

Right on Time: Revising Time Series Models by Constraining their Explanations

Maurice Kraus¹ *, David Steinmann^{1,2} *, Antonia Wüst¹, Andre Kokozinski⁵, and Kristian Kersting^{1,2,3,4}

¹ Artificial Intelligence and Machine Learning Group, TU Darmstadt
 {maurice.kraus,david.steinmann}@cs.tu-darmstadt.de

² Hessian Center for Artificial Intelligence (hessian.AI), Darmstadt
³ Centre for Cognitive Science, TU Darmstadt

⁴ German Center for Artificial Intelligence (DFKI)

⁵ Institute for Production Engineering and Forming Machines, TU Darmstadt

Abstract. Deep time series models often suffer from reliability issues due to their tendency to rely on spurious correlations, leading to incorrect predictions. To mitigate such shortcuts and prevent "Clever-Hans" moments in time series models, we introduce Right on Time (RioT), a novel method that enables interacting with model explanations across both the *time* and *frequency* domains. By incorporating feedback on explanations in both domains, RioT constrains the model, steering it away from annotated spurious correlations. This dual-domain interaction strategy is crucial for effectively addressing shortcuts in time series datasets. We empirically demonstrate the effectiveness of RioT in guiding models toward more reliable decision-making across popular time series classification and forecasting datasets, as well as our newly recorded dataset with naturally occurring shortcuts, P2S, collected from a real mechanical production line.

1 Introduction

Time series data is ubiquitous in today’s world. Everything that is measured over time generates some form of time series, for example, energy load [15], sensor measurements in industrial machinery [21] or recordings of traffic data [18]. Complex time series data is often analyzed using various neural models [3,28]. However, as in other domains, these can be subject to spurious factors ranging from simple noise or artifacts to complex shortcuts [16]. Intuitively, a shortcut, also called “Clever-Hans” moment, is a spurious pattern in the data that correlates with the target task during training but lacks true relevance. If a model learns to rely on such patterns rather than meaningful features, its generalizability suffers, performing well on data with the shortcut but failing on data without

*These authors share equal contribution.

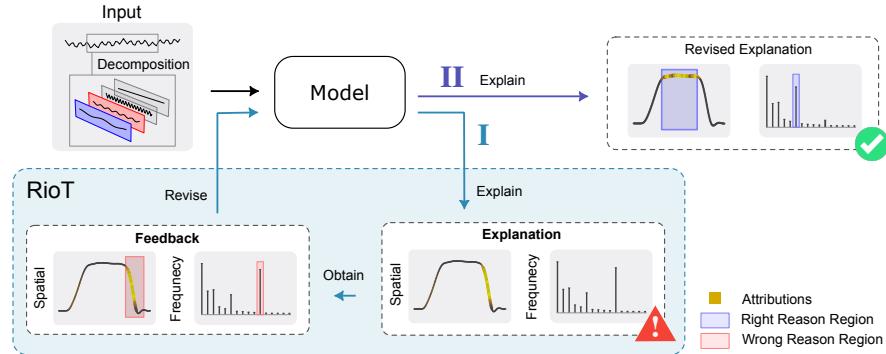


Fig. 1. Explanations reveal that the model relies on spurious factors in the input (red region) instead of relevant features (blue region). With RioT, the model can be guided away from these misleading patterns, whether they appear in the *spatial* or *frequency* domain. For this, RioT leverages feedback on *incorrect explanations* to steer the model toward more meaningful and *reliable reasoning*.

it, which poses a significant challenge in real-world deployment [11]. While model explanations can help uncover these shortcuts, they do not resolve the issue on their own (cf. Fig. 1 **I**). Despite extensive research in other domains [36], shortcuts in time series models remain underexplored. Existing studies often have specific assumptions about settings and data [4], leaving a gap in understanding and mitigating shortcut learning in broader time series applications. To address this, we introduce Right on Time (RioT), a new method grounded in the principles of explanatory interactive learning (XIL) [38], which leverages feedback on explanations to mitigate shortcuts (cf. Fig. 1 **II**). RioT uses traditional explanation methods, such as Integrated Gradients (IG) [37], to assess whether the model attends to the correct time steps. It then incorporates feedback on shortcut areas to refine the model, improving robustness and generalization.

However, spurious factors in time series data extend beyond the time domain. For example, a consistent noise frequency in an audio signal can act as a shortcut without being tied to a specific point in time. RioT can handle these types of shortcuts by incorporating feedback in the frequency domain. To highlight the importance of shortcuts in time series data, we introduce a new real-world dataset with naturally occurring shortcuts, called PRODUCTION PRESS SENSOR DATA (P2S). The dataset includes sensor measurements from an industrial high-speed press, essential to many manufacturing processes in the sheet metal working industry. The sensor data for detecting faulty production contains shortcuts and thus provokes incorrect predictions after training. Next to its industrial relevance, P2S is the first time series dataset that contains explicitly annotated shortcuts, enabling the evaluation of mitigation strategies on real data.

Altogether, we make the following contributions: (1) We show both on our newly introduced real-world dataset P2S and on several other datasets with manual shortcuts that SOTA neural networks on time series classification and

forecasting are affected by these shortcuts. (2) We introduce RioT to mitigate shortcuts for time series data. The method can incorporate feedback on the time domain and the frequency domain. (3) By incorporating explanations and feedback in the frequency domain, we enable a new perspective on XIL, overcoming the important limitation that shortcuts must be spatially separable.

The paper is structured as follows: Sec. 2 provides a brief overview of related work on explaining time series and correcting model mistakes. Sec. 3 introduces our approach, while Sec. 4 describes our decoy methods and P2S. We then present a detailed evaluation and discussion in Sec. 5. Finally, Sec. 6 concludes the paper and outlines directions for future research.

2 Related Work

Explanations for Time Series. Explainable artificial intelligence offers various techniques to interpret machine learning models, many of which originated in image or text data before being adapted for time series [26]. Attribution methods explain models directly in the input space, while approaches like symbolic aggregation [17] and shapelets [42] provide higher-level insights (cf. [26,29] for a broader discussion on time series explanations). While explanation methods help identify shortcuts, they alone do not enable model revision. Thus, our approach begins with explanations to detect shortcuts and integrates feedback to mitigate them. Specifically, we use Integrated Gradients (IG) [37], which computes attributions via model gradients and is widely used for time series data [22,39].

Explanatory Interactive Learning (XIL). Research on shortcuts and their mitigation is growing, though it primarily focuses on visual data [36]. One direction is explanatory interactive learning (XIL), which entails methods that revise a model’s decision-making based on human feedback [30,38]. A core aspect of XIL is using model explanations to correct mistakes, particularly to prevent Clever-Hans-like behavior, where models rely on spurious shortcuts [9,35]. Several XIL methods have been applied to image data. Right for the Right Reasons (RRR) [27] and Right for Better Reasons [32] penalize incorrect attributions, while HINT [31] rewards correct focus and [10] explore using multiple explainers. Despite their success in vision tasks, XIL approaches remain largely unexplored for time series. To address this, we introduce RioT, adapting XIL principles to the unique challenges of time series data.

Unconfounding Time Series. Apart from interactive learning approaches, some methods address confounding in time series models through causal inference [8]. Techniques like the Time Series Deconfounder [4], SqeDec [13], and LipCDE [5] estimate data while mitigating confounders in covariates of the target variable. They rely on causal analysis and specific assumptions about data generation. In contrast, our method focuses on shortcuts within the target variable itself, requiring no assumptions beyond the shortcut being detectable in model explanations - an area where existing causal methods are less applicable.

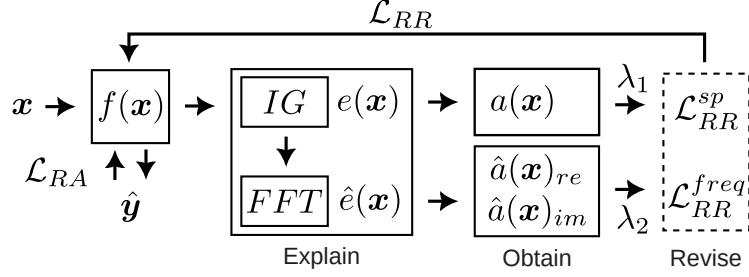


Fig. 2. RioT’s explanation-based revision process. Input data \mathbf{x} passes through the model $f(\mathbf{x})$ to generate explanations $e(\mathbf{x})$, receives annotated feedback $a(\mathbf{x})$, and is then fed back into the model. Integrated Gradients (IG) provides spatial explanations, while Fourier Transform (FFT) converts them into frequency-based explanations. Annotations can be applied in either or both domains and are leveraged by the right-reason loss (\mathcal{L}_{RR}^{sp} and \mathcal{L}_{RR}^{freq}) to steer the model away from shortcuts in time or frequency.

3 Right on Time (RioT)

The core idea of Right on Time (RioT) is to use feedback on model explanations to guide the model away from incorrect reasoning. Following the XIL paradigm, RioT is designed for seamless integration with other XIL methods. To ensure compatibility, we structure RioT around the four key steps identified by [9]: *Select*, *Explain*, *Obtain* and *Revise*. In *Select*, samples for feedback and model revision are selected. Following previous methods, we select all samples by default but also explore using subsets of the data. Afterwards, *Explain* covers model explanations before feedback is provided in *Obtain*. Lastly, in *Revise*, the feedback is integrated into the model to overcome the shortcuts. We introduce RioT along these steps in the following (as illustrated in Fig. 2).

Given a dataset $(\mathcal{X}, \mathcal{Y})$ and a model $f(\cdot)$ for time series classification or forecasting. The dataset consists of D many pairs of \mathbf{x} and \mathbf{y} . Thereby, $\mathbf{x} \in \mathcal{X}$ is a time series of length T , i.e., $\mathbf{x} \in \mathbb{R}^T$. For K class classification, the ground-truth output is $\mathbf{y} \in \{1, \dots, K\}$ and for forecasting, the ground-truth output is the forecasting window $\mathbf{y} \in \mathbb{R}^W$ of length W . The ground-truth output of the full dataset is described as \mathcal{Y} in both cases. For a datapoint \mathbf{x} , the model generates the output $\hat{\mathbf{y}} = f(\mathbf{x})$, where the dimensions of $\hat{\mathbf{y}}$ are the same as of \mathbf{y} .

3.1 Explain

Given a pair of input \mathbf{x} and model output $\hat{\mathbf{y}}$ for time series classification, the explainer generates an explanation $e_f(\mathbf{x}) \in \mathbb{R}^T$ in the form of attributions to explain $\hat{\mathbf{y}}$ w.r.t. \mathbf{x} , where a large attribution value means a large influence on the output. In the remainder of the paper, explanations refer to the model f , but we drop f from the notation to declutter it, resulting in $e(\mathbf{x})$. We use IG [37] (Eq. 1) as an explainer, an established gradient-based attribution method. This method integrates the gradient along the path (using the integration variable

α) from a baseline $\bar{\mathbf{x}}$ to the input \mathbf{x} . It multiplies the result with the difference between baseline and input. However, we make some adjustments to the base method to make it more suitable for time series and model revision, namely taking the absolute value of the difference between \mathbf{x} and $\bar{\mathbf{x}}$ (further details in Appx. A.1). In the following, we introduce the modifications to use attributions for forecasting and to obtain explanations in the frequency domain.

$$e(\mathbf{x}) = |\mathbf{x} - \bar{\mathbf{x}}| \cdot \int_0^1 \frac{\partial f(\tilde{\mathbf{x}})}{\partial \tilde{\mathbf{x}}} \Big|_{\tilde{\mathbf{x}}=\bar{\mathbf{x}}+\alpha(\mathbf{x}-\bar{\mathbf{x}})} d\alpha \quad (1) \quad e(\mathbf{x}) = \frac{1}{W} \sum_{i=1}^W e'_i(\mathbf{x}) \quad (2)$$

Attributions for Forecasting. In a classification setting, attributions are generated by propagating gradients back from the model output (of its highest activated class) to the model inputs. However, there is often no single model output in time series forecasting. Instead, the model simultaneously generates one output for each timestep of the forecasting window. Naively, one could use these W outputs and generate as many explanations $e'_1(\mathbf{x}), \dots, e'_W(\mathbf{x})$, where each $e'_i(\mathbf{x})$ is the IG explanation using the i -th time step from the forecasting window as a target instead of a classification label. This number of explanations would, however, make it even harder for humans to interpret the results, as the size of the explanation increases with W [23]. Therefore, we propose aggregating the individual explanations by averaging in Eq. 2. Averaging attributions over the forecasting window provides a simple yet robust aggregation of the explanations. Other means of combining them, potentially even weighted based on distance of the forecast in the future are also imaginable. Overall, this allows attributions for time series classification and forecasting to be generated similarly.

Attributions in the Frequency Domain. Time series data is often given in the frequency representation, and this format can sometimes be more intuitive for humans to understand than the typical spatial representation. Thus, providing explanations in this domain is essential. [40] showed how to obtain frequency attributions of the method Layerwise Relevance Propagation [1], even if the model does not operate directly on the frequency domain. We adapt this idea to IG: for an input sample \mathbf{x} , we generate attributions with IG, resulting in $e(\mathbf{x}) \in \mathbb{R}^T$ (Eq. 1 for classification or Eq. 2 for forecasting). We then interpret the explanation as a time series, with the attribution scores as values. To obtain the frequency explanation, we perform a Fourier transformation of $e(\mathbf{x})$, resulting in the frequency explanation $\hat{e}(\mathbf{x}) \in \mathbb{C}^T$ with \hat{E} for the entire set.

3.2 Obtain

The next step of RioT is to obtain feedback on shortcuts. For an input \mathbf{x} , feedback marks input parts via a binary mask $a(\mathbf{x}) \in \{0, 1\}^T$, where a 1 signals a potential shortcut at this time step. Thereby, masks $a(\mathbf{x}) = (0, \dots, 0)^T$ corresponds to no feedback for a sample. Similarly, feedback can also be given on the frequency explanation, marking which elements in the frequency domain are potential shortcuts. The resulting feedback mask $\hat{a}(\mathbf{x}) = (\hat{a}(\mathbf{x})_{re}, \hat{a}(\mathbf{x})_{im})$ can be

different for the real $\hat{a}(\mathbf{x})_{re} \in \{0, 1\}^T$ and imaginary part $\hat{a}(\mathbf{x})_{im} \in \{0, 1\}^T$. For the whole dataset, we then have spatial annotations A and frequency annotations \hat{A} . Obtaining annotated feedback masks can become costly, particularly if the feedback comes from human experts. However, as shortcuts often occur systematically, it can be possible to apply annotations to many samples, drastically reducing the number of annotations required in practice.

3.3 Revise

The last step of RioT is integrating the feedback into the model. We apply the general idea of using a loss-based model revision [27,30,35] based on the explanations and the annotation mask. Given the input data $(\mathcal{X}, \mathcal{Y})$, we define the original task (or right-answer) loss as $\mathcal{L}_{RA}(\mathcal{X}, \mathcal{Y})$. This loss measures the model performance and is the primary learning objective. To incorporate the feedback, we further use the right-reason loss $\mathcal{L}_{RR}(A, E)$. This loss aligns model explanations $E = \{e(\mathbf{x}) | \mathbf{x} \in \mathcal{X}\}$ and user feedback A by penalizing the model for explanations in the annotated areas. To achieve model revision and a good task performance, both losses are combined, where λ is a hyperparameter to balance both parts of the combined loss $\mathcal{L}(\mathcal{X}, \mathcal{Y}, A, E) = \mathcal{L}_{RA}(\mathcal{X}, \mathcal{Y}) + \lambda \mathcal{L}_{RR}(A, E)$. Together, the combined loss simultaneously optimizes the primary training objective (e.g. accuracy) and feedback alignment.

Time Domain Feedback. Masking parts of the time domain is an easy way to mitigate spatially locatable shortcuts (Fig. 1, left). We use the explanations E and annotations A in the spatial version of the right-reason loss:

$$\mathcal{L}_{RR}^{sp}(A, E) = \frac{1}{D} \sum_{\mathbf{x} \in \mathcal{X}} (e(\mathbf{x}) * a(\mathbf{x}))^2 \quad (3)$$

As the explanations and the feedback masks are element-wise multiplied, this loss minimizes the explanation values in marked parts of the input. This effectively trains the model to disregard the marked parts for its computation. Thus, using the loss in Eq. 3 as right-reason component for the combined loss allows to effectively steer the model away from points or intervals in time.

Frequency Domain Feedback. However, feedback in the time domain is insufficient to handle every type of shortcut. If the shortcut is not locatable in time, giving spatial feedback cannot be used to revise the models' behavior. Therefore, we utilize explanations and feedback in the frequency domain to tackle shortcuts like in Fig. 1, (right). Given the frequency explanations \hat{E} and annotations \hat{A} , the right-reason loss for the frequency domain is:

$$\mathcal{L}_{RR}^{fr}(\hat{A}, \hat{E}) = \frac{1}{D} \sum_{\mathbf{x} \in \mathcal{X}} \left((\text{Re}(\hat{e}(\mathbf{x})) * \hat{a}_{re}(\mathbf{x}))^2 + (\text{Im}(\hat{e}(\mathbf{x})) * \hat{a}_{im}(\mathbf{x}))^2 \right) \quad (4)$$

The Fourier transformation is invertible and differentiable, so we can backpropagate gradients to parameters directly from this loss. Intuitively, the frequency right-reason loss causes the masked frequency explanations of the model to be small while not affecting any specific point in time.

Depending on the problem, it is possible to use RioT either in the spatial or frequency domain. Moreover, it is also possible to combine feedback in both domains, including two right-reason terms in the final loss. This results in two parameters λ_1 and λ_2 to balance the right-answer and both right-reason losses.

$$\mathcal{L}(\mathcal{X}, \mathcal{Y}, A, E) = \mathcal{L}_{\text{RA}}(\mathcal{X}, \mathcal{Y}) + \lambda_1 \mathcal{L}_{\text{RR}}^{sp}(A, E) + \lambda_2 \mathcal{L}_{\text{RR}}^{fr}(\hat{A}, \hat{E}) \quad (5)$$

It is important to note that giving feedback in the frequency domain allows a new form of model revision through XIL. Even if we effectively still apply masking in the frequency domain, the effect in the original input domain is entirely different. Masking out a single frequency affects all time points without preventing the model from looking at any of them. In general, having an invertible transformation from the input domain to a different representation allows to give feedback more flexible than before. The Fourier transformation is a prominent example but not the only possible choice for this. Using other transformations like wavelets [12], is also possible.

Computational Costs. Including RioT in the training of a model increases the computational cost. Computing the right reason loss term requires the computation of a mixed partial derivative: $\frac{\partial^2 f_\theta(x)}{\partial \theta \partial x}$. Even though this is a second-order derivative, it does not result in any substantial cost increases, as the second-order component of our loss can be formalized as a Hessian-vector product (cf. Appx. A.3), which is known to be fast to compute [20]. We also observed this in our experimental evaluation, as even the naive implementation of our loss in PyTorch scales to large models.

Source of Feedback. A key aspect of RioT is the feedback incorporated in the *Obtain* step, which can come from various sources, including automated methods, rule-based systems, foundation models, or human annotations. Automated approaches, such as rule-based heuristics or pre-trained foundation models, provide scalable and consistent feedback, reducing the reliance on manual labeling. However, human annotations remain valuable for ensuring accuracy, especially in complex cases where automated methods may introduce biases or fail to capture nuanced patterns. RioT is agnostic to the feedback source, allowing flexibility in its application.

4 Shortcuts in Time Series

Shortcuts, like those in images, naturally occur in time series data but are often less apparent. Developing effective mitigation methods requires datasets where shortcuts are explicitly annotated, yet no existing datasets provide such annotations, despite known biases in popular datasets [2]. To address this gap, we introduce several time series dataset decoy variants inspired by prior work on decoy data [27], allowing for controlled evaluation of shortcut mitigation strategies. To further evaluate shortcut mitigation under real-world conditions, we present P2S, a real-world dataset where shortcuts arise from sensor recording processes.

4.1 Decoy Shortcuts

Classification. For both spatial and frequency cases, we introduce the shortcut as a class-specific pattern embedded in each training sample. The **spatial** pattern *replaces* m time steps with a sine wave defined as $s := \sin(t \cdot (2 + j)\pi) \cdot A$ where $t \in \{0, 1, \dots, m\}$ are the respective time steps, j represents the class index and A is the amplitude. In contrast, the **frequency** pattern is a similar sine wave, but it is *additively* applied to the full sequence ($m = T$).

Forecasting. For forecasting datasets, spatial decoys are more challenging due to window-based sampling and the complexity of the target. To address this, we design the shortcut as a "back-copy" of the forecast, where the decoy is equivalent to the actual solution. Due to the windowed sampling, the shortcut appears in every second sample. Given a sample of *lookback window* $\mathbf{x} \in \mathbb{R}^T$ and the *forecast horizon* $\mathbf{y} \in \mathbb{R}^W$. In the shortcut samples, we overwrite the first W entries of the lookback window with the future horizon values, yielding:

$$\mathbf{x}^s = \left(\underbrace{y_1, \dots, y_W}_{\text{future horizon}}, \underbrace{x_{W+1}, \dots, x_T}_{\text{remaining lookback}} \right).$$

while y remains unchanged. While this setting may seem constructed, similar patterns can emerge in real-world scenarios. For instance, in data transmission, glitches such as packet losses or duplications can subtly introduce irregularities into time series data, inadvertently creating forecasting shortcuts.

We model the forecasting frequency decoy as a recurring Dirac impulse with a specific frequency, added every k time steps: $i \in \{n \cdot k | n \in \mathbb{N} \wedge n \cdot k \leq T + W\}$ with a strength of A : $\text{interference} := A \cdot \Delta_i$. The impulse is present within the lookback and forecasting window during training, representing an effective decoy distracting the model from the actual forecast.

4.2 Real-World Shortcuts: Production Press Sensor Data (P2S)

RioT aims to mitigate shortcuts in time series data. While the decoys above provide a controlled evaluation setting, they do not capture the complexity of real-world shortcuts. To rectify this, we introduce PRODUCTION PRESS SENSOR DATA (P2S)⁶, a dataset of sensor recordings with naturally occurring shortcuts.

The sensor data stems from a high-speed press production line for metal parts, one of the sheet metal working industry's most economically significant processes. Based on the sensor data, the task is to predict whether a run is defective. The recordings include different production speeds, which, although not affecting part quality, influence process friction, and applied forces. Fig. 3 shows samples recorded at different speeds from normal and defect runs, highlighting variations even within the same class. A domain expert identified regions in the time series that vary with production speed, potentially distracting models from relevant features, especially when no defect and normal runs of the same speed

⁶ <https://huggingface.co/datasets/AIML-TUDA/P2S>

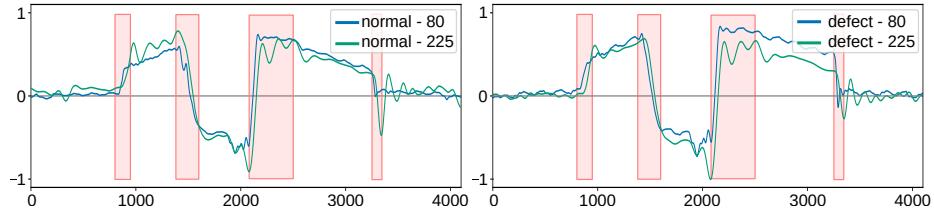


Fig. 3. Samples of P2S from normal (left) and defect (right) class at 80 and 225 strokes per minute. Areas of the time series that are especially sensitive to the stroke rate are considered a shortcut and marked red.

are in the training data. In these cases, the run’s speed is a shortcut and makes it difficult to generalize to other speeds than those present in training. P2S also includes a specifically curated setup that matches run speeds during training to avoid the shortcut. Further details on the dataset are available in Sec. B.

5 Experimental Evaluations

In this section, we investigate the effectiveness of RioT⁷ to mitigate shortcuts in time series classification and forecasting, including revision in the spatial domain (RioT_{sp}) and the frequency domain (RioT_{freq}), as well as both jointly.

5.1 Experimental Setup

Data. For classification, we use datasets from the UCR/UEA repository [7]. We select available datasets of a minimal size (cf. Appx. A.2), which results in FAULT DETECTION A, FORD A, FORD B, and SLEEP. For time series forecasting, we evaluate on three popular datasets from the Darts repository [14]: ETTM1, ENERGY, and WEATHER. We split the data into training and test sets using a 70/30 ratio and 20% of the training set are used for validation. We apply the previously described decoys to the training sets and simulate feedback based on the shortcuts to generate annotation masks. In the real-world experiment, we utilize our newly introduced dataset P2S. The mask is applied to all samples except in our feedback scaling experiment. For the real-world test case, we consider our newly introduced dataset P2S. We standardize all datasets as suggested by [41], i.e., rescaling the distribution to zero mean and a standard deviation of one.

Models. For time series classification, we use the FCN model of [19], with a slightly modified architecture for Sleep to achieve better performance (cf. Appx. A.1). Additionally, we use the OFA model [43]. For forecasting, we use TiDE [6], PatchTST [24] and NBEATS [25] to highlight the applicability of our method to a variety of model classes.

Metrics. In our evaluations, we compare model performance on datasets with and without shortcuts, as well as with and without RioT. For classification,

⁷ <https://github.com/ml-research/RioT>

Table 1. Applying RioT mitigates shortcuts in time series classification. Performance before and after applying RioT for spatial (Base_{sp}) and frequency ($\text{Base}_{\text{freq}}$) shortcuts. High training and low test accuracies indicate overfitting to the shortcut, which RioT successfully mitigates. *No Shortcut* represents the ideal scenario where the model is not affected by any shortcut.

Model	Config (ACC \uparrow)	Fault Detection A		FordA		FordB		Sleep	
		Train	Test	Train	Test	Train	Test	Train	Test
FCN	No Shortcut	0.99 \pm 0.00	0.99 \pm 0.00	0.92 \pm 0.01	0.91 \pm 0.00	0.93 \pm 0.00	0.76 \pm 0.01	0.68 \pm 0.00	0.62 \pm 0.00
	Base_{sp} + RioT _{sp}	1.00 \pm 0.00	0.74 \pm 0.06	1.00 \pm 0.00	0.71 \pm 0.08	1.00 \pm 0.00	0.63 \pm 0.03	1.00 \pm 0.00	0.10 \pm 0.03
	$\text{Base}_{\text{freq}}$ + RioT _{freq}	0.98 \pm 0.01	0.87 \pm 0.03	0.98 \pm 0.00	0.73 \pm 0.01	0.99 \pm 0.01	0.60 \pm 0.01	0.98 \pm 0.00	0.27 \pm 0.02
OFA	No Shortcut	1.00 \pm 0.00	0.98 \pm 0.02	0.92 \pm 0.01	0.87 \pm 0.04	0.95 \pm 0.01	0.70 \pm 0.04	0.69 \pm 0.00	0.64 \pm 0.01
	Base_{sp} + RioT _{sp}	1.00 \pm 0.00	0.53 \pm 0.02	1.00 \pm 0.00	0.50 \pm 0.00	1.00 \pm 0.00	0.52 \pm 0.01	1.00 \pm 0.00	0.21 \pm 0.05
	$\text{Base}_{\text{freq}}$ + RioT _{freq}	1.00 \pm 0.00	0.72 \pm 0.02	1.00 \pm 0.00	0.65 \pm 0.01	1.00 \pm 0.00	0.56 \pm 0.02	0.99 \pm 0.00	0.24 \pm 0.03

we report balanced (multiclass) accuracy (ACC), and mean squared error (MSE) for forecasting. The respective mean absolute error (MAE) results can be found in Appx. A.5. We report average and standard deviation over 5 runs.

5.2 Evaluations

Removing Shortcuts for Time Series Classification. We evaluate the effectiveness of RioT (spatial: RioT_{sp}, frequency: RioT_{freq}) in addressing shortcuts in classification tasks by comparing balanced accuracy with and without RioT. As shown in Tab. 1, without RioT, both FCN and OFA overfit to shortcuts, achieving \approx 100% training accuracy, while having poor test performance. Applying RioT significantly improves test performance for both models across all datasets. In some cases, RioT even reaches the performance of the ideal reference (no shortcut) scenario as if there would be no shortcut in the data. Even on FordB, where the drop in training-to-test performance highlights the distribution shift of that dataset [2], RioT_{sp} is still beneficial. Similarly, RioT_{freq} enhances performance on data with frequency shortcuts, though the improvement is less pronounced for FCN on Ford B, suggesting essential frequency information is sometimes obscured by RioT_{freq}. In summary, RioT successfully mitigates shortcuts in both domains, enhancing test generalization for FCN and OFA models.

Removing Shortcuts for Time Series Forecasting. Shortcuts are not exclusive to time series classification and can significantly impact other tasks, such as forecasting. In Tab. 2, we outline that spatial shortcuts cause models to overfit, but applying RioT_{sp} reduces MSE across datasets, especially for Energy, where MSE drops by up to 56%. In the frequency-shortcut setting, the training data includes a recurring Dirac impulse as a decoy (cf. Appx. A.4 for details). RioT_{freq} alleviates this distraction and improves the test performance significantly. For example, TiDE’s test MSE on ETTM1 decreases by 14% compared to the decoy setting.

Table 2. RioT can successfully overcome shortcuts in time series forecasting. MSE values (MAE values cf. Tab. 7) on the training set with and test set without shortcuts. *No Shortcut* is the ideal scenario where the model is not affected by shortcuts.

Model	Config (MSE \downarrow)	ETTM1		Energy		Weather	
		Train	Test	Train	Test	Train	Test
NBEATS	No Shortcut	0.30 \pm 0.02	0.47 \pm 0.02	0.34 \pm 0.03	0.26 \pm 0.02	0.08 \pm 0.01	0.03 \pm 0.01
	Base _{sp} + RioT _{sp}	0.24 \pm 0.01 0.30 \pm 0.01	0.55 \pm 0.01 0.50 \pm 0.01	0.33 \pm 0.03 0.45 \pm 0.03	0.94 \pm 0.02 0.58 \pm 0.01	0.09 \pm 0.01 0.11 \pm 0.01	0.16 \pm 0.04 0.09 \pm 0.02
	Base _{freq} + RioT _{freq}	0.30 \pm 0.02 0.31 \pm 0.02	0.46 \pm 0.01 0.45 \pm 0.01	0.33 \pm 0.04 0.33 \pm 0.04	0.36 \pm 0.04 0.34 \pm 0.04	0.11 \pm 0.02 0.81 \pm 0.48	0.32 \pm 0.09 0.17 \pm 0.01
PatchTST	No Shortcut	0.46 \pm 0.03	0.47 \pm 0.01	0.26 \pm 0.01	0.23 \pm 0.00	0.26 \pm 0.03	0.08 \pm 0.01
	Base _{sp} + RioT _{sp}	0.40 \pm 0.02 0.40 \pm 0.03	0.55 \pm 0.01 0.53 \pm 0.01	0.29 \pm 0.01 0.44 \pm 0.00	0.96 \pm 0.03 0.45 \pm 0.01	0.20 \pm 0.03 0.55 \pm 0.20	0.19 \pm 0.01 0.14 \pm 0.01
	Base _{freq} + RioT _{freq}	0.45 \pm 0.03 0.91 \pm 0.07	0.91 \pm 0.16 0.66 \pm 0.04	0.04 \pm 0.00 0.71 \pm 0.10	0.53 \pm 0.05 0.38 \pm 0.06	0.63 \pm 0.09 0.96 \pm 0.02	0.24 \pm 0.04 0.17 \pm 0.00
TiDE	No Shortcut	0.27 \pm 0.01	0.47 \pm 0.01	0.27 \pm 0.01	0.26 \pm 0.02	0.25 \pm 0.02	0.03 \pm 0.00
	Base _{sp} + RioT _{sp}	0.22 \pm 0.01 0.23 \pm 0.01	0.54 \pm 0.03 0.48 \pm 0.01	0.28 \pm 0.01 0.53 \pm 0.02	1.19 \pm 0.03 0.52 \pm 0.02	0.22 \pm 0.03 0.25 \pm 0.03	0.15 \pm 0.01 0.11 \pm 0.01
	Base _{freq} + RioT _{freq}	0.06 \pm 0.01 0.07 \pm 0.01	0.69 \pm 0.08 0.49 \pm 0.07	0.07 \pm 0.01 0.07 \pm 0.01	0.34 \pm 0.08 0.21 \pm 0.02	0.79 \pm 0.09 1.12 \pm 0.36	0.31 \pm 0.09 0.22 \pm 0.01

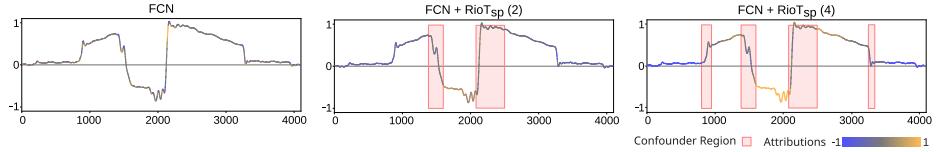


Fig. 4. Applying RioT lets the model ignore shortcut areas. While FCN primarily focuses on shortcuts, applying RioT with partial feedback (middle) or full feedback (bottom) causes the model to ignore the shortcut and focus on the remaining input.

In general, RioT effectively addresses spatial and frequency shortcuts in forecasting tasks. Interestingly, for TiDE on the Energy dataset, the performance with RioT_{freq} even surpasses the no shortcut model. Here, the added frequency acts as a form of data augmentation, enhancing model robustness. A similar behavior can be observed for NBEATS and ETTM1, where the decoy setting actually improves the model slightly, and RioT even improves upon that.

Removing Shortcuts in the Real-World. So far, our experiments have demonstrated RioT’s ability to counteract shortcuts within controlled environments. However, real-world shortcuts, as in our new dataset P2S, often have more complex structures. The model explanations in Fig. 4 (top) reveal a focus on distinct regions of the sensor curve, specifically the two middle regions. With domain knowledge, it is clear that these regions should not affect the model’s output. By applying RioT, we can redirect the model’s attention away from these regions. New model explanations highlight that the model still focuses on (other) incorrect regions, which can be mitigated by extending the annotated area. In Tab. 3, the model performance (exemplarily with FCN) in these settings is pre-

Table 3. Applying RioT overcomes shortcuts in P2S. Performance on the train set with and test set without shortcuts. FCN learns the train shortcut, resulting in lower test performance. Applying RioT with partial feedback (2) already yields good improvements, while adding feedback on the full shortcut area (4) is even better. *No Shortcut* is the ideal scenario, specifically curated so that there is no shortcut.

P2S (ACC \uparrow)	Train	Test
FCN _{No Shortcut}	0.97 \pm 0.01	0.95 \pm 0.01
FCN _{sp}	0.99 \pm 0.01	0.66 \pm 0.14
FCN _{sp} + RioT _{sp} (2)	0.96 \pm 0.01	0.78 \pm 0.05
FCN _{sp} + RioT _{sp} (4)	0.95 \pm 0.01	0.82 \pm 0.06

Table 4. RioT can combine spatial and frequency feedback. If the data contains time and frequency shortcuts, RioT can combine feedback on both domains to mitigate them, superior to feedback on only one domain. *No Shortcut* represents the ideal scenario when the model is not affected by any shortcuts.

	Sleep (Classification ACC \uparrow)	Train	Test
FCN _{No Shortcut}		0.68 \pm 0.00	0.62 \pm 0.00
FCN _{freq,sp}	1.00 \pm 0.00	0.10 \pm 0.04	
FCN _{freq,sp} + RioT _{sp}	0.94 \pm 0.00	0.24 \pm 0.02	
FCN _{freq,sp} + RioT _{freq}	1.00 \pm 0.00	0.04 \pm 0.00	
FCN _{freq,sp} + RioT _{freq,sp}	0.47 \pm 0.00	0.48 \pm 0.03	

	Energy (Forecasting MSE \downarrow)	Train	Test
TiDE _{No Shortcut}		0.28 \pm 0.01	0.26 \pm 0.02
TiDE _{freq,sp}	0.16 \pm 0.01	0.74 \pm 0.02	
TiDE _{freq,sp} + RioT _{sp}	0.20 \pm 0.01	0.61 \pm 0.02	
TiDE _{freq,sp} + RioT _{freq}	0.22 \pm 0.01	0.55 \pm 0.02	
TiDE _{freq,sp} + RioT _{freq,sp}	0.25 \pm 0.01	0.47 \pm 0.01	

sented. Without RioT, the model overfits to the shortcut. the test performance improves already with partial feedback (2) and even more with full feedback (4). These results highlight the effectiveness of RioT in real-world scenarios where not all shortcuts are initially known.

Removing Multiple Shortcuts at Once. In the previous experiments, we illustrated that RioT is suitable for addressing individual shortcuts, whether spatial or frequency-based. However, real-world time series data often presents a blend of multiple shortcuts that simultaneously influence model performance. Thus, we investigate the impact of applying RioT to both spatial and frequency shortcuts simultaneously (cf. Tab. 4), exemplarily using FCN and TiDE. When Sleep contains shortcuts in both domains, FCN without RioT overfits and fails to generalize. Addressing only one shortcut does not mitigate the effects, as the model adapts to the other. However, combining the respective feedback from both domains (RioT_{freq,sp}) significantly improves test performance, matching the frequency-shortcut scenario (cf. Tab. 1). Tab. 4 (bottom) shows the impact of multiple shortcuts on the Energy dataset, where the lower training MSE indicates overfitting. While applying either spatial or frequency feedback individually shows some effect, utilizing both types of feedback simultaneously (RioT_{freq,sp}) results in the largest improvements, as both decoys are addressed. The performance gap between RioT_{freq,sp} and the no shortcut setting is more pronounced than in single shortcut cases (cf. Tab. 2). This highlights again the known challenge of removing multiple shortcuts at once, which is generally more complex than individual shortcuts [36].

Handling Feedback. As the annotations are a crucial component of RioT, we conduct two different ablation studies to evaluate its impact using the classification data set Fault Detection A and the forecasting data set Energy. The first experiment examines the required amount of feedback, while the second assesses

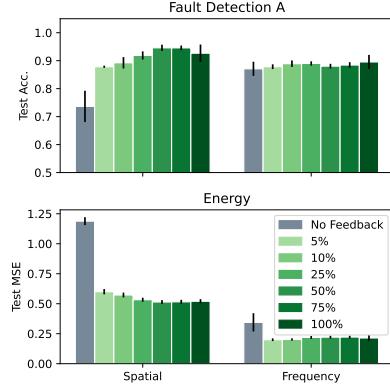


Fig. 5. RioT uses feedback efficiently. Even with feedback on only a small percentage of the data, RioT can overcome shortcuts.

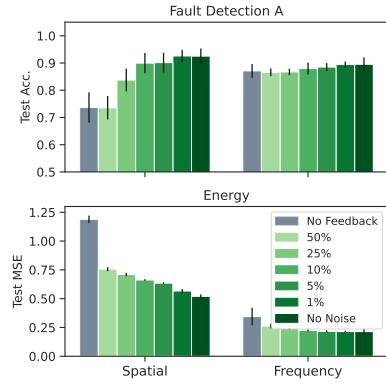


Fig. 6. RioT is robust against invalid feedback. Even with some percentage of random feedback, RioT overcomes the shortcuts.

robustness to noisy feedback. In particular if the feedback stems from domain experts, making excessive feedback requests is impractical. Thus our first experiment evaluates the performance of RioT when feedback is provided on only a portion of the dataset (Fig. 5). The findings reveal that full annotations are unnecessary. Even with minimal feedback, such as annotating just 5% of the samples, RioT significantly outperforms scenarios with no feedback. While previous experiments assumed entirely accurate feedback, real-world applications often involve some degree of error. Therefore, we test the resilience of RioT to increasing levels of incorrect feedback (Fig. 6). Instead of accurately marking shortcut areas, random time steps or frequency components are incorrectly labeled as shortcuts. The results show that RioT maintains strong performance even with up to 10% invalid feedback, presenting only slight performance declines. In certain cases, like forecasting with spatial shortcuts, RioT can still achieve notable improvements despite high levels of feedback noise. To further evaluate whether annotations in different settings can also be incorporated via RioT, we conduct an additional ablation where the feedback is based on shaplets instead of the input domain directly (cf. Tab. 11 in the appendix). The results show that RioT can be effective in this setup as well, and is not limited to the specific explanation method and annotation modality shown in the other experiments (more details in Appx. A.5). In summary, RioT effectively generalizes from small subsets of feedback and remains robust against a moderate amount of annotation noise. Additionally, RioT can incorporate feedback in other settings with other types of explanations as well. These qualities demonstrate that RioT is well-equipped to manage the practical challenges associated with incorporating feedback.

Qualitative Insights into Model Encodings. Lastly, we examine the inner workings of a model by analyzing its latent representations under various configurations. In Fig. 7, t-SNE plots show OFA’s feature encodings on Fault

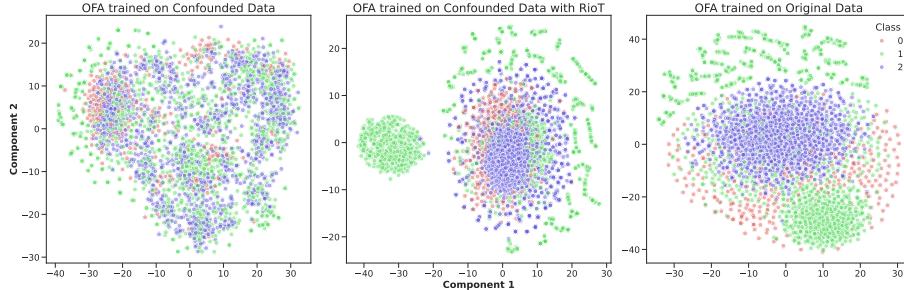


Fig. 7. t-SNE plots of OFA encodings for Fault Detection A. The left plot shows that a model trained with shortcuts shows minimal class separation. The middle plot shows the same setup but after RioT regularization, while the far right plot shows a model without shortcuts with clear class separation. Both RioT-regularized and model without shortcuts exhibit similar structures, highlighting the effectiveness of RioT.

Detection A in three settings: trained on shortcut data (left) with poor class separation; after RioT regularization (center), where structure and separation improve; and trained on clean data (right), yielding clear clusters. This reflects the scores of the models (Tab. 1): the shortcut model reaches $\approx 50\%$ accuracy, whereas RioT boosts it to nearly 100%, matching the reference scenario with clean data. This further demonstrates RioT’s ability to mitigate shortcuts and restore robust performance.

Limitations. A key aspect of RioT is the incorporation of feedback. While this is a major advantage of RioT, obtaining feedback can also present some challenges. Although we demonstrate that only a small fraction of annotated samples is needed, annotations remain essential. Moreover, like many interactive learning approaches, RioT assumes accurate feedback, making it important to consider potential issues from inaccuracies in practical applications. To reduce the need of manual feedback, one could explore automated feedback strategies instead or alongside manual feedback [34] (e.g., using an LLM to provide feedback or automated clustering of explanations to identify outliers). Such approaches may alleviate annotation costs but inevitably trade off some precision and can introduce new failure modes if the surrogate feedback is misaligned with task requirements. Another drawback of RioT is the increased training cost. Optimizing model explanations with gradient-based attributions requires computing a mixed-partial derivative. However, this can be efficiently handled using a Hessian-vector product, keeping the additional overhead manageable.

6 Conclusion

In this work, we present Right on Time (RioT) a method to mitigate shortcuts in time series data with the help of feedback. By revising the model, RioT significantly diminishes the influence of these factors, steering the model to align

with the correct reasons. Using popular time series models on several controlled decoy datasets and the newly introduced, real-world dataset P2S with naturally occurring shortcuts, showcases that SOTA models are indeed subject to shortcuts. Our results demonstrate that applying RioT to these models can mitigate shortcuts in the data. Furthermore, we have unveiled that addressing solely the time domain is insufficient for fully steering the model toward the correct reasons. To overcome this, we extended our method to incorporate feedback in the frequency domain, offering an additional mechanism for reducing reliance on shortcuts. Logical next steps are the extension of RioT to multivariate time series and the integration of various explainer types. Furthermore, exploring the usage of adaptive feedback mechanisms could prove to be beneficial, in particular in the context of multiple simultaneous shortcuts. Beyond time series, the application of RioT, especially RioT_{freq}, can also allow for a more nuanced approach to shortcut mitigation in other modalities.

Acknowledgments. This work was partly funded by the Federal Ministry of Education and Research (BMBF) project “XEI” (FKZ 01IS24079B), received funding by the EU project EXPLAIN, funded by the Federal Ministry of Education and Research (grant 01—S22030D). Additionally, it was funded by the project “The Adaptive Mind” from the Hessian Ministry of Science and the Arts (HMWK), the “ML2MT” project from the Volkswagen Stiftung, and the Priority Program (SPP) 2422 in the subproject “Optimization of active surface design of high-speed progressive tools using machine and deep learning algorithms” funded by the German Research Foundation (DFG). The latter also contributed the data for the P2S dataset. Furthermore, this work benefited from the HMWK project “The Third Wave of Artificial Intelligence – 3AI”.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE* **10**(7), e0130140 (2015)
2. Bagnall, A., Davis, L., Hills, J., Lines, J.: Transformation Based Ensembles for Time Series Classification. In: *Proceedings of the SIAM International Conference on Data Mining (SDM)* (2012)
3. Benidis, K., Rangapuram, S.S., Flunkert, V., Wang, Y., Maddix, D., Turkmen, C., Gasthaus, J., Bohlke-Schneider, M., Salinas, D., Stella, L., Aubet, F.X., Callot, L., Januschowski, T.: Deep Learning for Time Series Forecasting: Tutorial and Literature Survey. *ACM Computing Surveys* **55**(6), 1–36 (2023)
4. Bica, I., Alaa, A.M., Van Der Schaar, M.: Time series deconfounder: estimating treatment effects over time in the presence of hidden confounders. In: *Proceedings of the International Conference on Machine Learning (ICML)* (2020)
5. Cao, D., Enouen, J., Wang, Y., Song, X., Meng, C., Niu, H., Liu, Y.: Estimating treatment effects from irregular time series observations with hidden confounders. In: *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)* (2023)

6. Das, A., Kong, W., Leach, A., Mathur, S., Sen, R., Yu, R.: Long-term Forecasting with TiDE: Time-series Dense Encoder. *Transactions on Machine Learning Research (TMLR)* (2023)
7. Dau, H.A., Bagnall, A.J., Kamgar, K., Yeh, C.C.M., Zhu, Y., Gharghabi, S., Ratanamahatana, C.A., Keogh, E.J.: The UCR time series archive. *ArXiv:1810.07758* (2018)
8. Flanders, W.D., Klein, M., Darrow, L., Strickland, M., Sarnat, S., Sarnat, J., Waller, L., Winquist, A., Tolbert, P.: A Method for Detection of Residual Confounding in Time-Series and Other Observational Studies. *Epidemiology* **22**(1), 59–67 (2011)
9. Friedrich, F., Stammer, W., Schramowski, P., Kersting, K.: A typology for exploring the mitigation of shortcut behaviour. *Nature Machine Intelligence* **5**(3), 319–330 (2023)
10. Friedrich, F., Steinmann, D., Kersting, K.: One explanation does not fit XIL. *ArXiv:2304.07136* (2023)
11. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. *Nature Machine Intelligence* **2**(11), 665–673 (2020)
12. Graps, A.: An introduction to wavelets. *IEEE Computational Science and Engineering* **2**(2), 50–61 (1995)
13. Hatt, T., Feuerriegel, S.: Sequential deconfounding for causal inference with unobserved confounders. In: *Proceedings of the Conference on Causal Learning and Reasoning (CLeaR)* (2024)
14. Herzen, J., Lässig, F., Piazzetta, S.G., Neuer, T., Tafti, L., Raille, G., Pottelbergh, T.V., Pasieka, M., Skrodzki, A., Huguenin, N., Dumonal, M., Kościsz, J., Bader, D., Gusset, F., Benheddi, M., Williamson, C., Kosinski, M., Petrik, M., Grosch, G.: Darts: User-Friendly Modern Machine Learning for Time Series. *Journal of Machine Learning Research (JMLR)* **23**(124), 1–6 (2022)
15. Koprinska, I., Wu, D., Wang, Z.: Convolutional Neural Networks for Energy Time Series Forecasting. In: *Proceedings of the International Joint Conference on Neural Networks (IJCNN)* (2018)
16. Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., Müller, K.R.: Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications* **10**(1), 1096 (2019)
17. Lin, J., Keogh, E., Lonardi, S., Chiu, B.: A symbolic representation of time series, with implications for streaming algorithms. In: *Proceedings of the ACM SIGMOD workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD)* (2003)
18. Ma, C., Dai, G., Zhou, J.: Short-Term Traffic Flow Prediction for Urban Road Sections Based on Time Series Analysis and LSTM_bilstm Method. *IEEE Transactions on Intelligent Transportation (T-ITS)* **23**(6), 5615–5624 (2022)
19. Ma, Q., Liu, Z., Zheng, Z., Huang, Z., Zhu, S., Yu, Z., Kwok, J.T.: A survey on time-series pre-trained models. *ArXiv:2305.10716* (2023)
20. Martens, J.: Deep learning via hessian-free optimization. In: *Proceedings of the International Conference on Machine Learning (ICML)* (2010)
21. Mehdiyev, N., Lahann, J., Emrich, A., Enke, D., Fettke, P., Loos, P.: Time Series Classification using Deep Learning for Process Planning: A Case from the Process Industry. *Procedia Computer Science* **114**, 242–249 (2017)
22. Mercier, D., Bhatt, J., Dengel, A., Ahmed, S.: Time to Focus: A Comprehensive Benchmark Using Time Series Attribution Methods. *ArXiv:2202.03759* (2022)

23. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence (AIJ)* **267**, 1–38 (2019)
24. Nie, Y., H. Nguyen, N., Sinthong, P., Kalagnanam, J.: A time series is worth 64 words: Long-term forecasting with transformers. In: Proceedings of the International Conference on Learning Representations (ICLR) (2023)
25. Oreshkin, B.N., Carpov, D., Chapados, N., Bengio, Y.: N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. In: Proceedings of the International Conference on Learning Representations (ICLR) (2020)
26. Rojat, T., Puget, R., Filliat, D., Del Ser, J., Gelin, R., Díaz-Rodríguez, N.: Explainable Artificial Intelligence (XAI) on TimeSeries Data: A Survey. ArXiv:2104.00950 (2021)
27. Ross, A.S., Hughes, M.C., Doshi-Velez, F.: Right for the right reasons: Training differentiable models by constraining their explanations. In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI) (2017)
28. Ruiz, A.P., Flynn, M., Large, J., Middlehurst, M., Bagnall, A.: The great multivariate time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery (DMKD)* **35**(2), 401–449 (2021)
29. Schlegel, U., Arnout, H., El-Assady, M., Oelke, D., Keim, D.A.: Towards a Rigorous Evaluation of XAI Methods on Time Series. ArXiv:1909.07082 (2019)
30. Schramowski, P., Stammer, W., Teso, S., Brugger, A., Herbert, F., Shao, X., Luigs, H.G., Mahlein, A.K., Kersting, K.: Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nature Machine Intelligence* **2**(8), 476–486 (2020)
31. Selvaraju, R.R., Lee, S., Shen, Y., Jin, H., Ghosh, S., Heck, L., Batra, D., Parikh, D.: Taking a HINT: Leveraging Explanations to Make Vision and Language Models More Grounded. In: Proceedings of the International Conference on Computer Vision (ICCV) (2019)
32. Shao, X., Skryagin, A., Stammer, W., Schramowski, P., Kersting, K.: Right for Better Reasons: Training Differentiable Models by Constraining their Influence Functions. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI) (2021)
33. Shrikumar, A., Greenside, P., Shcherbina, A., Kundaje, A.: Not Just a Black Box: Learning Important Features Through Propagating Activation Differences. ArXiv:1605.01713 (2017)
34. Stammer, W., Friedrich, F., Steinmann, D., Brack, M., Shindo, H., Kersting, K.: Learning by self-explaining. *Transactions on Machine Learning Research (TMLR)* (2024)
35. Stammer, W., Schramowski, P., Kersting, K.: Right for the right concept: Revising neuro-symbolic concepts by interacting with their explanations. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
36. Steinmann, D., Divo, F., Kraus, M., Wüst, A., Struppek, L., Friedrich, F., Kersting, K.: Navigating Shortcuts, Spurious Correlations, and Confounders: From Origins via Detection to Mitigation. ArXiv:2412.05152 (2024)
37. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: Proceedings of the International Conference on Machine Learning (ICML) (2017)
38. Teso, S., Kersting, K.: Explanatory Interactive Machine Learning. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES) (2019)
39. Veerappa, M., Anneken, M., Burkart, N., Huber, M.F.: Validation of XAI explanations for multivariate time series classification in the maritime domain. *Journal of Computational Science* **58**, 101539 (2022)

40. Vielhaben, J., Lapuschkin, S., Montavon, G., Samek, W.: Explainable AI for Time Series via Virtual Inspection Layers. ArXiv:2303.06365 (2023)
41. Wu, H., Xu, J., Wang, J., Long, M.: Autoformer: Decomposition transformers with Auto-Correlation for long-term series forecasting. In: Proceedings of the Conference on Neural Information Processing Systems (NeurIPS) (2021)
42. Ye, L., Keogh, E.: Time series shapelets: a novel technique that allows accurate, interpretable and fast classification. Data Mining and Knowledge Discovery (DMKD) **22**, 149–182 (2011)
43. Zhou, T., Niu, P., Wang, X., Sun, L., Jin, R.: One fits all: Power general time series analysis by pretrained LM. In: Proceedings of the Conference on Neural Information Processing Systems (NeurIPS) (2023)

A Appendix

A.1 Implementation and Experimental Details

Adaption of Integrated Gradients (IG) A part of IG is a multiplication of the model gradient with the input itself, improving the explanation’s quality [33]. However, this multiplication makes some implicit assumptions about the input format. In particular, it assumes that there are no inputs with negative values. Otherwise, multiplying the attribution score with a negative input would flip the attribution’s sign, which is not desired. For images, this is unproblematic because they are always equal to or larger than zero. In time series, negative values can occur and normalization to make them all positive is not always suitable. To avoid this problem, we use only the input magnitude and not the input sign to compute the IG attributions.

Computing Explanations To compute explanations with Integrated Gradients, we followed the common practice of using a baseline of zeros. The standard approach worked well in our experiments, so we did not explore other baseline choices in this work. For the implementation, we utilized the widely-used Captum⁸ library, with patched `captum._utils.gradient.compute_gradients` function to allow for the propagation of the gradient with respect to the input to be propagated back into the parameters.

Model Training and Hyperparameters To find suitable parameters for model training, we performed a hyperparameter search over batch size, learning rate, and the number of training epochs. We then used these parameters for all model trainings and evaluations, with and without RioT. In addition, we selected suitable λ values for RioT with a hyperparameter selection on the respective validation sets. The exact values for the model training parameters and the λ values can be found in the provided code⁹.

To avoid model overfitting on the forecasting datasets, we performed shifted sampling with a window size of half the lookback window.

Code For the experiments, we based our model implementations on the following repositories:

- FCN: https://github.com/qianlima-lab/time-series-ptms/blob/master/model/tsm_model.py
- OFA: <https://github.com/DAMO-DI-ML/NeurIPS2023-One-Fits-All/>
- NBEATS: <https://github.com/unit8co/darts/blob/master/darts/models/forecasting/nbeats.py>
- TiDE: https://github.com/unit8co/darts/blob/master/darts/models/forecasting/tide_model.py

⁸ <https://github.com/pytorch/captum>

⁹ <https://github.com/ml-research/RioT>

- PatchTST: https://github.com/awslabs/gluonts/tree/dev/src/gluonts/torch/model/patch_tst

All experiments were executed using our Python 3.11 and PyTorch, which is available in the provided code. To ensure reproducibility and consistency, we utilized Docker. Configurations and Python executables for all experiments are provided in the repository.

Hardware To conduct our experiments, we utilized single GPUs from Nvidia DGX2 machines equipped with A100-40G and A100-80G graphics processing units. By maintaining a consistent hardware setup and a controlled software environment, we aimed to ensure the reliability and reproducibility of our experimental results.

A.2 UCR Dataset selection

We focused our evaluation on a subset of UCR datasets with a minimum size. Our selection process was as follows: First, we discarded all multivariate datasets, as we only considered univariate data in this paper. Then we removed all datasets with time series of different length or missing values. We further excluded all datasets of the category *SIMULATED*, to avoid datasets which were synthetic from the beginning. We furthermore considered only datasets with less than 10 classes, as having a per-class shortcuts on more than 10 classes would result in a very high number of different shortcuts, which would probably rarely happen. Besides these criteria, we discarded all datasets with less than 1000 training samples or a per sample length of less than 100, to avoid the small datasets of UCR, which leads to the resulting four datasets: Fault Detection A, Ford A, Ford B and Sleep.

A.3 Computational Costs of RioT

Training a model with RioT induces additional computational costs. The right-reason term requires computations of additional gradients. Given a model $f_\theta(x)$, parameterized by θ and input x , then computing the right reason loss with a gradient-based explanation method requires the computation of the mixed-partial derivative $\frac{\partial^2 f_\theta(x)}{\partial \theta \partial x}$, as a gradient-based explanation includes the derivative $\frac{\partial f_\theta(x)}{\partial x}$. While this mixed partial derivative is a second order derivative, this does not substantially increase the computational costs of our method for two main reasons. First, we are never explicitly materializing the Hessian matrix. Second, the second-order component of our loss can be formulated as a Hessian-vector product:

$$\frac{\partial \mathcal{L}}{\partial \theta} = g + \frac{\lambda}{2} H_{\theta x}(e(x) - a(x)) \quad (6)$$

where $g = \frac{\partial \mathcal{L}_{RA}}{\partial \theta}$ is the partial derivative of the right answer loss and if H is the full joint Hessian matrix of the loss with respect to θ and x , then $H_{\theta x}$ is the

sub-block of this matrix mapping x into θ (cf. Fig. 8), given by $H_{\theta x} = \frac{\partial^2 f_\theta(x)}{\partial \theta \partial x}$. Hessian-vector products are known to be fast to compute [20], enabling the right-reason loss computation to scale to large models and inputs.

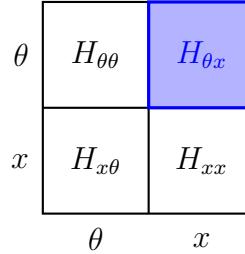


Fig. 8. Illustration of the Hessian matrix with its respective sub-blocks. The mapping from x into θ is highlighted in blue.

A.4 Details on Shortcuts

In the datasets which are not P2S, we added synthetic shortcuts to evaluate the effectiveness of shortcuts. Here, we provide some examples and more detailed descriptions of the decoy shortcuts. In Fig. 9 an example for a spatial decoy for classification is shown and in Fig. 10 an examples for the frequency decoy for classification.

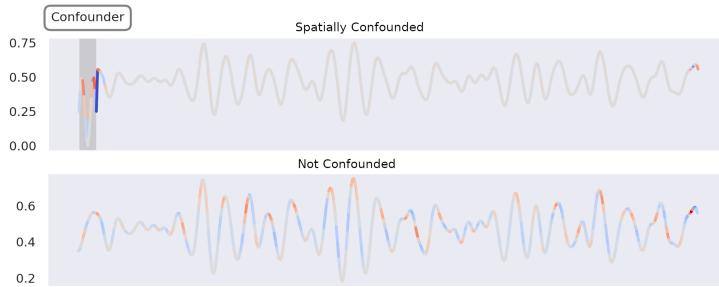


Fig. 9. Example of the added spatial shortcut in the Fault Detection dataset.

For forecasting datasets, spatial decoy are shortcuts that act as the actual solution to the forecasting problem. Periodically, data from the time series is copied back in time. This “back-copy” is a shortcut for the forecast, as it resembles the time steps of the forecasting window. Due to the windowed sampling from the time series, this shortcut occurs at every second sample. An exemplary

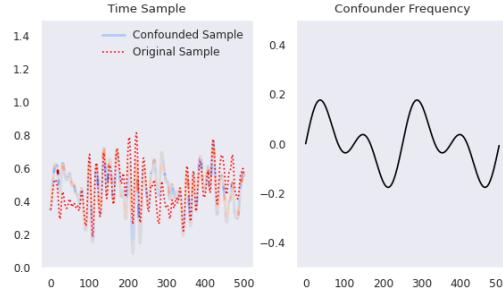


Fig. 10. Example of an added frequency shortcut in the Fault Detection dataset.

visualization of this setting is shown in Fig. 11, with an exemplary lookback length of 9, forecasting horizon of 3 and window stride of 6. This results in a shortcut in samples 1 and 3 (marked red) and overlapping in sample 2 (marked orange).

1. Sample		Lookback	Horizon
No Shortcut:		0 1 2 3 4 5 6 7 8	9 10 11
Shortcut:		9 10 11 3 4 5 6 7 8	9 10 11
Feedback:		1 1 1 0 0 0 0 0 0	
2. Sample			
No Shortcut:		6 7 8 9 10 11 12 13 14	15 16 17
Shortcut:		6 7 8 9 10 11 21 22 23	15 16 17
Feedback:		0 0 0 0 0 0 0 0 0	
3. Sample			
No Shortcut:		12 13 14 15 16 17 18 19 20	21 22 23
Shortcut:		21 22 23 15 16 17 18 19 20	21 22 23
Feedback:		1 1 1 0 0 0 0 0 0	
Shortcut:	■	Overlapping Shortcut	■

Fig. 11. Schematic overview of spatial decoys for the forecasting experiments.

Table 5. Feedback percentage for forecasting across all datasets, reported for the TiDE model. Corresponding to (test) results shown in Fig. 5, a higher percentage indicates more feedback, lower is better.

Metric	Feedback	ETTM1		Energy		Weather	
		Spatial	Freq	Spatial	Freq	Spatial	Freq
MAE (↓)	0%	0.54 ± 0.01	0.74 ± 0.06	0.85 ± 0.01	0.53 ± 0.07	0.29 ± 0.01	0.49 ± 0.09
	5%	0.52 ± 0.00	0.63 ± 0.03	0.62 ± 0.01	0.40 ± 0.02	0.28 ± 0.01	0.43 ± 0.03
	10%	0.52 ± 0.00	0.63 ± 0.03	0.61 ± 0.01	0.40 ± 0.02	0.27 ± 0.01	0.43 ± 0.03
	25%	0.52 ± 0.00	0.63 ± 0.03	0.58 ± 0.01	0.41 ± 0.01	0.25 ± 0.01	0.43 ± 0.04
	50%	0.52 ± 0.00	0.63 ± 0.03	0.57 ± 0.01	0.41 ± 0.01	0.24 ± 0.01	0.44 ± 0.05
	75%	0.52 ± 0.01	0.63 ± 0.03	0.57 ± 0.01	0.41 ± 0.01	0.24 ± 0.01	0.45 ± 0.06
	100%	0.51 ± 0.01	0.60 ± 0.05	0.58 ± 0.01	0.40 ± 0.03	0.24 ± 0.01	0.41 ± 0.02
MSE (↓)	0%	0.54 ± 0.03	0.69 ± 0.08	1.19 ± 0.03	0.34 ± 0.08	0.15 ± 0.01	0.31 ± 0.09
	5%	0.54 ± 0.01	0.52 ± 0.03	0.60 ± 0.02	0.20 ± 0.01	0.14 ± 0.01	0.24 ± 0.02
	10%	0.53 ± 0.01	0.52 ± 0.03	0.57 ± 0.02	0.20 ± 0.01	0.14 ± 0.01	0.24 ± 0.02
	25%	0.53 ± 0.01	0.52 ± 0.03	0.53 ± 0.02	0.22 ± 0.01	0.11 ± 0.01	0.24 ± 0.03
	50%	0.53 ± 0.01	0.52 ± 0.03	0.51 ± 0.02	0.22 ± 0.01	0.11 ± 0.01	0.25 ± 0.04
	75%	0.52 ± 0.01	0.51 ± 0.03	0.52 ± 0.02	0.22 ± 0.01	0.11 ± 0.01	0.26 ± 0.05
	100%	0.48 ± 0.01	0.49 ± 0.07	0.52 ± 0.02	0.21 ± 0.02	0.11 ± 0.01	0.22 ± 0.01

Table 6. Feedback (FB) percentage for classification across all datasets, reported for the FCN model. Corresponding to results shown in Fig. 5, a higher percentage indicates more feedback, higher is better.

FB	Fault Detection A (ACC ↑)		FordA (ACC ↑)		FordB (ACC ↑)		Sleep (ACC ↑)	
	Spatial	Freq	Spatial	Freq	Spatial	Freq	Spatial	Freq
0%	0.74 ± 0.06	0.87 ± 0.03	0.71 ± 0.08	0.73 ± 0.01	0.63 ± 0.03	0.60 ± 0.01	0.10 ± 0.03	0.27 ± 0.02
5%	0.88 ± 0.00	0.88 ± 0.01	0.81 ± 0.03	0.80 ± 0.03	0.66 ± 0.03	0.66 ± 0.02	0.53 ± 0.03	0.49 ± 0.00
10%	0.89 ± 0.02	0.89 ± 0.01	0.82 ± 0.04	0.79 ± 0.02	0.66 ± 0.03	0.64 ± 0.03	0.48 ± 0.09	0.48 ± 0.02
25%	0.92 ± 0.01	0.89 ± 0.01	0.83 ± 0.02	0.78 ± 0.01	0.67 ± 0.02	0.65 ± 0.01	0.49 ± 0.08	0.42 ± 0.08
50%	0.95 ± 0.01	0.88 ± 0.01	0.82 ± 0.03	0.81 ± 0.05	0.67 ± 0.02	0.65 ± 0.00	0.55 ± 0.03	0.44 ± 0.07
75%	0.95 ± 0.01	0.88 ± 0.01	0.81 ± 0.03	0.80 ± 0.04	0.65 ± 0.03	0.64 ± 0.00	0.54 ± 0.04	0.44 ± 0.07
100%	0.93 ± 0.03	0.90 ± 0.03	0.84 ± 0.02	0.83 ± 0.02	0.68 ± 0.02	0.65 ± 0.01	0.54 ± 0.05	0.45 ± 0.07

A.5 Additional Experimental Results

This section provides further insights into our experiments, covering both forecasting and classification tasks. Specifically, it showcases performance through various metrics such as MAE, MSE, and accuracy, qualitative insights about the influence of shortcuts, and explores different feedback configurations.

Feedback Generalization Tab. 6 and Tab. 5 detail provided feedback percentages for forecasting and classification across all datasets, respectively. These tables report the performance of the TiDE and FCN models, highlighting how different levels of feedback impact model outcomes on various datasets. Tab. 5 focuses on MAE and MSE for forecasting, while Tab. 6 presents ACC for classification.

Removing Shortcuts for Time Series Forecasting Tab. 7 reports the MAE results for our forecasting experiment across different models, datasets and configurations. It emphasizes how well each model performs on both the training

Table 7. RioT can successfully overcome shortcuts in time series forecasting. MAE values on the training set with and the test set without shortcuts. *No Shortcut* is the ideal scenario where the model is not affected by any shortcut.

Model	Config (MAE \downarrow)	ETTM1		Energy		Weather	
		Train	Test	Train	Test	Train	Test
NBEATS	No Shortcut	0.39 \pm 0.01	0.48 \pm 0.01	0.44 \pm 0.02	0.38 \pm 0.01	0.21 \pm 0.01	0.12 \pm 0.01
	Base _{sp} + RioT _{sp}	0.34 \pm 0.01 0.40 \pm 0.01	0.54 \pm 0.01 0.52 \pm 0.01	0.44 \pm 0.03 0.53 \pm 0.02	0.77 \pm 0.01 0.62 \pm 0.01	0.21 \pm 0.01 0.23 \pm 0.01	0.30 \pm 0.04 0.22 \pm 0.01
	Base _{freq} + RioT _{freq}	0.39 \pm 0.01 0.40 \pm 0.01	0.47 \pm 0.01 0.47 \pm 0.01	0.45 \pm 0.03 0.45 \pm 0.03	0.45 \pm 0.03 0.44 \pm 0.02	0.21 \pm 0.03 0.59 \pm 0.22	0.45 \pm 0.06 0.39 \pm 0.01
PatchTST	No Shortcut	0.50 \pm 0.01	0.49 \pm 0.01	0.39 \pm 0.00	0.38 \pm 0.01	0.38 \pm 0.03	0.18 \pm 0.00
	Base _{sp} + RioT _{sp}	0.46 \pm 0.00 0.46 \pm 0.01	0.53 \pm 0.01 0.52 \pm 0.01	0.41 \pm 0.00 0.51 \pm 0.00	0.78 \pm 0.01 0.53 \pm 0.01	0.32 \pm 0.04 0.54 \pm 0.12	0.33 \pm 0.00 0.28 \pm 0.00
	Base _{freq} + RioT _{freq}	0.53 \pm 0.01 0.92 \pm 0.05	0.81 \pm 0.07 0.80 \pm 0.02	0.15 \pm 0.00 0.97 \pm 0.86	0.64 \pm 0.03 0.57 \pm 0.02	0.58 \pm 0.03 0.65 \pm 0.01	0.41 \pm 0.05 0.40 \pm 0.01
TiDE	No Shortcut	0.36 \pm 0.01	0.48 \pm 0.01	0.40 \pm 0.01	0.38 \pm 0.02	0.36 \pm 0.02	0.13 \pm 0.00
	Base _{sp} + RioT _{sp}	0.32 \pm 0.01 0.34 \pm 0.01	0.54 \pm 0.01 0.51 \pm 0.01	0.40 \pm 0.01 0.57 \pm 0.01	0.85 \pm 0.01 0.58 \pm 0.01	0.32 \pm 0.03 0.35 \pm 0.03	0.29 \pm 0.01 0.24 \pm 0.01
	Base _{freq} + RioT _{freq}	0.18 \pm 0.01 0.19 \pm 0.01	0.74 \pm 0.06 0.60 \pm 0.05	0.18 \pm 0.01 0.18 \pm 0.01	0.53 \pm 0.07 0.40 \pm 0.03	0.65 \pm 0.05 0.79 \pm 0.16	0.49 \pm 0.09 0.41 \pm 0.02

set with and test set without shortcuts. Both setups are shown with and without applying RioT. The *No Shortcut* configuration representing the ideal scenario where the model is unaffected by any shortcuts.

Removing Multiple Shortcuts at Once Tab. 8 reports the MAE values and illustrates the effectiveness of combining spatial and frequency feedback via RioT for the TiDE model. The results demonstrate significant improvements in forecasting accuracy when integrating both feedback domains compared to using them separately.

Table 8. RioT can combine spatial and frequency feedback. MAE results when applying feedback in time and frequency with RioT. Combining both feedback domains is superior to feedback on only one of the domains. *No Shortcut* values represent the ideal scenario when the model is not affected by any shortcut (mean and std over 5 runs).

Energy (MAE \downarrow)	Train	Test
TiDE _{No Shortcut}	0.40 \pm 0.01	0.38 \pm 0.02
TiDE _{freq,sp}	0.30 \pm 0.01	0.70 \pm 0.02
TiDE _{freq,sp} + RioT _{sp}	0.34 \pm 0.01	0.64 \pm 0.01
TiDE _{freq,sp} + RioT _{freq}	0.36 \pm 0.01	0.60 \pm 0.01
TiDE _{freq,sp} + RioT _{freq,sp}	0.39 \pm 0.01	0.55 \pm 0.01

Early Stopping as Shortcut Mitigation Baseline In this experiment, we compare the performance of a model with RioT to a model regularized via early stopping (which is decided on a validation set without shortcuts). In that, we stop model training if there are no improvements in the validation set for several epochs in the hope that it thus does not overfit to the shortcut. The results are presented in Tab. 9 for classification and Tab. 10 for forecasting. We can observe that early stopping can help in some instances to achieve performances similar to RioT (e.g. PatchTST with a frequency shortcut or FCN with a spatial shortcut). However, for the majority of cases the performance with early stopping is substantially lower than the performance with RioT, signaling that early stopping alone is not a sufficient approach to overcome shortcuts.

Table 9. Early stopping on classification datasets.

Config	FCN		OFA	
	Train	Test	Train	Test
No Shortcut	0.99 \pm 0.00	0.99 \pm 0.00	1.00 \pm 0.00	0.98 \pm 0.02
Base _{sp}	1.00 \pm 0.00	0.74 \pm 0.06	1.00 \pm 0.00	0.53 \pm 0.02
+RioT _{sp}	0.98 \pm 0.01	0.93 \pm 0.03	0.96 \pm 0.08	0.98 \pm 0.01
+ES _{sp}	0.87 \pm 0.01	0.91 \pm 0.03	0.69 \pm 0.02	0.67 \pm 0.04
Base _{freq}	0.98 \pm 0.01	0.87 \pm 0.03	1.00 \pm 0.00	0.72 \pm 0.02
+RioT _{freq}	0.94 \pm 0.00	0.90 \pm 0.03	0.96 \pm 0.02	0.98 \pm 0.02
+ES _{freq}	0.83 \pm 0.01	0.86 \pm 0.02	0.81 \pm 0.01	0.75 \pm 0.02

Table 10. Early stopping on forecasting datasets.

Config	PatchTST		TiDE	
	Train	Test	Train	Test
No Shortcut	0.26 \pm 0.01	0.23 \pm 0.00	0.27 \pm 0.01	0.26 \pm 0.02
Base _{sp}	0.29 \pm 0.01	0.96 \pm 0.03	0.28 \pm 0.01	1.19 \pm 0.03
+RioT _{sp}	0.44 \pm 0.00	0.45 \pm 0.01	0.53 \pm 0.02	0.52 \pm 0.02
+ES _{sp}	0.48 \pm 0.05	0.68 \pm 0.03	1.20 \pm 0.25	0.81 \pm 0.08
Base _{freq}	0.04 \pm 0.00	0.53 \pm 0.05	0.07 \pm 0.01	0.34 \pm 0.08
+RioT _{freq}	0.71 \pm 0.10	0.38 \pm 0.06	0.07 \pm 0.01	0.21 \pm 0.02
+ES _{freq}	0.48 \pm 0.09	0.49 \pm 0.08	0.21 \pm 0.08	0.36 \pm 0.09

Incorporating Alternative Explanations based on Shapelets Attribution-based methods are commonly used to provide explanations at the individual

time-step level. However, such explanations can often appear fuzzy or ambiguous, leading to potential misinterpretations and complicating feedback. Humans naturally perceive and interpret shapes or patterns more intuitively than isolated time steps. To leverage this strength, we explored an alternative approach by extracting shapelets [42] from one of our benchmark datasets. In doing so, we demonstrate that our method’s effectiveness is not limited to deep learning models; it can also address shortcut learning in simpler, interpretable models based on shapelets.

Specifically, we extracted shapelets (cf. Fig. 12) from the dataset, trained a classifier based on these shapelets, and analyzed their attributions. Our observations revealed that the identified shortcut was captured within the activated shapelets. Crucially, we showed that our method (RioT) successfully mitigates this shortcut without compromising the activation of shapelets that represent meaningful features.

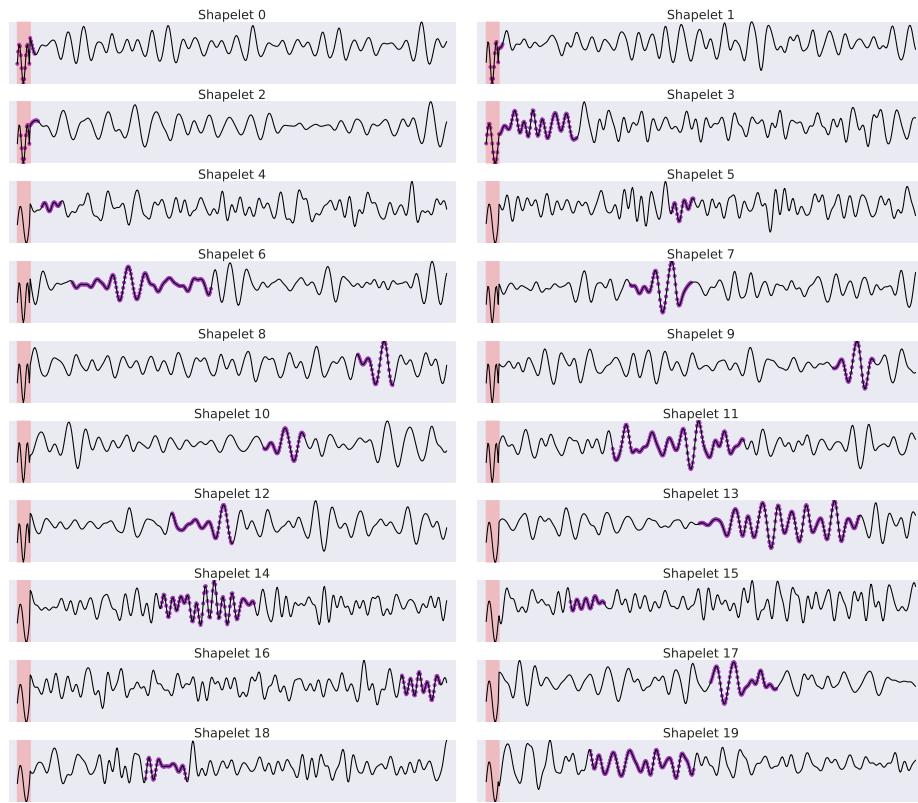


Fig. 12. Computed shapelets of FordA. The purple segments represent the discovered shapelets overlaid on their corresponding training samples. The red region marks the included shortcut, and shapelets 0–3 explicitly model this shortcut.

We observe that in the baseline model, the first four shapelets—which cover the shortcut (highlighted in red)—are predominantly activated. In contrast, with RioT, the activation shifts away from the shapelets containing the shortcut. This observation is further quantitatively outlined in Tab. 11, which summarizes the accuracy scores and aligns with our previous experimental results.

Table 11. Accuracy on FordA when solving the task based on shapelets.

Config	Train	Test
No Shortcut	0.90 \pm 0.00	0.89 \pm 0.00
Shapelets	0.99 \pm 0.00	0.72 \pm 0.02
+RioT _{sp}	0.92 \pm 0.00	0.82 \pm 0.00

B Dataset from a High-speed Progressive Tool with Natural Shortcuts

The presence of shortcuts is a common challenge in practical settings, affecting models in diverse ways. As the research community strives to identify and mitigate these issues, it becomes imperative to test our methodologies on datasets that mirror the complexities encountered in actual applications. However, for the time domain, datasets with explicitly labeled shortcuts are not present, highlighting the challenge of assessing model performance against the complex nature of practical shortcuts.

To bridge this gap, we introduce P2S, a dataset that represents a significant step forward by featuring explicitly identified shortcuts. This dataset originates from experimental work on a production line for deep-drawn sheet metal parts, employing a progressive die on a high-speed press. The sections below detail the experimental approach and the process of data collection.

B.1 Real-World setting

The production of parts in multiple progressive forming stages using stamping, deep drawing and bending operations with progressive dies is generally one of the most economically significant manufacturing processes in the sheet metal working industry and enables the production of complex parts on short process routes with consistent quality. For the tests, a four-stage progressive die was used on a Bruderer BSTA 810-145 high-speed press with varied stroke speed. The strip material to be processed is fed into the progressive die by a BSV300 servo feed unit, linked to the cycle of the press, in the stroke movement while the tools are not engaged. The part to be produced remains permanently connected to the sheet strip throughout the entire production run. The stroke height of

the tool is 63 mm and the material feed per stroke is 60 mm. The experimental setup with the progressive die set up on the high-speed press is shown in Fig. 13.

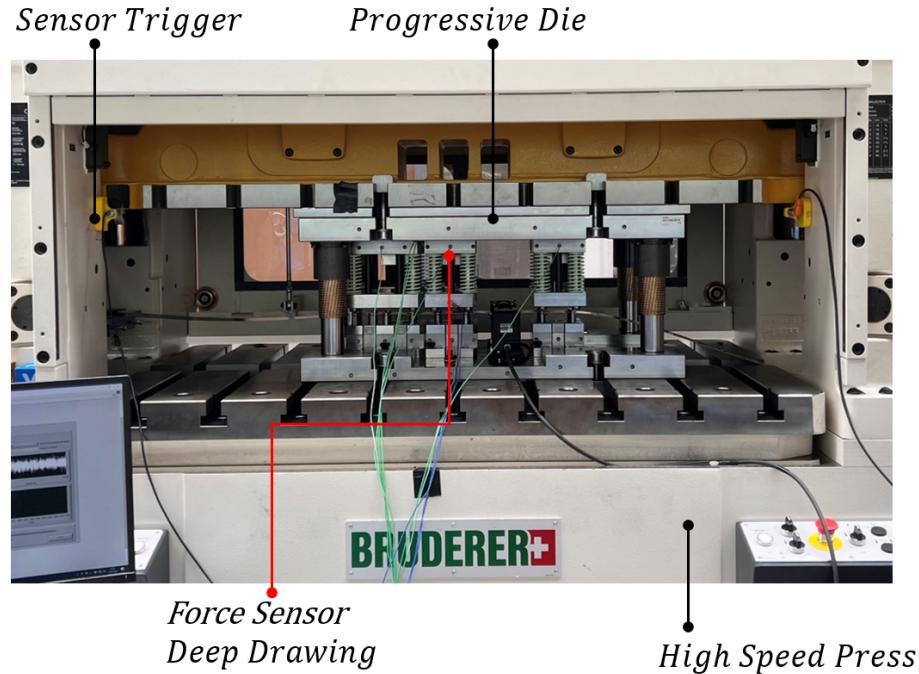


Fig. 13. Experimental setup with high-speed press and tool as well as trigger for stroke-by-stroke recording of the data

The four stages include a pilot punching stage, a round stamping stage, deep drawing and a cut-out stage. In the first stage, a 3 mm hole is punched in the metal strip. This hole is used by guide pins in the subsequent stages to position the metal strip. During the stroke movement, the pilot pin always engages in the pilot hole first, thus ensuring the positioning accuracy of the components. In the subsequent stage, a circular blank is cut into the sheet metal strip. This is necessary so that the part can be deep-drawn directly from the sheet metal strip. This is a round geometry that forms small arms in the subsequent deep-drawing step that hold the component on the metal strip. In the final stage, the component is then separated from the sheet metal strip and the process cycle is completed. The respective process steps are performed simultaneously so that each stage carries out its respective process with each stroke and therefore a part is produced with each stroke. Fig. 14 shows the upper tool unfolded and the forming stages associated with the respective steps on the continuous sheet metal strip.

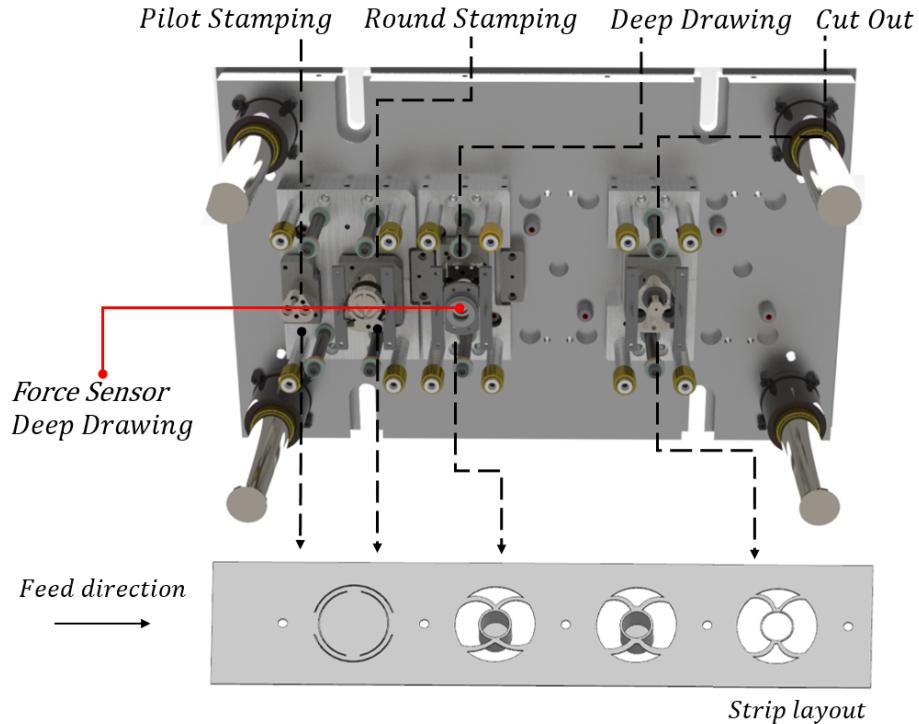


Fig. 14. Upper tool unfolded and the forming stages associated with the respective steps on the passing sheet metal strip as well as the positions of the piezoelectric force sensors.

B.2 Data collection

An indirect piezoelectric force sensor (Kistler 9240A) was integrated into the upper mould mounting plate of the deep-drawing stage for data acquisition. The sensor is located directly above the punch and records not only the indirect process force but also the blank holder forces which are applied by spring assemblies between the upper mounting plate and the blank holder plate. The data is recorded at a sampling frequency of 25 kHz. The material used is DC04 with a width of 50 mm and a thickness of 0.5 mm. The voltage signals from the sensors are digitised using a "CompactRIO" (NI cRIO 9047) with integrated NI 9215 measuring module (analogue voltage input ± 10 V). Data recording is started via an inductive proximity switch when the press ram passes below a defined stroke height during the stroke movement and is stopped again as it passes the inductive proximity switch during the return stroke movement. Due to the varying process speed caused by the stroke speeds that have been set, the recorded time series have a different number of data points. Further, there are slight variations in the length of the time series within one experiment. For this reason, all time series are interpolated to a length of 4096 data points and

we discard any time series that deviate considerably from the mean length of a run (i.e., threshold of 3). A total of 12 series of experiments, shown in Tab. 12, were carried out with production rates from 80 to 225 spm. To simulate a defect, the spring hardness of the blank holder was manipulated in the test series that were marked as *defect*. The manipulated experiments result in the component bursting and tearing during production. In a real production environment, this would lead directly to the parts being rejected.

B.3 Data characteristics

Fig. 15 shows the progression of the time series recorded with the indirect force sensor. The force curve characterises the process cycle during a press stroke. The measurement is started by the trigger which is activated by the ram moving downwards. The downholder plates touch down at point A and press the strip material onto the die. Between point A and point B, the downholder springs are compressed, causing the applied force to increase linearly. The deep drawing process begins at point B. At point C, the press reaches its bottom dead centre and the reverse stroke begins so that the punch moves out of the material again. At point D, the deep-drawing punch is released from the material and now the hold-down springs relax linearly up to point E. At point E, the downholder plate lifts off again and the component is fed to the last process step.

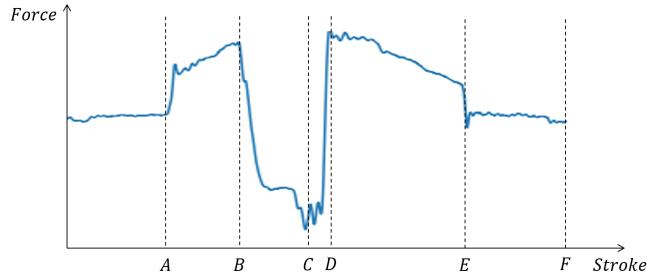


Fig. 15. Force curve for one stroke. A) set down downholder plate B) start of deep drawing C) bottom dead centre D) deep drawing process completed E) downholder plates lift off F) measurement stops.

B.4 Shortcuts

In P2S, the operation speed of the progressive tool acts as a shortcut to solve the task. The higher the stroke rate of the press, the more friction occurs, and the higher the impact of the downholder plate. The differences can be observed in Fig. 3. Since we are aware of these physics-based shortcuts, we are able to annotate them in our dataset. As the process speed increases, the friction that

Table 12. Overview of conducted runs on the high-speed press with normal and defect states at different stroke rates.

Experiment #	State	Stroke Rate	Samples
1	Normal	80	193
2	Normal	100	193
3	Normal	150	189
4	Normal	175	198
5	Normal	200	194
6	Normal	225	188
7	Defect	80	149
8	Defect	100	193
9	Defect	150	188
10	Defect	175	196
11	Defect	200	193
12	Defect	225	190
Total			2264

occurs between the die and the material in the deep-drawing stage changes, as the frictional force is dependent on the frictional speed. This is particularly evident in the present case, as there are no deep-drawing oils used in the experiments, which could optimize the friction condition. The areas affected by the punch's friction are from 1380 to 1600 (start of deep drawing) and from 2080 to 2500 (end of deep drawing). In addition, the impulse of the downholder plate affects the die increases due to increased process dynamics. If the process speed is increased, the process force also increases in the ranges of the time series from 800 to 950 (downholder plate sets down) and 3250 to 3550 (downholder plate lifts).

In the experiment setting of Tab. 3, the stroke rate in the training data correlates with the class label, i.e., there are only normal experiments with small stroke rates and defective ones with high stroke rates: Experiment 1, 2, 3, 10, 11, 12 are the training data and the remaining experiments are the test data. To obtain a reference setting without shortcuts, normal and defect experiments are in the same set (training or test). This results in experiments 1, 3, 5, 7, 9, 11 in the training set and the remaining in the test set.