Imbalanced Classification through the Lens of Spurious Correlations

Jakob Hackstein and Sidney Bender [*]

Machine Learning Group, TU Berlin, 10587 Berlin, Germany

Abstract. Class imbalance poses a fundamental challenge in machine learning, frequently leading to unreliable classification performance. While prior methods focus on data- or loss-reweighting schemes, we view imbalance as a data condition that amplifies Clever Hans (CH) effects by underspecification of minority classes. In a counterfactual explanations-based approach, we propose to leverage Explainable AI to jointly identify and eliminate CH effects emerging under imbalance. Our method achieves competitive classification performance on three datasets and demonstrates how CH effects emerge under imbalance, a perspective largely overlooked by existing approaches.

1 Introduction

Classification under imbalance is a long-standing challenge in machine learning and remains highly relevant due to its prevalence in real-world applications. Class imbalance destabilizes training, biases feature learning [1], and causes overfitting to majority classes, altogether limiting classifier reliability. Existing methods typically address these issues by implementing loss-reweighting schemes [2] to emphasize learning minority classes. While these approaches stabilize training, they leave a central challenge untouched: the minority class often provides insufficient data to accurately model its underlying semantics, leaving it fundamentally underspecified.

We propose to view class imbalance through the lens of spurious correlations, arguing that insufficient information on minority classes encourages seemingly discriminative yet non-causal classification strategies. Given their inductive bias to favor simple features [3] and tendency to rely on spurious cues, classifiers are prone to adopt Clever Hans (CH) solutions [4]. This perspective reveals a limitation in existing methods: they emphasize learning the correct *classification outcome* but do not explicitly encourage a causal *classification strategy*. Inspired by this insight, we aim to address imbalance by mitigating CH solutions that arise from minority classes. We study this approach in isolation by considering binary image classification tasks, which provide a controlled setting to analyze the interrelated effects of imbalance and spurious correlations.

Enforcing causal behavior is challenging, as classification strategies are deeply entangled and categorical annotations often ambiguous. However, recent advances in Explainable AI (XAI) have introduced various methods to analyze classifier behavior [5–7]. Therefore, to access and influence behavior beyond
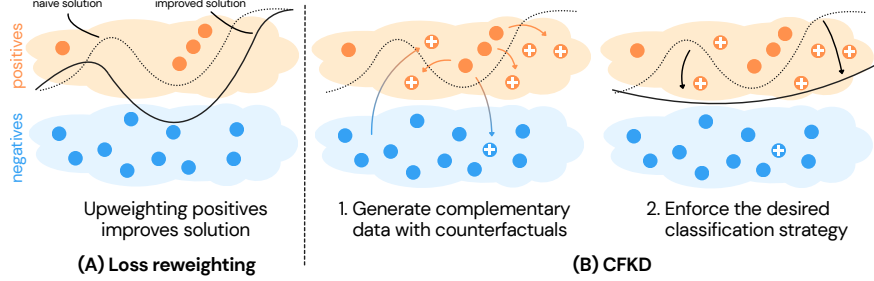
Fig. 1: **(A)** Upweighting positives yields a better decision boundary, but CH solutions persist. **(B)** CFKD first generates complementary data using counterfactuals. All counterfactuals cross the initial decision boundary, but may not flip the true class label (false counterfactuals). Then, the desired classification strategy is distilled into the classifier, explicitly prohibiting CH solutions.

mere outcome, we employ counterfactual explanations [8, 9]. Counterfactuals are interpretable, minimally altered samples that suffice to flip a classifier's prediction, thereby exposing the decisive features behind its decision. Our method builds on Counterfactual Knowledge Distillation (CFKD) [10, 11], which fine-tunes classifiers on domain expert-annotated counterfactuals. In contrast to reweighting schemes, this approach jointly uncovers CH effects and effectively eliminates spurious correlations, as illustrated in Figure 1. By extending the counterfactual-based CFKD framework, we take a first step toward harnessing XAI to improve classification performance under imbalance and jointly ensure trustworthy employment in safety-critical settings. Furthermore, we investigate the interplay between imbalance and spurious correlations through a series of experiments in both confounder-controlled and real-world settings.

2   Theoretical Motivation

To better understand how positive and negative samples corresponding to the minority and majority class, respectively, affect the emergence of spurious correlations, we examine a simplified example. Consider a dataset with n positives and an imbalance ratio $k \geq 1$, yielding nk negatives. Each sample has two binary features, a causal X_c aligned with class Y by construction, and a spurious X_s independent of class with $\Pr(X_s = 1 \mid Y = y) = 0.5$ for $y \in \{0, 1\}$.

We ascribe spurious correlations under imbalance to the underspecification of positives, as small n induces high variance in empirical estimates. For $n = 1$, X_s appears perfectly predictive as its pattern matches the single positive sample. More generally, the probability that X_s aligns perfectly with all n positive samples is 2^{1-n}, halving with every additional positive. Thus, with few positives, arbitrary features can mimic X_c and induce CH solutions. In contrast, negatives can refute the predictivity of non-causal features that small n alone might suggest. For X_s, the empirical negative prevalence follows $p_s \sim 1/nk \operatorname{Bin}(nk, 0.5)$.

As k increases, p_s concentrates around its true value 0.5, revealing X_s as uninformative. Additionally, surplus negatives introduce sample variability and promote learning task-relevant features.

However, in practice, increasing k quickly yields diminishing returns as severe imbalance causes practical issues during optimization and biases feature learning towards the majority class. Further, X_s may rarely occur in negatives (e.g., copyright tags [4]), limiting k's value in refuting CH solutions. Thus, while negatives assist in filtering futile features and refining representations, the trade-off between mitigating spurious correlations and avoiding severe imbalance allows CH effects to persist.

3   Method

Mitigating CH effects is a two-step process, as it involves (1) detecting reliance on spurious correlations and (2) subsequent removal of undesired feature reliance. In our setting, designing a suitable method is challenging: For (1), we must assume the emergence of multiple spurious correlations of arbitrary feature complexity. These undesired feature dependencies are not known a priori and complicate (2), as we cannot rely on confounder labels during training. Importantly, inherent ambiguity in the dataset limits the effectiveness of cost-sensitive techniques or data reweighting schemes, underscoring the need for a principled approach.

To account for these challenges, we utilize counterfactual explanations. Formally, for a given image-class pair (x, y) and a classifier f with $f(x) = y$, a counterfactual $\tilde{x}$ is a semantically manipulated x such that f changes its prediction to a desired label $\tilde{y}$, i.e., $f(\tilde{x}) = \tilde{y}$. A *true counterfactual* flips the prediction by altering a causal feature, while a *false counterfactual* alters a confounding feature, thereby revealing CH effects. Counterfactuals address (1) by detecting undesired feature reliance agnostic to the number and complexity of spurious features, and (2) by providing data to eliminate predictivity of spurious cues.

3.1   Counterfactual Knowledge Distillation

To effectively harness counterfactuals for imbalanced classification, we apply CFKD as implemented in Algorithm 1. CFKD receives a base classifier f trained on an imbalanced dataset $\mathcal{D}$, where f is assumed to exhibit undesired behavior due to imbalance, along with a counterfactual explainer $\mathcal{E}$ and a teacher $\mathcal{T}$. In practice, $\mathcal{T}$ is a domain expert while our experiments rely on oracle models. We draw a subset $\mathcal{S} \subseteq \mathcal{D}$ of image-class pairs and task $\mathcal{E}$ to generate counterfactual explanations $\tilde{\mathcal{S}}$ according to the beliefs of f. Since convincing f to flip its prediction does not necessarily reflect ground-truth behavior, we query $\mathcal{T}$ to annotate each counterfactual $\tilde{x}$ with a correct label $\tilde{y}^*$. If $\mathcal{T}$ also flips its prediction (i.e., $\tilde{y}^* = \tilde{y}$), the teacher agrees with the manipulated feature to cause a label flip and attests desired classifier behavior to f. However, if the prediction of $\mathcal{T}$ remains (i.e., $\tilde{y}^* = y$), then f must base its classification behavior on a spurious correlation and a CH solution is detected. A refined classifier f' is obtained by fine-tuning f on $\mathcal{D}$ and $\tilde{\mathcal{S}}$, which forces classification behavior to align with $\mathcal{T}$.

CFKD is particularly suitable as it unifies the two-step process of (1) CH detection and (2) mitigation. Intuitively, CFKD reveals classification behavior semantically using counterfactuals and leverages ground-truth annotations to provide explicit feedback on this behavior. Thus, annotated counterfactuals serve as a proxy to correct the erroneous predictivity of arbitrary spurious correlations.

Algorithm 1: Counterfactual Knowledge Distillation

Input: Classifier f trained on $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$, explainer $\mathcal{E}$, teacher $\mathcal{T}$
Draw subset $\mathcal{S} \subseteq \mathcal{D}$ to be explained by $\mathcal{E}$, $\tilde{\mathcal{S}} = \emptyset$
for $(x, y) \in \mathcal{S}$ **do**
$\quad$| Generate counterfactual for f with $\mathcal{E}$; $\qquad \tilde{x} \leftarrow \mathcal{E}(f, x, 1-y)$
$\quad$| Query $\mathcal{T}$ for true $\tilde{y}^*$ and add to data; $\qquad \tilde{y}^* \leftarrow \mathcal{T}(\tilde{x}), \ \tilde{\mathcal{S}}.\texttt{add}\big((\tilde{x}, \tilde{y}^*)\big)$
end
Output: Corrected classifier $f' \leftarrow f.\texttt{finetune}(\mathcal{D} \cup \tilde{\mathcal{S}})$

---

4 Experiments

We study CFKD's effectiveness in mitigating CH effects emerging under imbalance on Camelyon17 [12], C-Smile, and C-Male, where the latter two are variants of CelebA [13] binarized for *Male* and *Smiling*. Medical datasets provide safety-critical, confounder-rich settings, while CelebA is notoriously known for CH effects. To compose datasets, we adjust minority class size n and imbalance ratio k to create challenging dataset instances. For Camelyon17, the hospital source site as a natural confounder is eliminated unless stated otherwise.

As a naive baseline, we train ImageNet-pretrained ResNet18 classifiers by optimizing cross-entropy (CE) with L2 regularization and early-stopping. For the base classifier to be corrected by CFKD, we merely add batch-balancing (BB) to prevent collapsed solutions. Advanced reweighting schemes are omitted intentionally to isolate correcting CH effects. We compare CFKD to Focal Loss (FL) [2] due to its robustness and widespread use in imbalanced classification. For CFKD, we train oracle classifiers on well-conditioned datasets to substitute the domain expert during experiments. To correct classifiers, we follow the implementation of CFKD and task smooth counterfactual explorers (SCE) [14] with generating counterfactual explanations for 1000 samples.
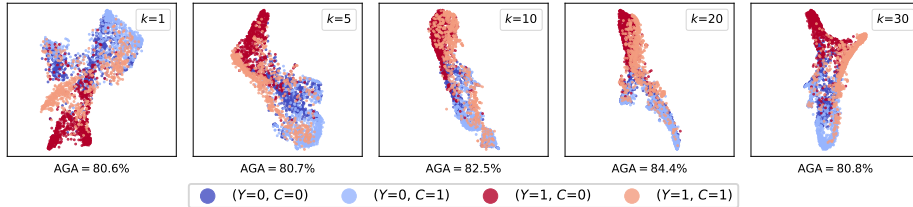

$k=1$	$k=5$	$k=10$	$k=20$	$k=30$
AGA = 80.6%	AGA = 80.7%	AGA = 82.5%	AGA = 84.4%	AGA = 80.8%

● (Y=0, C=0) ● (Y=0, C=1) ● (Y=1, C=0) ● (Y=1, C=1)

Fig. 2: AGA scores and t-SNE plots for Camelyon17 classifiers under moderate confounding C when increasing the imbalance ratio k.

4.1 Simulation on Confounder-Controlled Datasets

In a preliminary experiment, we empirically complement our analysis from Section 2 by simulating a spurious correlation emerging under imbalance. In particular, we study how surplus negatives influence classification performance as a trade-off between severe imbalance and refined features that might overcome CH effects. To this end, we utilize the hospital source site in Camelyon17 as a confounder C and compose confounder-controlled datasets. In the minority class, C co-occurs disproportionately with an observed prevalence of 90% while it only occurs in 10% of majority class samples. We set $n = 100$ to deliberately underspecify the positive class and then ablate k. Figure 2 shows the resulting representations for all four (Y, C) groups obtained by base classifiers (CE+BB) and corresponding average group accuracy (AGA). The results agree with our analysis since, for $k = 1$, limited data yields weak features and the lowest AGA, as the classifier partly relies on C (color saturation) rather than Y (color temperature). For moderate k, (Y, C) clusters become more distinct and AGA rises, indicating reduced CH effects. However, for $k = 30$, performance drops as severe imbalance adversely impacts training. Thus, surplus negatives yield diminishing returns as refined representations cannot effectively overcome CH effects.

4.2 Results

We present our main classification results in Table 1. As expected, the performance of naive CE classifiers degrades as k increases. When BB is employed, increasing k tends to improve performance slightly but yields diminishing returns, which agrees with our theoretical analysis and the previous trial. Applying CFKD to base classifiers substantially boosts performance across all datasets, and outperforms FL in most cases. The results demonstrate that annotated counterfactuals successfully identify and eliminate spurious correlations, thereby improving both performance and reliability by preventing CH solutions.

5 Conclusion

In this work, we address spurious correlations emerging under imbalance. By applying the XAI-driven approach CFKD, we successfully identify and mitigate

Table 1: F1-Score (%) for classification on C-Smile, C-Male, and Camelyon17.

n	k	C-Smile				C-Male				Camelyon17			
		CE	CE+BB	FL	CFKD	CE	CE+BB	FL	CFKD	CE	CE+BB	FL	CFKD
	10	77.2	77.6	83.3	**89.6**	82.3	80.1	85.7	**88.1**	86.9	87.8	90.6	**92.0**
100	20	76.2	78.5	77.8	**84.9**	81.9	81.1	83.8	**87.2**	85.6	86.6	90.4	**91.0**
	30	74.9	81.3	80.1	**88.4**	78.3	82.2	84.9	**87.0**	85.3	86.8	89.4	**91.5**
	10	84.5	85.0	83.5	**88.0**	86.7	88.4	86.2	**92.1**	92.0	92.7	92.9	**93.6**
200	20	84.2	85.7	86.0	**88.3**	85.3	86.9	82.4	**89.4**	91.5	88.4	91.3	90.0
	30	82.6	87.3	82.6	**88.7**	84.4	89.1	87.3	**90.2**	90.2	**91.5**	90.9	**91.5**

the arising CH effects leveraging teacher-annotated counterfactuals. Thereby, we outperform popular baselines in imbalanced classification and jointly ensure trustworthy deployment in safety-critical environments. We hope this work inspires further analysis on the interrelated effects of imbalance and CH solutions.

References

[1] E. Francazi, M. Baity-Jesi, and A. Lucchi. A theoretical analysis of the learning dynamics under class imbalance. In *International Conference on Machine Learning*, pages 10285–10322, 2023.

[2] T.Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988, 2017.

[3] H. Shah, K. Tamuly, A. Raghunathan, P. Jain, and P. Netrapalli. The pitfalls of simplicity bias in neural networks. *Advances in Neural Information Processing Systems*, 33:9573–9585, 2020.

[4] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, 10(1):1096, 2019.

[5] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K.-R. Müller. Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 109(3):247–278, 2021.

[6] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

[7] M. T. Ribeiro, S. Singh, and C. Guestrin. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD*, pages 1135–1144, 2016.

[8] A.-K. Dombrowski, J. E. Gerken, K.-R. Müller, and P. Kessel. Diffeomorphic counterfactuals with generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):3257–3274, 2024.

[9] G. Jeanneret, L. Simon, and F. Jurie. Adversarial counterfactual visual explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16425–16435, 2023.

[10] S. Bender, C. J. Anders, P. Chormai, H. A. Marxfeld, J. Herrmann, and G. Montavon. Towards fixing clever-hans predictors with counterfactual knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2607–2615, 2023.

[11] S. Bender, O. Delzer, J. Herrmann, H. A. Marxfeld, K.-R. Müller, and G. Montavon. Mitigating clever hans strategies in image classifiers through generating counterexamples. *arXiv preprint arXiv:2510.17524*, 2025.

[12] P. W. Koh, S. Sagawa, H. Marklund, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664, 2021.

[13] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738, 2015.

[14] S. Bender, J. Herrmann, K.-R. Müller, and G. Montavon. Towards desiderata-driven design of visual counterfactual explainers. *arXiv preprint arXiv:2506.14698*, 2025.