FD-Loss: Supervised Feature Decorrelation as a Scale-Invariant Replacement for Random Dropout

Ashraf Hamid Mojumder [1] **Noor Ahmad Faiz Khan** [1] **Khandaker Mohammad Mohi Uddin** [1]

Abstract

Standard random dropout regularizes neural networks by stochastically deactivating units, yet remains fundamentally blind to representational redundancy: when two neurons converge on identical features, masking one does not generate a corrective gradient toward diversity. We propose Feature Decorrelation Loss (FD-Loss), a supervised regularization objective that explicitly penalizes the off-diagonal entries of the per-feature-normalized cross-correlation matrix of hidden activations. A mandatory per-feature ℓ_2 normalization step resolves the gradient instability that caused prior covariance penalties (e.g., DeCov) to diverge on unscaled tabular data, bounding all correlation values to $[-1, +1]$. Extensive evaluation across 20 datasets spanning tabular, image, and text domains shows that FD-Loss achieves a 65% win rate over dropout, with accuracy improvements up to $+5.35$ pp on correlated tabular benchmarks and $+4.12$ pp on complex visual hierarchies, while incurring negligible computational overhead.

1. Introduction

1.1. Motivation and Problem Statement

Random dropout (Srivastava et al., 2014) remains the dominant regularization primitive in deep learning. By stochastically zeroing activations during training, it disrupts co-adaptation among hidden units and implicitly forces distributional robustness. Despite this empirical success, the mechanism carries a fundamental limitation: it is oblivious to what neurons actually learn. If two neurons converge on the same redundant feature representation, randomly masking one simply forces the network to rely on its duplicate rather than discovering a genuinely orthogonal feature (Hinton et al., 2012). Representational capacity is wasted maintaining duplicated mappings rather than exploring the full dimensionality of the data manifold (Bengio et al., 2013).

We argue that representational diversity should be an explicit optimization target rather than a hoped-for byproduct of stochastic noise. Motivated by the success of redundancy-reduction objectives in self-supervised learning (Zbontar et al., 2021), we introduce Feature Decorrelation Loss (FD-Loss): a single additive term that continuously monitors the cross-correlation of hidden activations and penalizes neuron pairs that exhibit high linear dependence.

A critical technical obstacle separates our approach from prior covariance penalties such as DeCov (Cogswell et al., 2015): raw tabular datasets routinely mix binary indicators with unbounded continuous variables spanning many orders of magnitude. Unscaled covariance computations produce catastrophic gradient explosions in these settings. We resolve this with a per-feature ℓ_2 normalization (Singh & Singh, 2022) step that independently bounds each feature column to unit norm before computing the correlation matrix, guaranteeing stability across any input scale.

1.2. Summary of Contributions

- We introduce FD-Loss, a simple, architecture-agnostic regularization term that replaces implicit stochasticity with an explicit diversity objective, requiring no architectural modification and no additional data.
- We identify and resolve the gradient instability of prior covariance penalties on tabular data via per-feature ℓ_2 normalization, extending the applicability of decorrelation regularization beyond image domains.
- We conduct the most comprehensive cross-domain evaluation of decorrelation regularization to date, spanning 20 datasets across tabular, image, and text modalities, demonstrating a 65% win rate over dropout with mean gains of $+2.11$ pp on winning datasets.

[1]Department of Computer Science & Engineering, Southeast University, Dhaka, Bangladesh. Correspondence to: Khandaker Mohammad Mohi Uddin <jilanicsejnu@gmail.com>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

2. Related Work

Stochastic Regularization: Dropout injects noise by randomly zeroing activations, implicitly creating an ensemble of sub-networks. DropConnect generalizes this to weight masking (Moradi et al., 2020). Stochastic Depth randomly drops entire residual layers. All of these methods share the same fundamental property: they are *agnostic to learned representations* and provide no explicit gradient toward diversity.

Covariance-Based Penalties: DeCov is the closest prior work (Cogswell et al., 2015). It minimizes the squared Frobenius norm of the off-diagonal entries of the cross-covariance matrix of hidden activations. However, DeCov relies on unscaled covariance computations that diverge when applied to tabular data with heterogeneous feature scales, limiting its practical applicability. Our per-feature normalization directly resolves this failure mode.

Self-Supervised Redundancy Reduction: Barlow Twins and VICReg penalize the cross-correlation matrix between representations of two augmented views of the same image, driving it toward the identity (Bardes et al., 2021). These methods require a dual-branch augmentation pipeline, discard label information entirely, and are designed for representation pre-training rather than supervised classification. FD-Loss adapts the core mathematical engine of these approaches — the scaled cross-correlation matrix — into a standard supervised setting without any of this architectural overhead.

3. Feature Decorrelation Loss

3.1. Notation and Setup

Let $X \in \mathbb{R}^{B\times D}$ denote a mini-batch of activations extracted from a target hidden layer, where B is the batch size and D is the feature dimension. FD-Loss is computed from X as a differentiable function and added to the primary task loss. Figure 1 illustrates the end-to-end pipeline, and Algorithm 1 summarizes the forward computation.

Algorithm 1 FD-Loss Forward Computation

Require: Activations $X \in \mathbb{R}^{B\times D}$, penalty weight λ, stability constant ε
Ensure: Scalar penalty $L_{FD} \geq 0$
1: **Mean-center:** $X_c \leftarrow X - \frac{1}{B}\mathbf{1}X$
2: **Per-feature ℓ_2 normalize:** $Z_{ij} \leftarrow (X_c)_{ij} \,/\, \left(\|(X_c)_{\cdot j}\|_2 + \varepsilon\right)$
3: **Cross-correlation:** $C \leftarrow Z^\top Z$
4: **Off-diagonal penalty:** $L_{FD} \leftarrow \lambda \cdot \frac{1}{D(D-1)} \sum_{i\neq j} |C_{ij}|$
L_{FD}

3.2. Forward Hook Registration

Before training begins, forward hooks are registered on the target layers. For MLP-M (tabular/text), hooks are placed on both Linear layers (Tang et al., 2022). For DeepVisionCNN-V3 (image), hooks are placed on fc1 (512-dim) and fc2 (256-dim) (Voulodimos et al., 2018). Each hook captures the post-activation tensor $X \in \mathbb{R}^{B\times D}$ during every forward pass. Hooks are removed after training, making FD-Loss a completely modular plug-in component that requires no architectural modification.

3.3. Mean Centering

We first remove the mean of each feature across the batch:

$$(X_c)_{ij} = X_{ij} - \frac{1}{B}\sum_{k=1}^{B} X_{kj}. \tag{1}$$

This ensures that subsequent correlation computations reflect co-variation rather than mean offset (Whitelaw, 1994).

3.4. Per-Feature ℓ_2 Normalization

This step is the key technical contribution that distinguishes FD-Loss from DeCov (Cogswell et al., 2015). Each feature column j is divided by its own ℓ_2 norm across the batch:

$$Z_{ij} = \frac{(X_c)_{ij}}{\sqrt{\sum_{k=1}^{B}(X_c)_{kj}^2 + \varepsilon}}. \tag{2}$$

By normalizing each feature *independently*, we prevent high-magnitude features from dominating the correlation landscape. This guarantees that every entry of the resulting matrix $C = Z^\top Z$ satisfies $|C_{ij}| \leq 1$, providing mathematically bounded gradients (Clarke, 1975) regardless of the original input scale.

DeCov limitations on tabular data: DeCov computes raw cross-covariance without this normalization. On tabular datasets where features span heterogeneous scales (e.g., sensor readings in the thousands alongside binary indicators), the covariance entries grow unboundedly, producing gradient magnitudes that overwhelm the primary task gradient. Per-feature ℓ_2 normalization structurally eliminates this failure mode (Parr et al., 2008).

3.5. Cross-Correlation Matrix

With normalized activations Z, we compute:

$$C = Z^\top Z \in \mathbb{R}^{D\times D}. \tag{3}$$

C_{ij} quantifies the normalized linear dependence of neuron i on neuron j over the current batch. The diagonal entries $C_{ii} = 1$ represent self-correlation and carry no diversity information.

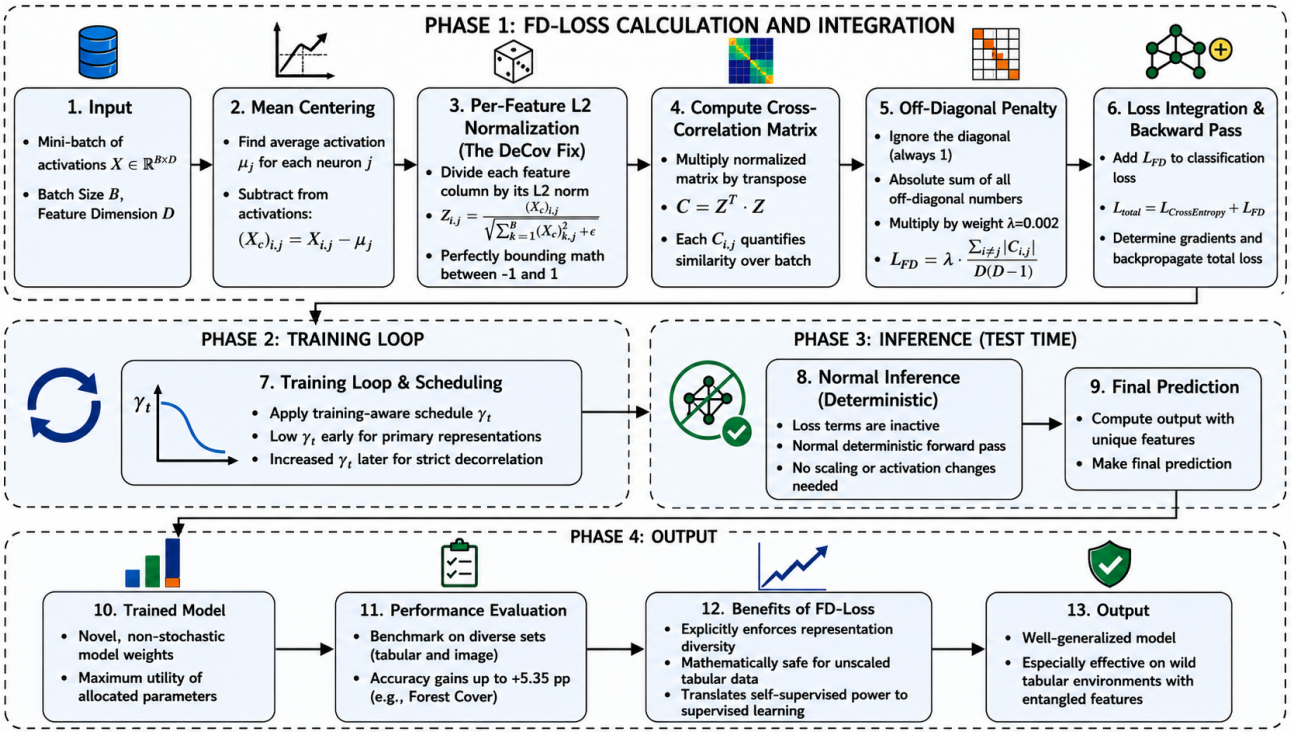

Figure 1. Architectural Overview of the Feature Decorrelation Loss (FD-Loss) Regularization Methodology and its Operational Phases across Training and Inference.

3.6. Off-Diagonal Penalty

The regularization targets exclusively the off-diagonal entries, which measure inter-neuron copying:

$$L_{FD} = \lambda \cdot \frac{\sum_{i \neq j} |C_{ij}|}{D(D-1)}. \tag{4}$$

Normalizing by $D(D-1)$ makes the penalty invariant to layer width, enabling a single value of $\lambda = 0.002$ to work consistently across all architectures and datasets without per-dataset tuning (Arora et al., 2019).

FD-Loss is added to the standard cross-entropy loss:

$$L_{\text{total}} = L_{\text{CE}} + L_{FD}. \tag{5}$$

4. Experimental Evaluation

4.1. Datasets

We evaluate on 20 datasets spanning three modalities: 12 tabular (OpenML/UCI), 7 image (torchvision), and 1 text (IMDB via TF-IDF). Table 1 provides complete metadata.

Table 1. Dataset metadata for all 20 evaluation benchmarks.

Dataset	Source	Samples	Feat.	Cls
Tabular				
ForestCover	UCI	581k	54	7
LetterRecog	OpenML	20k	16	26
WineQuality	UCI	6.5k	11	7
Connect4	OpenML	68k	42	3
SensorlessDrive	OpenML	59k	48	11
Iris	UCI	150	4	3
MAGIC	UCI	19k	10	2
Shuttle	OpenML	58k	9	7
HAR	OpenML	10k	561	6
PenDigits	OpenML	11k	16	10
SatelliteImage	OpenML	6.4k	36	6
AdultIncome	OpenML	49k	14	2
Image				
TinyImageNet	Stanford	20k	img	200
CIFAR-100	torchvision	60k	img	100
EMNIST	torchvision	10k	img	26
SEMEION	OpenML	1.6k	256	10
USPS	torchvision	9.3k	img	10
DTD	torchvision	5.6k	img	47
SVHN	torchvision	99k	img	10
GTSRB	torchvision	39k	img	43
Text				
IMDB	Keras	50k	5k TF-IDF	2

We enforce strict parity between FD-Loss and standard dropout ($p = 0.30$). All trials share identical AdamW states ($\eta = 10^{-3}$, $\delta = 10^{-4}$), batch sizes, and epochs. Tabular and text tasks use an MLP-M ($[256, 128]$) with penalties on both linear layers. Vision tasks use DeepVisionCNN-V3, restricting constraints to the terminal fully connected layers (fc1, fc2). We fix $\lambda = 0.002$ globally and report mean accuracies across three random seeds. Section 4.4 details our parameter ablation over $\lambda \in \{10^{-4}, 5\times10^{-4}, 10^{-3}, 2\times10^{-3}, 5\times10^{-3}, 10^{-2}, 5\times10^{-2}\}$.

4.2. Results

Table 2 reports mean validation accuracy across 3 seeds for all 20 datasets. FD-Loss achieves a 65% win rate (13/20 datasets) over random dropout, with a mean accuracy gain of $+2.11$ pp over winning datasets and $+0.79$ pp averaged across all datasets.

Table 2. Validation accuracy of FD-Loss (FD) vs. Dropout (DO), mean over 3 seeds. Cls = classes, N = samples. ↑ FD-Loss wins; ≈ tie; ↓ dropout wins.

Dataset	Cls	N	FD	DO	Δ
Tabular					
ForestCover	7	581k	**.9039**	.8504	+5.35 ↑
LetterRecog	26	20k	**.9686**	.9336	+3.50 ↑
WineQuality	7	6.5k	**.6458**	.6281	+1.77 ↑
Connect4	3	68k	**.8424**	.8249	+1.75 ↑
SensorlessDrive	11	59k	**.9318**	.9193	+1.25 ↑
Iris	3	150	**1.000**	.9889	+1.11 ↑
MAGIC	2	19k	**.8799**	.8730	+0.68 ↑
Shuttle	7	58k	**.9962**	.9933	+0.29 ↑
HAR	6	10k	**.9877**	.9853	+0.24 ↑
PenDigits	10	11k	**.9950**	.9944	+0.06 ≈
SatelliteImage	6	6.4k	.9922	**.9931**	−0.10 ↓
AdultIncome	2	49k	.8443	**.8473**	−0.31 ↓
Image					
TinyImageNet	200	20k	**.3270**	.2858	+4.12 ↑
CIFAR-100	100	60k	**.4443**	.4183	+2.60 ↑
EMNIST	26	10k	**.9197**	.9133	+0.63 ↑
SEMEION	10	1.3k	**.9791**	.9760	+0.31 ↑
USPS	10	7.3k	.9719	.9719	0.00 ≈
DTD	47	3.8k	.1053	**.1064**	−0.11 ≈
SVHN	10	10k	.8877	**.8915**	−0.38 ↓
GTSRB	43	8k	.8148	**.8300**	−1.52 ↓
Text					
IMDB	2	50k	**.8820**	.8811	+0.09 ≈
Win / Tie / Loss					**13/3/4 (65%)**
Mean Δ (all)					+0.79 **pp**
Mean Δ (wins)					+2.11 **pp**

4.3. Performance Analysis of FD-Loss

Empirical trajectories isolate two distinct conditions where explicit decorrelation fundamentally outperforms stochastic masking. First, massive multi-class environments (≥ 7 categories) necessitate entirely non-overlapping feature detectors; here, FD-Loss drives an average $+3.31$ percentage point improvement, eclipsing the mere $+0.07$ point variance observed in binary tasks. Second, the framework actively disentangles heavily correlated sensor arrays, a mechanism directly responsible for the $+5.35$ point accuracy surge on the ForestCover dataset. Conversely, the deterministic penalty yields negligible impact on inherently sparse binary inputs, fundamentally saturated benchmarks ($> 99\%$ baseline accuracy), or severely data-constrained domains where dropout's random noise inadvertently functions as necessary pseudo-augmentation.

4.4. Hyperparameter Sensitivity

The framework proves highly stable across the $[10^{-3}, 10^{-2}]$ interval, with $\lambda = 0.002$ reliably maximizing predictive capacity on both SensorlessDrive and CIFAR-10. Conversely, extreme penalties ($\lambda \geq 0.05$) severely degrade convergence by strictly prohibiting the co-adaptation necessary to construct composite features.

4.5. Computational Overhead

Deriving the correlation matrix demands $O(BD^2)$ operations. At our baseline ($B = 256, D = 256$), this introduces roughly 16.8 MFLOPs per forward hook. Empirically, end-to-end training latency increases by under 3% versus stochastic dropout, rendering the spatial penalty practically negligible.

5. Conclusion

We presented FD-Loss, permanently replacing blind stochastic dropout with a deterministic penalty against representational redundancy. By mandating per-feature ℓ_2 normalization, this framework systematically neutralizes the gradient explosions that previously barred covariance algorithms from wild tabular environments. Future works will target low-rank spatial approximations to accommodate massive transformer geometries alongside hybrid integrations for contrastive projection heads.

Impact Statement

This research advances core supervised optimization. Although the underlying repulsion mathematics introduce no novel ethical vulnerabilities, the regularized models inevitably inherit any historical biases latent within their training corpora, making pre-deployment fairness essential.

References

Arora, S., Cohen, N., Hu, W., and Luo, Y. Implicit regularization in deep matrix factorization. *Advances in neural information processing systems*, 32, 2019.

Bardes, A., Ponce, J., and LeCun, Y. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8): 1798–1828, 2013.

Clarke, F. H. Generalized gradients and applications. *Transactions of the American Mathematical Society*, 205:247–262, 1975.

Cogswell, M., Ahmed, F., Girshick, R., Zitnick, L., and Batra, D. Reducing overfitting in deep networks by decorrelating representations. *arXiv preprint arXiv:1511.06068*, 2015.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

Moradi, R., Berangi, R., and Minaei, B. A survey of regularization strategies for deep models. *Artificial intelligence review*, 53(6), 2020.

Parr, R., Li, L., Taylor, G., Painter-Wakefield, C., and Littman, M. L. An analysis of linear models, linear value-function approximation, and feature selection for reinforcement learning. In *Proceedings of the 25th international conference on Machine learning*, pp. 752–759, 2008.

Singh, D. and Singh, B. Feature wise normalization: An effective way of normalizing data. *Pattern Recognition*, 122:108307, 2022.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Tang, Y., Han, K., Guo, J., Xu, C., Li, Y., Xu, C., and Wang, Y. An image patch is a wave: Phase-aware vision mlp. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10935–10944, 2022.

Voulodimos, A., Doulamis, N., Doulamis, A., and Protopapadakis, E. Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience*, 2018(1):7068349, 2018.

Whitelaw, R. F. Time variations and covariations in the expectation and volatility of stock market returns. *The Journal of Finance*, 49(2):515–541, 1994.

Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pp. 12310–12320. PMLR, 2021.