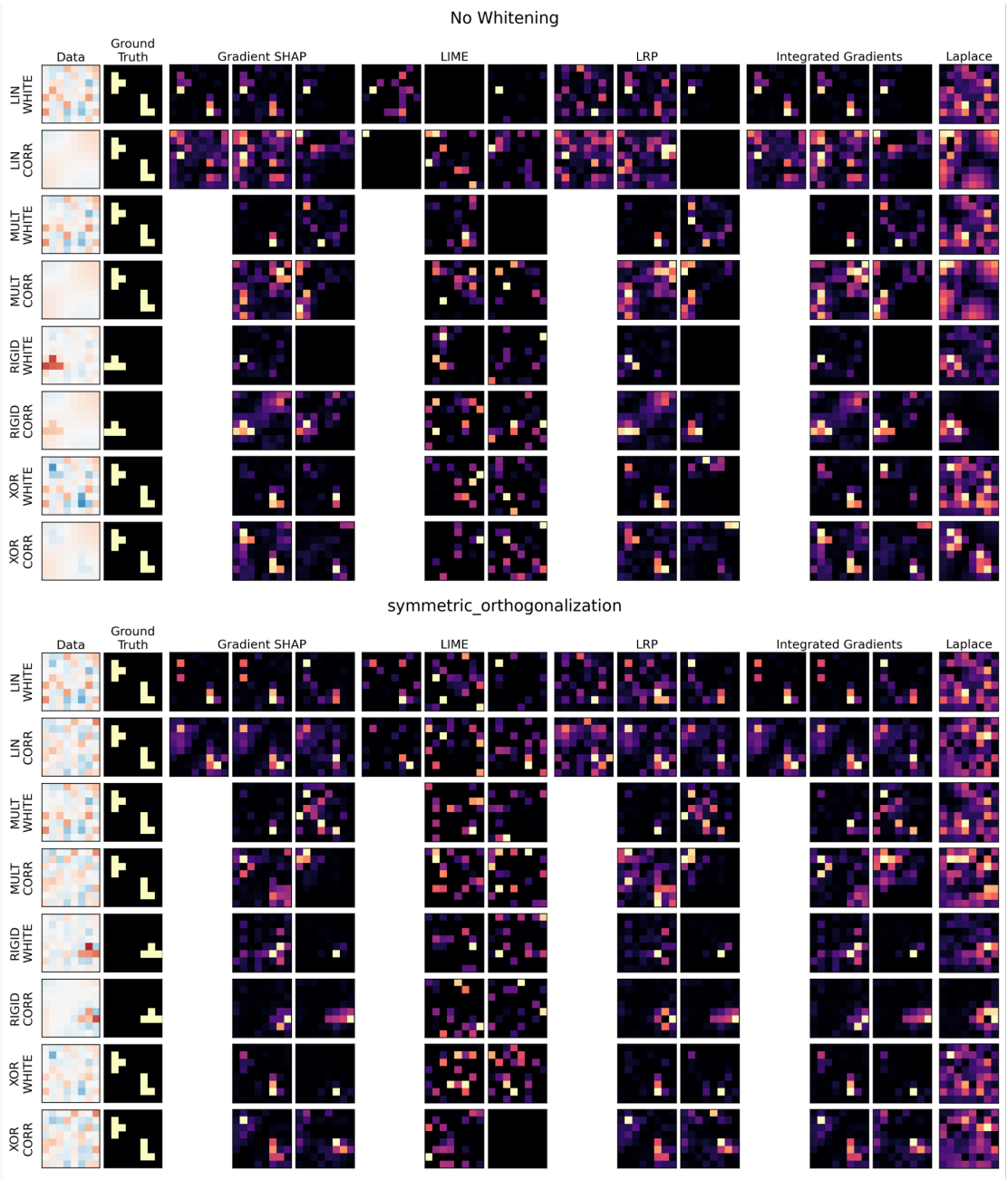

Figure 2: Absolute-valued importance maps obtained for a random correctly-predicted data sample, for XAI-TRIS (Clark et al., 2024a) scenario data with no whitening applied and data for which the symmetric orthogonalization whitening method was applied. Note, different samples are visualized for both cases, and the WHITE results are shown as a form of sanity check towards the 'ideal' results in the top plot where no whitening is applied. Visually comparing CORR results between the two plots, the results are mixed with some visual improvement in some problem scenarios such as LIN and XOR. This is seen in the form of reduced 'importance' attributed to background pixels and some increased attribution to the foreground tetromino features. The RIGID results show perhaps less noisy attributions to background pixels but little improvement (or even deterioration) in foreground tetromino importance attribution. Due to the variable position of the class-defining tetrominoes in this case, this may not be unexpected when compared to the other scenarios with fixed-position and perhaps more easily learnable correlation structures.

3.2 Quantitative Analysis

From Figure 3, it can be observed that the precision of XAI methods tends to decline when operating on the

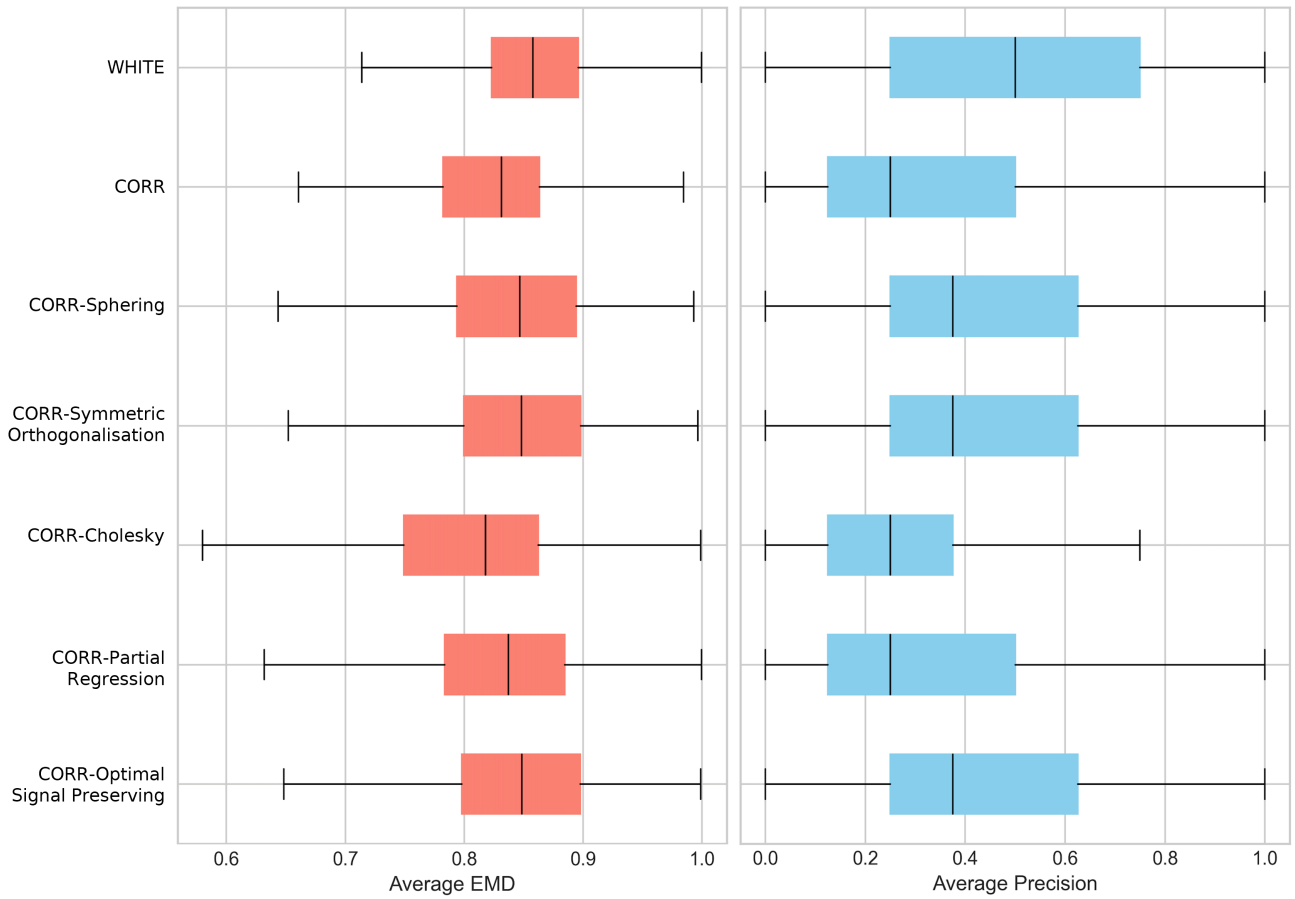

Figure 3: Average earth mover's distance (left) and precision (right) across all samples, XAI methods, and XAI-TRIS (Clark et al., 2024a) scenarios. This is split by background type, where the WHITE background serves as a 'baseline' with the CORR background type serving as the base for whitening. Each subsequent row therefore shows the application of different whitening methods to the underlying CORR background scenarios. Both metrics follow a similar trend of which whitening methods improve the correctness, where the sphering, symmetric orthogonalization, and optimal signal preserving methods perform the best – nearly reaching the performance levels of the WHITE results.

correlated background (compared to the uncorrelated case), confirming the notion that correlated noise negatively impacts the ability of XAI methods to accurately identify features of importance, due to the induction of suppressor variables. This decline is rectified to varying extents by the application of whitening techniques, which aim to remove the correlation between features in the dataset, thereby mitigating these potential suppressor variables that could lead to false attributions of importance. In this regard, Sphering, Optimal Signal Preserving, and Symmetric Orthogonalization stand out as the most effective techniques in restoring precision, indicating their strength in clarifying the data's structure and enhancing the ability of XAI methods to discern true signals from noise. As with the precision metric, sphering, symmetric orthogonalization and the optimal signal preserving transformation demonstrate the highest EMD values, reinforcing their effectiveness. For both metrics, it can be observed that partial regression and Cholesky whitening, while sometimes improving upon the correlated scenario, fall short compared to the other techniques. This suggests that while they do have a positive impact, they might not be as capable of dealing with complex correlations or might introduce artifacts that prevent the XAI methods from reaching the accuracy levels of the other techniques.

Supplementary Materials Figures 5 and 6 expand on the results of Figure 3 by splitting up results for each problem scenario and background type, and by the four main XAI methods studied. Supplementary Materials Figures 7 and 8 go even further by illustrating the EMD and precision results for all sixteen XAI methods

studied and four baselines for the non-whitened case, compared to the top performing whitening technique as identified by the qualitative analysis – symmetric orthogonalization. While we can see improvement in explanation performance in many cases where whitening is used, the results are not consistent across all XAI techniques.

4 Discussion and Implications

The presented analysis highlights the intricate relationship between data preprocessing techniques, specifically whitening, and the explanation performance provided by various XAI methods across different ML models. While whitening aims to simplify model training and improve numerical stability, its impacts on XAI interpretability are multifaceted and were the main point of investigation in this paper.

The observed results demonstrate that whitening does not offer a fundamental protection against spurious attribution to suppressor variables. Such a general effect could only be expected if the observed features are linear combinations of at most as many independent underlying signal or noise factors as there are features. For more underlying signals than features, whitening will inevitable need to mix discriminative and non-discriminative signal components into novel features, which could lead to worse explanations. Future work will look to find analytic expressions of the impact of whitening on explanations in the presence of suppressors, extending previous work of Wilming et al. (2023).

Despite these considerations, whitening did in some cases have a positive effect on explanation performance, depending on the method used. Each technique modifies the data in distinct ways, leading to unique alterations in the heatmaps maps generated by the XAI methods. This is evident by the consistent trend where whitening techniques both lead to a shift from diffused to localized importance patterns (that better match the ground truth as seen in the global absolute-valued importance maps) and produced better quantitative results compared to the correlated background case in which no whitening method was applied. Specifically, optimal signal preserving whitening and symmetric orthogonalization appear to be the most effective in this context, while Cholesky whitening and partial regression seem to be the least effective among the investigated whitening techniques. The complexities introduced by suppressor variables in XAI interpretations reinforce the need for careful consideration of background noise and its correlation structures when evaluating the performance of XAI methods. XAI methods may require additional mechanisms to distinguish between true predictors and correlated suppressors to maintain adequate explanation performance.

While the need for better and more theoretically justified methods is clear, application of data whitening methods can potentially help to improve explanation performance in the meantime. Caution is still advised, as explanations produced can still be noisy and misleading, with false positive attribution to non-important features still present.

Moreover, the full results in the supplementary materials show a mixed story of performance, and as such we cannot state that one particular XAI method or whitening transformation are outright the 'best' in these evaluations. This inconsistency in performance again highlights the need for better XAI methods. First, we need more ground truth benchmarks based not only on evaluating explanation correctness, but also other studied aspects like robustness (for example, Shah et al., 2021). With this, a more holistic view of the performance of XAI methods can be achieved, in order to guide future method development.

It is also important to apply these analyses to larger and more 'real-world' focused datasets with more complicated and less consistent background correlation structures present. However, the difficulty with these datasets is that there is no well-defined ground truth to directly measure explanation performance in terms of correctness, so this line of future work should first be auxiliary to a ground truth focused study.

5 Conclusion

The findings advocate for a tailored approach to data preprocessing, aimed at aligning whitening techniques with specific interpretative goals for the user's problem. It becomes apparent that achieving clear and understandable AI systems necessitates context-sensitive preprocessing strategies that do not compromise the depth and accuracy of explanations. The findings also call for continued exploration into the interplay between suppressor variables, model architecture, whitening techniques, and XAI method efficacy, with the goal of fostering the development of balanced AI systems that are both high-performing and interpretable. This balance is crucial for measuring that the developed systems are not only accurate and efficient but also transparent and understandable, ensuring their responsible and ethical application in real-world scenarios.

References

Agarwal, C., Krishna, S., Saxena, E., Pawelczyk, M., Johnson, N., Puri, I., Zitnik, M., and Lakkaraju, H. OpenXAI: Towards a transparent evaluation of

model explanations. *Advances in Neural Information Processing Systems*, 35:15784–15799, 2022.

Alber, M., Lapuschkin, S., Seegerer, P., Hägele, M., Schütt, K. T., Montavon, G., Samek, W., Müller, K.-R., Dähne, S., and Kindermans, P.-J. iNNvestigate Neural Networks! arXiv: 1808.04260, August 2018.

Annavarapu, R. N. Singular Value Decomposition and the Centrality of Löwdin Orthogonalizations. *American Journal of Computational and Applied Mathematics*, 2013:33–35, January 2013. doi: 10.5923/j.ajcam.20130301.06.

Arras, L., Osman, A., and Samek, W. CLEVR-XAI: A benchmark dataset for the ground truth evaluation of neural network explanations. *Information Fusion*, 81:14–40, 2022. ISSN 1566-2535.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K., and Samek, W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7):e0130140, 2015. doi: 10.1371/journal.pone.0130140.

Bonneel, N., Van De Panne, M., Paris, S., and Heidrich, W. Displacement interpolation using Lagrangian mass transport. In *Proceedings of the 2011 SIGGRAPH Asia conference*, pp. 1–12, 2011.

Castro, J., Gómez, D., and Tejada, J. Polynomial calculation of the Shapley value based on sampling. *Computers & Operations Research*, 36(5):1726–1730, 2009. ISSN 0305-0548.

Clark, B., Wilming, R., Dox, A., Eschenbach, P., Hached, S., Wodke, D. J., Zewdie, M. T., Bruila, U., Oliveira, M., Schulz, H., Cornils, L. M., Panknin, D., Boubekki, A., and Haufe, S. EXACT: Towards a platform for empirically benchmarking Machine Learning model explanation methods, 2024a. _eprint: 2405.12261.

Clark, B., Wilming, R., and Haufe, S. Xai-tris: non-linear image benchmarks to quantify false positive post-hoc attribution of feature importance. *Machine Learning*, pp. 1–40, 2024b.

Conger, A. A Revised Definition for Suppressor Variables: A Guide to Their Identification and Interpretation. *Educational and Psychological Measurement*, 34(1):35–46, 1974.

Fisher, A., Rudin, C., and Dominici, F. All Models Are Wrong, but Many Are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019.

Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., Gautheron, L., Gayraud, N. T. H., Janati, H., Rakotomamonjy, A., Redko, I., Rolet, A., Schutz, A., Seguy, V., Sutherland, D. J., Tavenard, R., Tong, A., and Vayer, T. POT: Python Optimal Transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.

Friedman, J. and Wall, M. Graphical Views of Suppression and Multicollinearity in Multiple Linear Regression. *The American Statistician*, 59(2):127–136, 2005.

Gevaert, A., Rousseau, A.-J., Becker, T., Valkenborg, D., Bie, T. D., and Saeys, Y. Evaluating Feature Attribution Methods in the Image Domain. *arXiv e-prints*, art. arXiv:2202.12270, 2022.

Golomb, S. W. *Polyominoes: Puzzles, Patterns, Problems, and Packings*, volume 111. Princeton University Press, 1996.

Hao, N., Dong, B., and Fan, J. Sparsifying the Fisher linear discriminant by rotation. *J. R. Statist. Soc. B*, 77:827–851, 2015.

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., and Bießmann, F. On the Interpretation of Weight Vectors of Linear Models in Multivariate Neuroimaging. *NeuroImage*, 87:96–110, 2014.

Kessy, A., Lewin, A., and Strimmer, K. Optimal Whitening and Decorrelation. *The American Statistician*, 72(4):309–314, October 2018. arXiv: 1512.00809.

Kindermans, P.-J., Schütt, K. T., Alber, M., Müller, K.-R., Erhan, D., Kim, B., and Dähne, S. Learning how to explain neural networks: PatternNet and PatternAttribution. In *International Conference on Learning Representations*, 2018.

Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S., and Reblitz-Richardson, O. Captum: A Unified and Generic Model Interpretability Library for PyTorch. arXiv: 2009.07896, 2020.

Li, X.-H., Shi, Y., Li, H., Bai, W., Cao, C. C., and Chen, L. An experimental study of quantitative evaluations on saliency methods. In *Proceedings of the 27th ACM sigkdd conference on knowledge discovery & data mining*, pp. 3200–3208, 2021.

Lundberg, S. M. and Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 4765–4774. Curran Associates, Inc., 2017.

Mamalakis, A., Barnes, E. A., and Ebert-Uphoff, I. Carefully choose the baseline: Lessons learned from applying XAI attribution methods for regression tasks in geoscience, 2022.

Montavon, G., Bach, S., Binder, A., Samek, W., and Müller, K. Explaining NonLinear Classification Decisions with Deep Taylor Decomposition. *Pattern Recognition*, 65:211–222, 2017.

Oliveira, M., Wilming, R., Clark, B., Budding, C., Eitel, F., Ritter, K., and Haufe, S. Benchmarking the influence of pre-training on explanation performance in MR image classification. *Frontiers in Artificial Intelligence*, 7, 2024. ISSN 2624-8212. doi: 10.3389/frai.2024.1330919. URL https://www.frontiersin.org/articles/10.3389/frai.2024.1330919.

Pourahmadi, M. Covariance estimation: the GLM and regularization perspectives. *Statistical Science*, 26: 369–387, 2011.

Ribeiro, M. T., Singh, S., and Guestrin, C. " Why Should I Trust You?" Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017.

Shah, H., Jain, P., and Netrapalli, P. Do input gradients highlight discriminative features? *CoRR*, abs/2102.12781, 2021. URL https://arxiv.org/abs/2102.12781.

Shrikumar, A., Greenside, P., and Kundaje, A. Learning Important Features Through Propagating Activation Differences. *arXiv: Machine Learning*, 2017.

Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In *Workshop at International Conference on Learning Representations*, 2014.

Springenberg, J., Dosovitskiy, A., Brox, T., and Riedmiller, M. Striving for simplicity: The all convolutional net. In *ICLR (workshop track)*, 2015.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic Attribution for Deep Networks. *Proceedings of the 34th International Conference on Machine Learning*, 2017.

Tjoa, E. and Guan, C. Quantifying Explainability of Saliency Methods in Deep Neural Networks. arXiv: 2009.02899, 2020.

Velleman, P. F. and Welsch, R. E. Efficient computing of regression diagnostics. *The American Statistician*, 35(4):234–242, 1981. ISSN 00031305, 15372731. URL http://www.jstor.org/stable/2683296.

Wilming, R., Budding, C., Müller, K.-R., and Haufe, S. Scrutinizing XAI Using Linear Ground-Truth Data with Suppressor Variables. *Machine Learning*, 2022.

Wilming, R., Kieslich, L., Clark, B., and Haufe, S. Theoretical Behavior of XAI Methods in the Presence of Suppressor Variables. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 37091–37107. PMLR, July 2023. URL https://proceedings.mlr.press/v202/wilming23a.html.

Wilming, R., Dox, A., Schulz, H., Oliveira, M., Clark, B., and Haufe, S. Gecobench: A gender-controlled text dataset and benchmark for quantifying biases in explanations. *arXiv preprint arXiv:2406.11547*, 2024.

Zeiler, M. D. and Fergus, R. Visualizing and Understanding Convolutional Networks. In Fleet, D., Pajdla, T., Schiele, B., and Tuytelaars, T. (eds.), *Computer Vision – ECCV 2014*, Lecture Notes in Computer Science, pp. 818–833. Springer International Publishing, 2014. ISBN 978-3-319-10590-1. Place: Cham.

Zhou, Y., Booth, S., Ribeiro, M. T., and Shah, J. Do Feature Attribution Methods Correctly Attribute Features? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 9623–9633, 2022.

Zuber, V. and Strimmer, K. Gene ranking and biomarker discovery under correlation. *Bioinformatics*, 25:2700–2707, 2009.

Checklist

The checklist follows the references. For each question, choose your answer from the three possible options: Yes, No, Not Applicable. You are encouraged to include a justification to your answer, either by referencing the appropriate section of your paper or providing a brief inline description (1-2 sentences). Please do not modify the questions. Note that the Checklist section does not count towards the page limit. Not including the checklist in the first submission won't result in desk rejection, although in such case we will ask you to upload it during the author response period and include it in camera ready (if accepted).

In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [**Yes**]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [**Not Applicable. We focus on the results of the application of whitening transforms to XAI methods in terms of explanation performance itself, rather than analyzing the listed qualities of XAI methods or Whitening methods.**]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [**Yes**]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [**Not Applicable**]
   (b) Complete proofs of all theoretical results. [**Not Applicable**]
   (c) Clear explanations of any assumptions. [**Not Applicable**]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [**Yes**]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [**Yes**]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [**Yes**]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [**Yes**]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [**Yes**]
   (b) The license information of the assets, if applicable. [**Yes**]
   (c) New assets either in the supplemental material or as a URL, if applicable. [**Yes**]
   (d) Information about consent from data providers/curators. [**Yes**]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [**Not Applicable**]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [**Not Applicable**]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [**Not Applicable**]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [**Not Applicable**]

A Data Generation

Following from Clark et al. (2024b), each dataset consists of images of size 8×8, formulated as $D = \{(x^{(n)}, y^{(n)})\}_{n=1}^{N}$, containing independent and identically distributed observations $(x^{(n)} \in \mathbb{R}^D, y^{(n)} \in \{0,1\})_{n=1}^{N}$ with $N = 10000$ and the dimensionality of the feature space $D = 64$. The entities $x^{(n)}$ and $y^{(n)}$ represent instances of the stochastic variables X and Y, governed by a joint probability density function $p_{X,Y}(x, y)$. In each defined scenario, the instance $x^{(n)}$ is synthesized by integrating a signal pattern $a^{(n)} \in \mathbb{R}^D$, which encapsulates the critical features that constitute the ground truth for a model explanation, with background noise $\eta^{(n)} \in \mathbb{R}^D$. In the analysis, a scenario is also considered where the signal pattern $a^{(n)}$ undergoes a random spatial rigid body transformation (involving translation and rotation of the tetromino) $R^{(n)} : \mathbb{R}^D \to \mathbb{R}^D$. In all other scenarios, the identity transformation is utilized, such that $R^{(n)} \circ a^{(n)} = a^{(n)}$. The transformed signal and noise components, $(R^{(n)} \circ a^{(n)})$ and $(G \circ \eta^{(n)})$, are horizontally concatenated into matrices $A = \left\{(R^{(1)} \circ a^{(1)}), \ldots, (R^{(N)} \circ a^{(N)})\right\}$ and $E = \left\{(G \circ \eta^{(1)}), \ldots, (G \circ \eta^{(N)})\right\}$. The signal and background components are then normalized by the Frobenius norms of A and E: $(R^{(n)} \circ a^{(n)}) \leftarrow (R^{(n)} \circ a^{(n)})/\|A\|_F$ and $(G \circ \eta^{(n)}) \leftarrow (G \circ \eta^{(n)})/\|E\|_F$, where the Frobenius norm of a matrix A is defined as $\|A\|_F := \left(\sum_{n=1}^{N} \sum_{d=1}^{D} (a_d^{(n)})^2\right)^{1/2}$. Additionally, the weighted sum of the signal and background components is computed, where the scalar parameter $\alpha \in [0, 1]$ determines the SNR. Two distinct generative models are adopted, diverging based on their method of combining these two elements either additively or multiplicatively. For data generated through either process, each sample $x^{(n)} \in \mathbb{R}^D$ is scaled to the range $[-1, 1]^D$, such that $x^{(n)} \leftarrow x^{(n)} / \max |x^{(n)}|$, where $\max |x^{(n)}|$ denotes the maximum absolute value of the sample $x^{(n)}$.

A.1 Additive Generation

In scenarios where the model is additive, the data generation formula for the n-th sample is defined as:

$$x^{(n)} = \alpha(R^{(n)} \circ a^{(n)}) + (1 - \alpha)(G \circ \eta^{(n)}) \tag{4}$$

where the signal pattern $a^{(n)} \in \mathbb{R}^D$ varies, embodying tetromino shapes based on the binary class label $y^{(n)}$ which is distributed according to a Bernoulli process with a success probability of 0.5. The noise component $\eta^{(n)}$, indicative of a non-informative background, is derived from a multivariate normal distribution $\mathcal{N}(0, I_D)$, resulting in white Gaussian noise with zero mean and an identity covariance matrix I_D. This setup ensures that noise in each feature dimension is independent and follows a standard-normal distribution, designated as the WHITE scenario. In each classification task, an alternate background context, termed CORR, is specified where a two-dimensional Gaussian spatial smoothing filter $G : \mathbb{R}^D \to \mathbb{R}^D$ modifies the noise element $\eta^{(n)}$, with the smoothing parameter (spatial standard deviation of the Gaussian) set to $\sigma_{\text{smooth}} = 3$.

A.2 Multiplicative Generation

In scenarios where the model is multiplicative, the sample-wise data generation process is defined as:

$$x^{(n)} = \left(1 - \alpha \left(R^{(n)} \circ a^{(n)}\right)\right) \left(G \circ \eta^{(n)}\right) \tag{5}$$

where $a^{(n)}$, $\eta^{(n)}$, $R^{(n)}$, and G are defined as previously stated, with A and E being Frobenius-normalized, and $\mathbf{1} \in \mathbb{R}^D$. This elaborate approach in generating datasets ensures the creation of a controlled setting crucial for the accurate and systematic assessment of XAI methods. Such an approach also serves to certify that the generated data accurately simulates various realistic scenarios while clearly separating signal from noise, which is pivotal for the analysis and interpretation phases that follow Clark et al. (2024b).

A.3 Suppressors Emergence

In the scenarios where background noise is correlated, the presence of suppressor variables is induced in both the additive and the multiplicative data generation cases. A suppressor, in this context, is identified as a pixel not part of the foreground $R^{(n)} \circ a^{(n)}$, while its activity still finds correlation with a foreground pixel through the application of the smoothing operator G. Drawing on characteristics of suppressor variables previously reported Conger (1974); Friedman & Wall (2005); Haufe et al. (2014); Wilming et al. (2023), it is anticipated that XAI

methods might erroneously attribute importance to suppressor features in both linear and non-linear settings. This misattribution can lead to decreased explanation performance when compared to scenarios involving white noise backgrounds.

A.4 Scenarios

Four distinct types of scenarios are introduced using tetrominoes Golomb (1996), which are geometric shapes consisting of four features. They are then utilized to define each signal pattern $a^{(n)} \in \mathbb{R}^{8\times 8}$. Tetrominos were chosen as the basis for signal patterns as they allow a fixed and controllable amount of features (pixels) per sample. Specifically, the T-shaped and L shaped tetrominoes were selected due to their four unique appearances under 90-degree rotations. These tetrominos are used to induce statistical associations between the features and the target in the previously mentioned four different binary classification problems Clark et al. (2024b).

Linear (LIN) In the linear case, the additive generation model from equation (4) is employed, where $R^{(n)}$ represents the identity transformation, combining the pure signal pattern and the Gaussian white noise background additively. T-shaped tetromino patterns a_T and L-shaped tetromino patterns a_L are utilized for signal patterns, positioned near the top-left corner if $y = 0$ and near the bottom-right corner if $y = 1$, respectively, thus constituting the binary classification problem. Each four-pixel pattern is encoded such that for each pixel in the tetromino pattern, positioned at the i-th row and j-th column, $a_{i,j}^{T/L} = 1$, and zero otherwise.

Multiplicative (MULT) The multiplicative generation process (5) with signal patterns a_T, a_L is defined with the same tetrominoes as in the linear case, while transformation $R^{(n)}$ remains the identity transform. In this scenario, a degree of non-linearity is introduced as the foreground tetromino pattern, when overlaying the background noise, is reduced towards zero. Therefore, values either increase or decrease depending on their original sign. The complexity introduced by the non-linearity renders linear classifiers unable to solve this classification problem effectively Clark et al. (2024b). This configuration is meant to evaluate how different machine learning methods can adjust to and manage intricate, interconnected data presentations that are not linear.

Translations and rotations (RIGID) In the RIGID scenario, the defining tetrominoes for each class, denoted as $a^{T/L}$, undergo random translations and rotations. This alteration adheres to a rigid body transform $R^{(n)}$, with the requirement that the entire 4-pixel tetromino must remain within the confines of the image space. Such a constraint ensures that despite the randomness of movement and orientation, the integrity of the tetromino shape is preserved within the visible boundaries of the dataset samples. This process is classified as an additive manipulation, consistent with the guidelines established in equation (4). In this context, the complexity introduced by the spatial transformations prevents the effective application of standard linear methods for resolving the classification challenges presented. Instead, such intricate scenarios often necessitate the usage of more sophisticated solutions, typically involving specialized neural network architectures such as Convolutional Neural Networks (CNNs). These architectures are specifically engineered to address the challenges posed by spatial variations within image data, making them better suited for capturing and interpreting the nuanced shifts and rotations applied to the tetromino shapes within the RIGID framework.

Exclusive or (XOR) In the XOR configuration, an additive challenge is presented where both tetromino variants, denoted as $a^{T/L}$, are utilized in each sample, with the transformation $R^{(n)}$ maintaining its role as the identity transform. Within this setup, the class membership is defined such that for the first class (where $y = 0$), a combination of both tetromino shapes is superimposed on the image background, either in a positive or negative overlay, expressed as $a^{XOR++} = a^T + a^L$ and $a^{XOR--} = -a^T - a^L$. Conversely, for the second class (where $y = 1$), the tetromino shapes are displayed in a contrasting manner; one shape is overlaid positively, and the other negatively, denoted as $a^{XOR+-} = a^T - a^L$ and $a^{XOR-+} = -a^T + a^L$. This ensures that all four XOR configurations are represented with equal frequency within the dataset.

B Whitening Techniques

B.1 Cholesky Whitening

Cholesky whitening utilizes the Cholesky decomposition to transform a dataset into one where all features are uncorrelated and possess unit variance. This technique ensures that the transformed features have a simpler structure, facilitating more stable numerical computations. The Cholesky whitening procedure encompasses the following steps:

1. Compute the covariance of the data matrix Σ

2. Perform Cholesky decomposition on Σ, which results in:

$$\Sigma = LL^\top \tag{6}$$

   Here, L is a lower triangular matrix with real and positive diagonal entries.

3. Apply the whitening transformation to obtain the decorrelated feature matrix X_{white}, computed as:

$$X_{\text{white}} = L^{-1}(X - \overline{X}) \tag{7}$$

   where X denotes the data matrix and $\overline{X}$ is the mean vector of the columns.

The utilization of the Cholesky whitening matrix leads to the formation of both a cross-covariance matrix and a cross-correlation matrix. These matrices are distinctive for being lower-triangular with positive diagonal elements (Kessy et al., 2018). The adoption of Cholesky factorization for whitening purposes inherently implies a specific ordering of the variables involved. This ordering is particularly beneficial for time series analysis, as it facilitates the incorporation of auto-correlation effects as highlighted by Pourahmadi (2011). The Cholesky whitening process is also recognized for its computational efficiency. Compared to alternative methods such as eigenvalue or singular value decompositions, Cholesky decomposition is generally quicker due to its simpler computational requirements. This efficiency makes Cholesky whitening an ideal choice for real-time processing tasks and scenarios where there are constraints on computational resources.

B.2 Partial Regression Whitening

The primary objective of this technique of whitening, is to modify each independent variable to isolate its unique variance, minimizing the influence of other variables. The procedure begins by first centering the data, which is an important step for ensuring that each variable contributes equally to the analysis by removing mean bias. Then follows the iterative residual calculation step which aims to reduce the influence of other features on each target feature, thereby whitening the dataset:

1. For each feature, separate it as the target (to be considered as a temporary dependent variable) from the matrix of remaining features (treated as independent variables).

2. Compute regression weights by applying the pseudo-inverse of the matrix of independent variables to the target feature. This identifies the extent to which other features predict the target.

3. Calculate and extract the residuals, which are the portions of the target feature not explained by its linear relationship with the other features. These residuals represent the "whitened" feature, emphasizing its unique variance.

By isolating the unique contributions of each variable, this approach facilitates clearer interpretation and identification of predictive features, essential for robust modeling and inference.

B.3 Optimal Signal Preservation Whitening

Optimal Signal Preservation (OSP) Whitening is a variant of whitening designed to reduce correlations among variables in a dataset while preserving the signal structure as effectively as possible. In contrast to traditional whitening methods, which may lead to significant signal distortion or dimensionality reduction, OSP whitening strives to retain the original characteristics of the data. The process involves the following steps:

1. **Data Centering:** The mean from each feature of the dataset is subtracted to ensure that the data is centered around zero. This step is critical for removing any bias that could distort the correlation analysis.
2. **Correlation Matrix Computation:** The correlation matrix from the centered data is calculated, instead of the covariance matrix. This method focuses on the normalized measure of the variables' linear relationships, providing a standardized basis for decorrelation.
3. **Transformation Matrix Construction:** The transformation matrix is computed by inverting the square root of the regularized correlation matrix and scaling it by the inverse of the square root of the variance of the centered data. This creates a whitening matrix that decorrelates the variables and equalyzes their variance without relying on the eigenvalue decomposition.
4. **Data Transformation:** The whitening transformation is applied to the centered data, resulting in a set of uncorrelated variables with unit variance. This step effectively whitens the data while aiming to preserve the original signal structure as much as possible.

OSP whitening, as described here, diverges from conventional methods by utilizing the correlation matrix and its direct transformations rather than relying on eigenvalue decomposition. This approach can provide a balance between reducing data redundancy (through decorrelation) and maintaining the integrity of the original signal.

B.4 Symmetric Orthogonalization

Symmetric orthogonalization, specifically Löwdin symmetric orthogonalization, is a method designed to convert a set of linearly independent, non-orthogonal vectors into an orthonormal set. This procedure is critical in quantum chemistry for orthogonalizing hybrid electron orbits, among other applications in computer science, mathematics, statistics, and biology (Annavarapu, 2013). The steps are:

1. **Overlap Matrix Calculation:** The first step involves computing the overlap matrix S through the equation $S = X^\top X$, where X represents the matrix of basis vectors. The overlap matrix S quantifies the non-orthogonality among the basis vectors.
2. **Hermitian Metric Matrix and Its Decomposition:** Introduce a Hermitian metric matrix M, related to S, and perform its decomposition. According to Löwdin's method, S can be diagonalized, leading to the formation of $S = UDU^\top$, where U contains the eigenvectors, and D is a diagonal matrix of eigenvalues.
3. **Orthogonalization Matrix Formation:** The orthogonalization matrix P is then formed as $P = UD^{-\frac{1}{2}}U^\top$. Applying P to the initial set of basis vectors yields an orthonormal set, aligning with the principle of minimizing deformation from the original vectors in the least-squares sense (Annavarapu, 2013).

Löwdin symmetric orthogonalization stands apart from sequential methods like Gram-Schmidt by treating all vectors simultaneously, thereby preserving symmetry and ensuring minimal deformation of the basis vectors. This "democratic" approach underpins its widespread utility across various scientific disciplines, underlining the method's centrality to orthogonalization techniques (Annavarapu, 2013).

C Defining Ground Truth Feature Importance

Ground truth feature importance is quantitatively defined through the identification of significant pixels, where the significance of a pixel is determined by its statistical relationship with the target outcome (Wilming et al., 2023). This leads to the establishment of ground truth sets for significant pixels, considering the positions occupied by tetromino patterns within the dataset, formalised as:

$$F^+(x^{(n)}) := \{d | \left(R^{(n)} \circ a^{(n)}\right)_d \neq 0, d \in \{1, \ldots, 64\}\}. \tag{8}$$

In the contexts of both LIN and MULT, each dataset sample includes either a T or an L shaped tetromino, each anchored at predetermined positions, corresponding respectively to the patterns a_T and a_L. This structured approach ensures that the absence of a tetromino shape at one specific location is considered as informative as the presence of the alternate shape in a different location, enhancing the comprehensive nature of the pixel importance set in these contexts as:

$$F^+(x^{(n)}) := \{d | (H \circ a_T)_d \neq 0 \vee (H \circ a_L)_d \neq 0, d \in \{1, \ldots, 64\}\}. \tag{9}$$

This conceptual framework is identical to equation 8 for the XOR challenge and adheres to the operational definition of feature importance as established by Wilming et al. (2023), applied uniformly across the LIN, MULT, and XOR scenarios. In these analyses, a feature is recognized as significant if it demonstrates a statistical relationship with the target outcome across the dataset under review. Consequently, the most important criterion for any optimal explanation method within this framework is to assign significance exclusively to elements within the set $F^+(x^{(n)})$, thereby ensuring that the attribution of importance is directly tied to statistically relevant features (Clark et al., 2024b).

D Classifiers

Convolutional layers in the CNN architecture are defined with parameters set to enable comprehensive feature analysis: four filters, a kernel size of two, a stride of one, and padding designed to preserve the dimensional integrity between input and output shapes. This padding not only enhances pixel utilization throughout each convolution but also serves to prevent the reduction of output sizes from the already compact images by introducing zero-value filler pixels at the peripheries Clark et al. (2024b). Some widely recognized CNN features like batch normalization are omitted due to compatibility issues with various XAI methodologies. For the parameterization θ and the training dataset D_{train}, classifiers denoted as $f_\theta : \mathbb{R}^D \to Y$ are trained. The training of each network spans over 500 epochs, utilizing the Adam optimiser without regularization. A distinct learning rate is applied based on the scenario: 0.004 for the LIN, MULT, and XOR scenarios, and a reduced rate of 0.0004 for the RIGID scenario to account for its increased complexity. During training, the validation dataset D_{val} plays a crucial role at each epoch, offering insights into the model's generalization capabilities on unseen data. The validation loss, computed at every epoch, serves as a marker for assessing when the classifier has attained its optimal performance. This is determined by recording the model state at the epoch where the validation loss is at its minimum, a strategy that aids in circumventing typical issues of model overfitting. Upon concluding the training phase, the test dataset D_{test} is employed to evaluate the finalised model's performance, which is also pivotal in the subsequent analysis of XAI methodologies. A classifier is considered to have effectively generalised the classification challenge if it achieves a test accuracy that meets or surpasses an 80% threshold. To accommodate experimentation across a diverse array of XAI methods, each network is constructed within both PyTorch and Keras environments, leveraging a TensorFlow backend. This dual-implementation approach allows for compatibility with a wide range of XAI tools, including those supported by the Captum Kokhlikyan et al. (2020) and iNNvestigate Alber et al. (2018) frameworks.

E XAI Methods

F Results

F.1 Qualitative Results

F.2 Quantitative Results

Table 1: Summary of XAI Methods Analyzed as per Clark et al. (2024b)

XAI Method	**Description**	**Reference, Framework, Parameterization**
Permutation Feature Importance (PFI)	Measures the change in prediction error of the model after permuting each feature's value.	Fisher et al. Fisher et al. (2019), Captum, Default,
Integrated Gradients	Computes gradients along the path from a baseline input to the input sample and cumulates these through integration to form an explanation.	Sundararajan et al. Sundararajan et al. (2017), Captum, Default, Zero input baseline
Saliency	Computes the gradients with respect to each input feature.	Simonyan et al. Simonyan et al. (2014), Captum, Default
Guided Backpropagation	Computes the gradient of the output with respect to the input but ensures only non-negative gradients of ReLU functions are backpropagated.	Springenberg et al. Springenberg et al. (2015), Captum, Default
Guided GradCAM	Computes the element-wise product of guided backpropagation attributions with respect to a class-discriminative localization map in the final convolution layer of a CNN.	Selvaraju et al. Selvaraju et al. (2017), Captum, Default
Deconvolution	Uses a Deconvolutional network to map features to pixels, ensuring only non-negative gradients of ReLU functions are backpropagated.	Zeiler and Fergus Zeiler & Fergus (2014), Captum, Default
DeepLift	Compares the activation of each neuron to its 'reference activation' and produces an explanation based on this difference.	Shrikumar et al. Shrikumar et al. (2017), Captum, Default, Zero input baseline
Shapley Value Sampling	Approximates Shapley values by repeatedly sampling random permutations of input features and calculating the contribution of each feature to the prediction.	Castro et al. Castro et al. (2009), Captum, Default, Zero input baseline
Gradient SHAP	Approximates Shapley values by computing the expected values of gradients when randomly sampled from the distribution of baseline samples.	Lundberg and Lee Lundberg & Lee (2017), Captum, Default, Zero input baseline
Kernel SHAP	Approximates Shapley values through the use of LIME, setting the loss function weighting kernel and regularization term in accordance with the SHAP framework.	Lundberg and Lee Lundberg & Lee (2017), Captum, Default, Zero input baseline
Deep SHAP	Approximates Shapley values through the use of DeepLift, computing the DeepLift attribution for each input sample with respect to each baseline sample.	Lundberg and Lee Lundberg & Lee (2017), Captum, Default, Zero input baseline
Locally-interpretable Model Agnostic Explanations (LIME)	Learns a linear surrogate model locally to an individual prediction, perturbing and weighting the dataset in the process, then builds an explanation by interpreting this local model.	Ribeiro et al. Ribeiro et al. (2016), Captum, Default
Layer-wise Relevance Propagation (LRP)	Propagates the model output back through the network as a measure of relevance, decomposing this score for each model layer.	Bach et al. Bach et al. (2015), Captum, Default
Deep Taylor Decomposition (DTD)	Applies a Taylor decomposition from a specified root point to approximate the network's sub-functions, building explanations backward from the output to input variables.	Montavon et al. Montavon et al. (2017), iNNvestigate, Default
PatternNet	Estimates activation patterns per neuron through signal estimator and back-propagates this through the network.	Kindermans et al. Kindermans et al. (2018), iNNvestigate, Default
PatternAttribution	utilizes the theory of PatternNet to estimate the root point for Deep Taylor Decomposition and yields the attribution for weight vector and positive activation patterns.	Kindermans et al. Kindermans et al. (2018), iNNvestigate, Default

Figure 4: Absolute-valued global importance maps calculated as the mean importance value over all correctly predicted samples, for selected XAI methods and baselines. While some attributions appear qualitatively better when comparing the CORR columns between the 'No Whitening' plot in the top left and the plots for each whitening method, attributions still appear noisy and potentially misleading. Cholesky whitening looks to focus on one or two pixels in the LIN CORR and XOR CORR cases respectively, and produces much sparser heatmaps than for the no whitening cases.

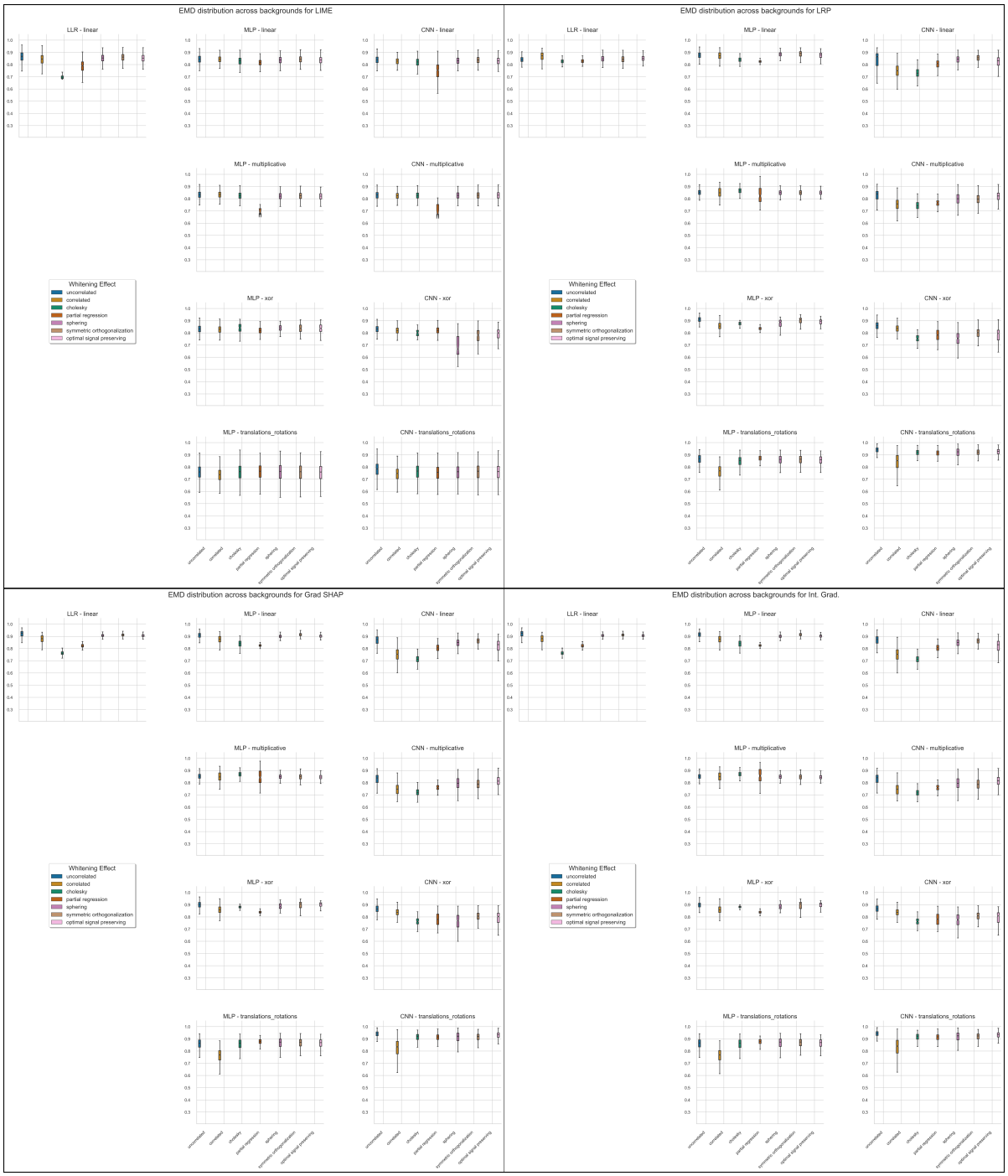

Figure 5: Boxplots of EMD scores across all problem scenarios and background types, where each plot is separated for each of the four main XAI methods studied. For each individual box plot, we can compare the correlated results to the whitened versions (applied to the CORR data) and then to the 'ideal' of the uncorrelated (WHITE) cases. Generally, the resulting trends follow similarly to the results of Figure 3, but we can see some individual cases of distinctly worse performance, such as for LIME with Cholesky whitening in the linear case with the linear model, or partial regression in the multiplicative cases. Comparing across box plots of a particular row, we can see how model choice affects the strength of improvement in explanation performance when whitening is applied. Generally, whitening looks to have the most variation in results in the LIN and XOR cases, which is partially to be expected due to the linear nature of the underlying transformations.

Figure 6: Precision scores across all problem scenarios and background types, where each plot is separated for each of the four main XAI methods studied. Similarly to Figure 5, we can compare the correlated results to the whitened versions (applied to the CORR data) and then to the 'ideal' of the uncorrelated (WHITE) cases. Generally, the resulting trends follow similarly to the results of Figure 3 and Figure 5, but we can see some more extreme cases than the narrowly spread EMD results. For example, partial regression in the multiplicative cases for all methods performs very well, and near perfect with the MLP model. Comparing across box plots of a particular row, we can see how model choice affects the strength of improvement in explanation performance when whitening is applied. Generally, whitening looks to have the most variation in results in the LIN and XOR cases, which is partially to be expected due to the linear nature of the underlying transformations. Comparing across the rows also allows us to see that LIME presents little variation in Precision results compared to the other three methods shown, where in the latter cases the CORR (and sometimes WHITE) results have higher precision. In a lot of these cases, there is also a higher disparity when comparing Cholesky and partial regression to the other three whitening methods studied.

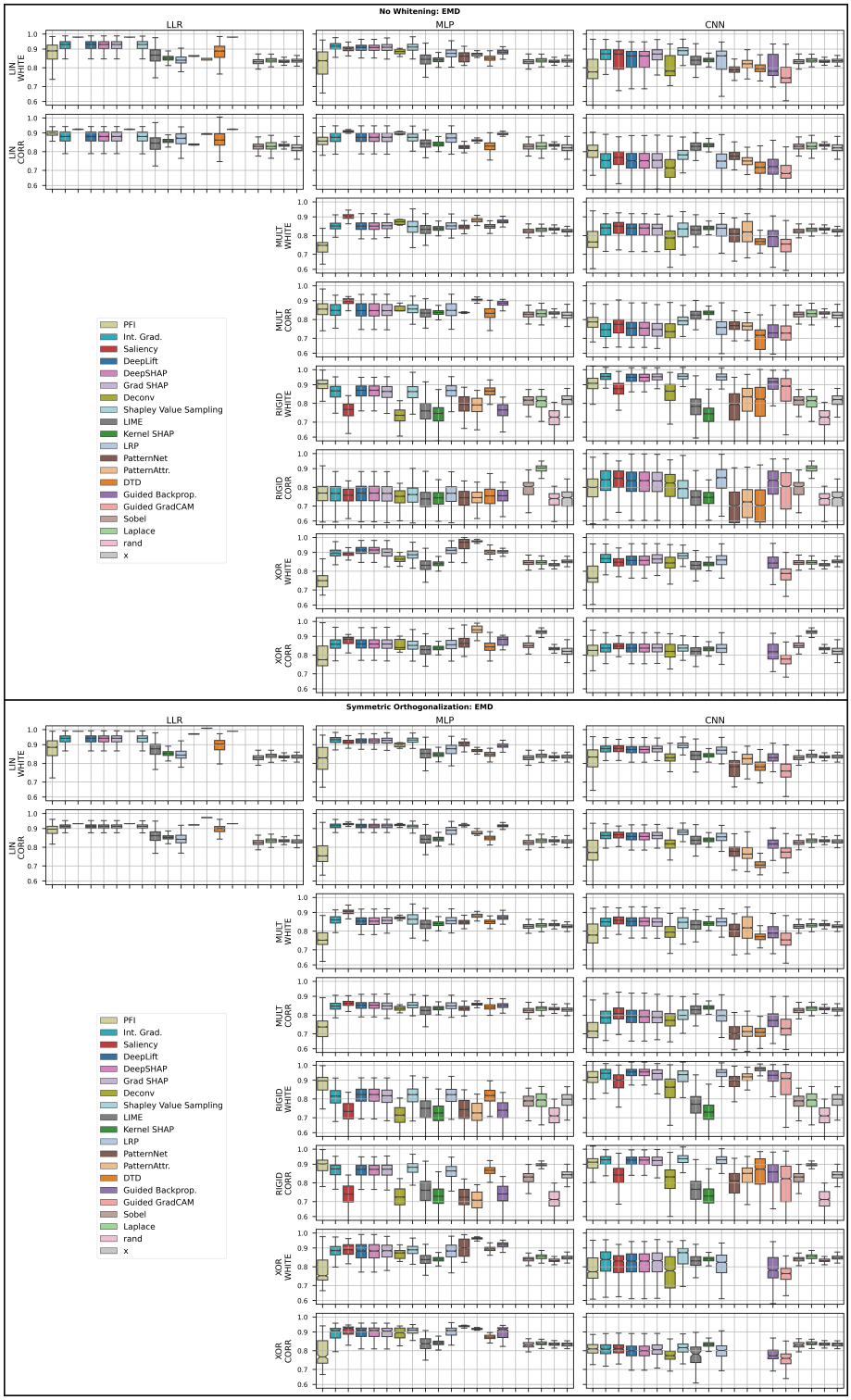

Figure 7: EMD results for all investigated XAI methods and baselines visualized as boxplots of median and quartile scores. The top plot shows the case where no whitening methods are applied to any scenario, and the bottom shows the equivalent where Symmetric Orthogonalization is applied to every scenario, even the WHITE background scenarios. A slight increase in EMD performance can be seen when whitening is applied, whilst retaining the same general trend in XAI method results.

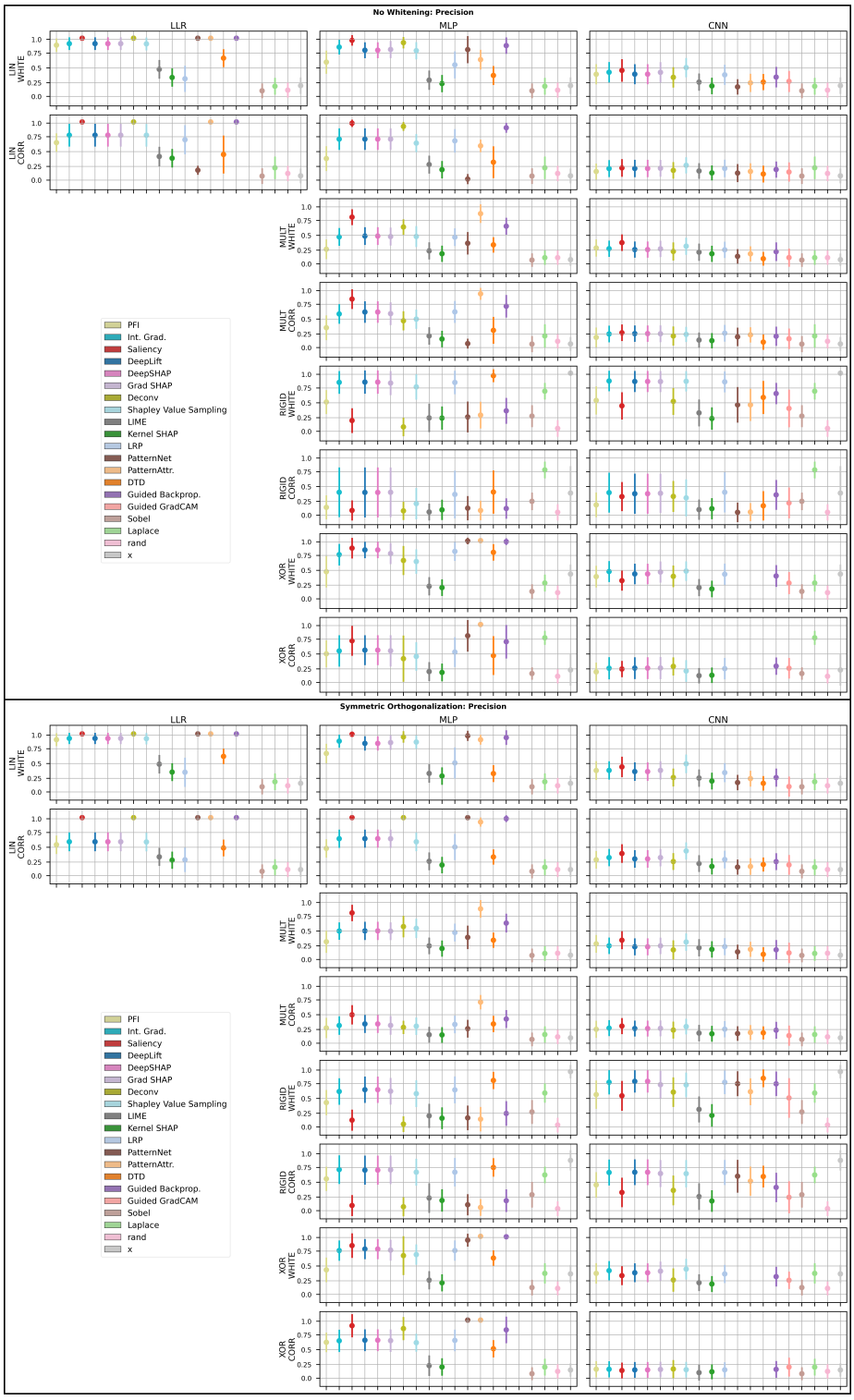

Figure 8: Mean and standard deviation Precision results for all investigated XAI methods and baselines. The top plot shows the case where no whitening methods are applied to any scenario, and the bottom shows the equivalent where Symmetric Orthogonalization is applied to every scenario, even the WHITE background scenarios. A slight increase in Precision performance can be seen when whitening is applied, whilst retaining the same general trend in XAI method results.