New Metrics for Disambiguating Feature Overlap and Catastrophic Forgetting in Incremental Learning Contexts

Niklas M. Melton[1], **Leonardo Enzo Brito da Silva**[2], **Donald C. Wunsch**[1]

[1] Missouri University of Science and Technology, Rolla, MO, USA
[2] Universidade Federal do Rio Grande do Norte, Natal, RN, Brazil
niklasmelton@mst.edu, leonardo.brito@imd.ufrn.br, dwunsch@mst.edu

Abstract

Catastrophic forgetting remains a central challenge in lifelong learning, where newly acquired knowledge interferes with previously learned tasks, degrading performance over time. Mitigation strategies such as rehearsal and regularization have been proposed, but both introduce limitations, either by retaining old data or by constraining model updates in ways that may impair learning. Complicating matters, recent findings show that feature-space overlap between tasks can produce similar performance drops even in models that memorize data, making it difficult to distinguish true forgetting from representational interference. Current accuracy-based metrics fail to disentangle these effects, undermining diagnostic clarity. In this work, we introduce the Overlap Index, an incremental cluster validity index adapted from the inter-cluster component of the iCONN index, which quantifies overlap between feature representations in input or latent space. We then introduce the Overshadowing and Forgetting Index, an online meta-metric that leverages the Overlap Index to attribute performance degradation to catastrophic forgetting, class overshadowing, or both. Our experimental results demonstrate that these tools enable more precise online and batch-mode evaluation of continual learning systems, paving the way for more targeted mitigation strategies.

Introduction

Lifelong learning (L2) aims to develop models capable of continually adapting as new data becomes available throughout their operational lifetimes (Parisi et al. 2019). This capability is essential for applications in evolving environments, where storing and repeatedly retraining on all previously encountered data is infeasible (Diethe et al. 2019). Unfortunately, standard gradient-based models typically optimize primarily for recent training examples, leading to progressive degradation of performance on earlier tasks – a phenomenon known as catastrophic forgetting (CF) (Grossberg 1987).

CF remains a critical barrier preventing the realization of truly adaptive and biologically plausible learning systems. To detect and mitigate CF, practitioners commonly rely on meta-metrics that monitor performance degradation over time, employing strategies such as regularization or replay (Kirkpatrick et al. 2017). However, these methods inherently compromise learning efficiency by increasing memory and computational demands or by restricting model adaptability. Even outside of L2, CF adversely affects offline model training, necessitating shuffled datasets and repeated presentations. Thus, accurately diagnosing CF and understanding its underlying causes are crucial for optimizing both lifelong and offline learning scenarios.

Recent research has highlighted a fundamental shortcoming of existing CF metrics: they cannot reliably differentiate between true CF and feature-space interference due to overlapping class representations (Melton et al. 2025). Traditional CF meta-metrics typically analyze per-class recall over time, identifying performance drops as indicators of forgetting (Díaz-Rodríguez et al. 2018). However, recall-based metrics inherently conflate genuine forgetting with a phenomenon known as *memory overshadowing*, which occurs when a model's internal representation of one class suppresses others due to overlapping features. This issue is particularly obvious in prototype-based models, such as k-Nearest Neighbors (kNN), which, due to their inherently greedy decision rules, produce singular predictions even in inherently ambiguous regions with significant feature overlap.

In this work, we address this diagnostic gap by introducing the *Overlap Index (OI)*, a novel incremental Cluster Validity Index (iCVI) (Brito Da Silva, Melton, and Wunsch 2020) designed to efficiently quantify the extent of feature overlap between class representations. We demonstrate the effectiveness of OI through comprehensive empirical evaluations on real-world datasets, comparing its interpretability and robustness to other state-of-the-art iCVIs.

Building upon the OI, we also introduce a novel, model-agnostic meta-metric termed the *Overshadowing and Forgetting Index (OFI)*. The OFI combines the insights provided by OI with traditional performance-based L2 measures to produce separate sub-indices for catastrophic forgetting and class overshadowing. Each sub-index measures the proportion of performance degradation attributed to the respective cause (forgetting or overshadowing). This contribution provides a foundation for more precise diagnostic capabilities in L2, enabling practitioners to implement targeted and efficient mitigation strategies.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

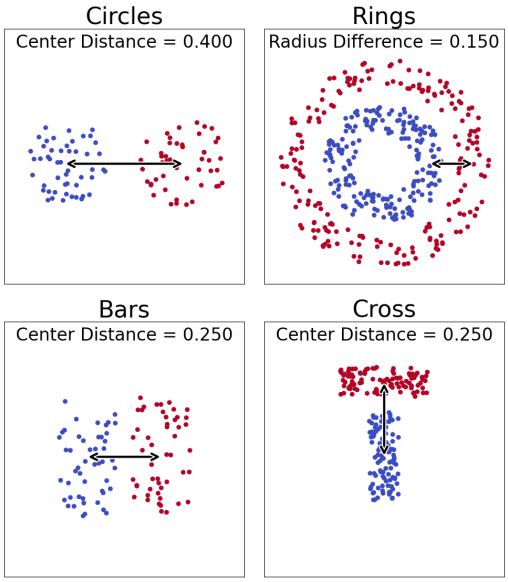

Figure 1: Synthetic Dataset exemplars used for OI experiments. All datasets contain two classes with varying degrees of overlap. In the case of the Rings dataset, overlap is controlled by varying the radius of one of the two rings, while in all others, overlap is controlled by changing the distance between class centroids. The direction of movement for all classes is indicated by the black lines on each plot, which also indicate the degree of separation for the exemplars.

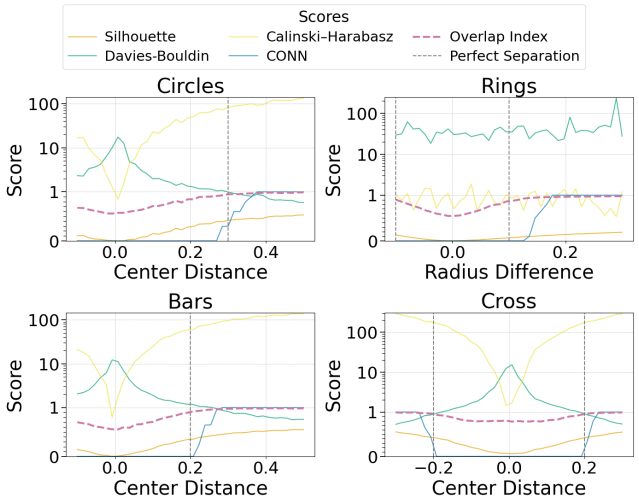

Figure 2: Cluster Validity indices for synthetic datasets as a function of separation. Silhouette, Davies-Bouldin, Calinski-Harabasz, CONN, and Overlap Indices are plotted for each synthetic dataset over a range of separation values. The vertical dashed lines on each plot indicate the point of perfect separation of the classes.

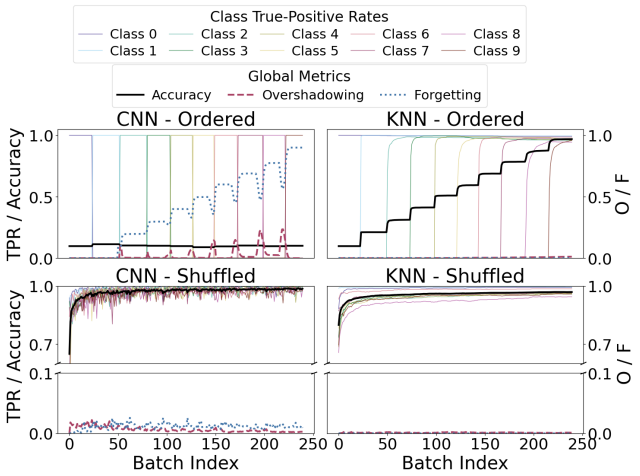

Figure 3: Overlap and Forgetting Index for both a CNN and a KNN trained on the MNIST dataset, under shuffled and class-ordered data presentations. In the ordered CNN plot, forgetting increases with each new class introduction, accompanied by brief spikes in overshadowing—potentially indicating poorly calibrated decision boundaries. The shuffled CNN plot shows only minor levels of forgetting and overshadowing, both of which diminish as training progresses. Neither KNN plot exhibits forgetting, and overshadowing appears only faintly toward the end of training, likely due to gradual recall degradation.

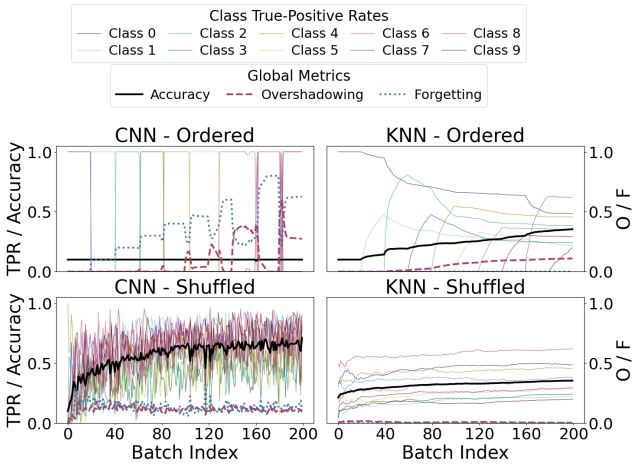

Figure 4: Overlap and Forgetting Index for both a CNN and a KNN trained on the CIFAR-10 dataset, under shuffled and class-ordered data presentations. In the ordered CNN plot, forgetting increases with each new class introduction, accompanied by brief spikes in overshadowing – possibly reflecting poorly calibrated decision boundaries. In the shuffled CNN plot, both forgetting and overshadowing are present but gradually taper off and stabilize as training progresses. Neither KNN plot shows forgetting; however, overshadowing appears in both. It is especially pronounced in the ordered KNN case, where it grows alongside recall degradation. The shuffled KNN plot, by contrast, exhibits only minor overshadowing, primarily early in training.

Acknowledgments

This research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-22-2-0209. This research was also supported by NSF grant 2420248 and by the Kummer Institute, Mary Finley Endowment, and Intelligent Systems Center of the Missouri University of Science and Technology.

The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

Brito Da Silva, L. E.; Melton, N. M.; and Wunsch, D. C. 2020. Incremental Cluster Validity Indices for Online Learning of Hard Partitions: Extensions and Comparative Study. *IEEE Access*, 8: 22025–22047.

Diethe, T.; Twomey, N.; Kull, M.; and Flach, P. 2019. Continual learning in practice. In *ESANN*.

Díaz-Rodríguez, N.; Lomonaco, V.; Filliat, D.; and Maltoni, D. 2018. Don't forget, there is more than forgetting: new metrics for Continual Learning. arXiv:1810.13166.

Grossberg, S. 1987. Competitive learning: From interactive activation to adaptive resonance. *Cognitive Science*, 11(1): 23–63.

Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwińska, A.; et al. 2017. Overcoming catastrophic forgetting in neural networks. In *Proceedings of the National Academy of Sciences*, volume 114, 3521–3526.

Melton, N. M.; Petrenko, S.; Brito da Silva, L. E.; and Wunsch II, D. C. 2025. The Need for More Nuanced Metrics of Catastrophic Forgetting. In *Proceedings of the International Conference on Soft Computing & Machine Intelligence*.

Parisi, G. I.; Kemker, R.; Part, J. L.; Kanan, C.; and Wermter, S. 2019. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113: 54–71.