

Supplementary Material for “Lightweight and Explainable Anomaly Detection in CAN Bus Traffic via Non-negative Matrix Factorization”

Anandkumar Balasubramaniam and Seung Yeob Nam

This document provides supplementary material (figures and tables) referenced in the main manuscript.

1 Detailed Cross-Attack Generalization Results

This section contains additional cross-attack heatmaps and delta analyses for all models, metrics, and feature representations discussed in Section 5.7 of the main manuscript. For each tabular (RF, IF) and sequence (CNN, LSTM, Attention) model, we show heatmaps for AUC-PR, AUC-ROC, and F1-score under both Raw and NMF-W input spaces. We also provide difference (delta) heatmaps showing performance shifts between NMF-W and Raw features.

Fig. S1–S5 present the full cross-attack heatmaps for RF, IF, CNN, LSTM, and Attention models, while Fig. S6 and S7 show the corresponding difference heatmaps. Each figure displays three metrics (AUC-PR, AUC-ROC, F1-score) for both Raw features (top row) and NMF-W representations (bottom row). Delta heatmaps show the difference (NMF-W – Raw) for each metric, and all plots include 95% bootstrapped confidence intervals.

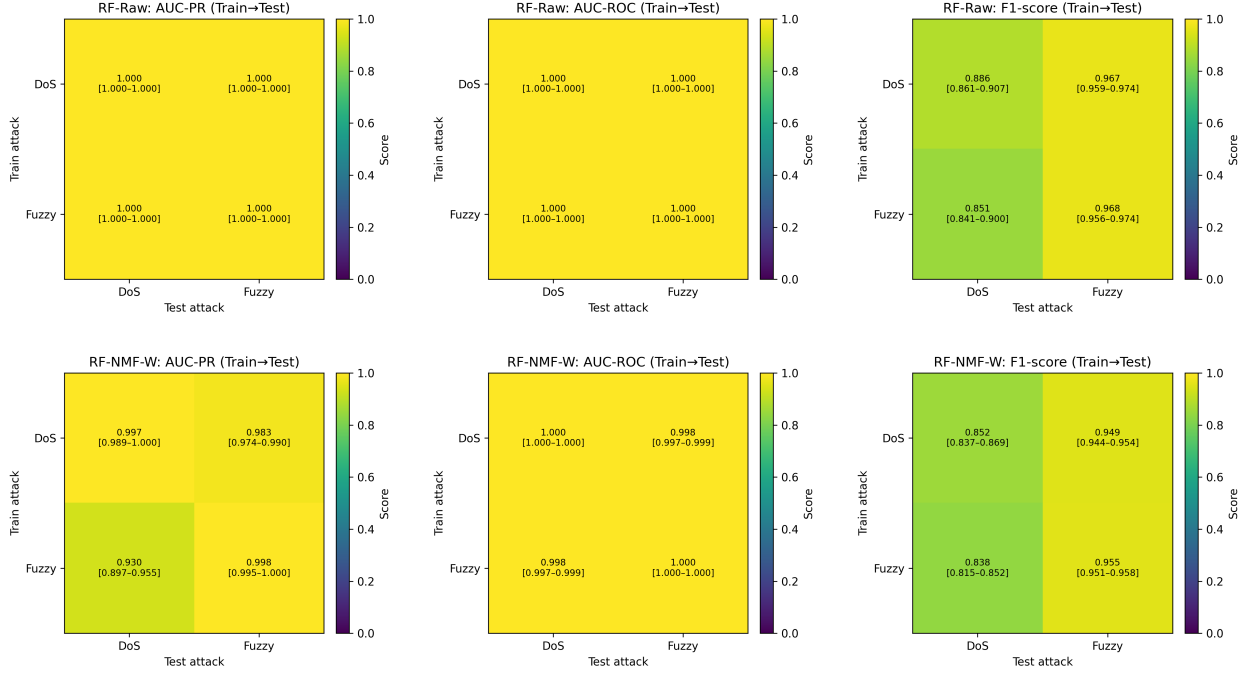


Figure S1: Cross-attack generalization of RF on DoS and Fuzzy attacks. Each heatmap shows the mean score with a 95% confidence interval, for both Raw features (top row) and NMF-W representations (bottom row).

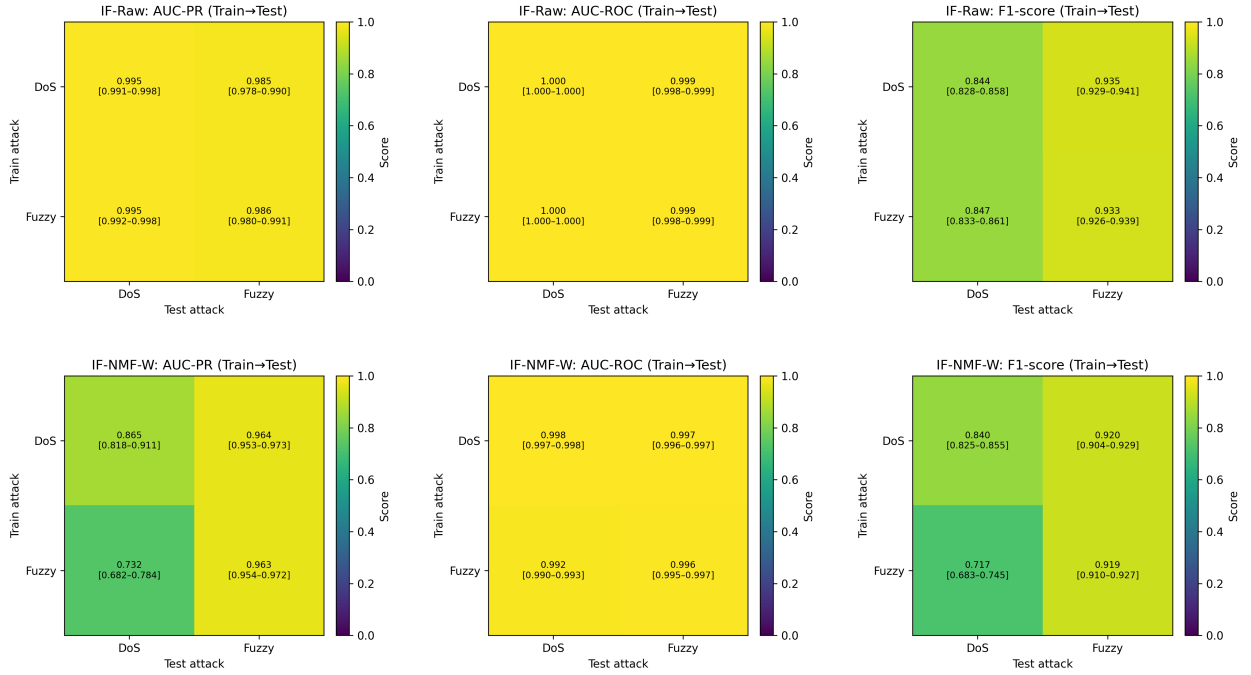


Figure S2: Cross-attack generalization of IF on DoS and Fuzzy attacks using Raw features (top row) and NMF-W representations (bottom row).

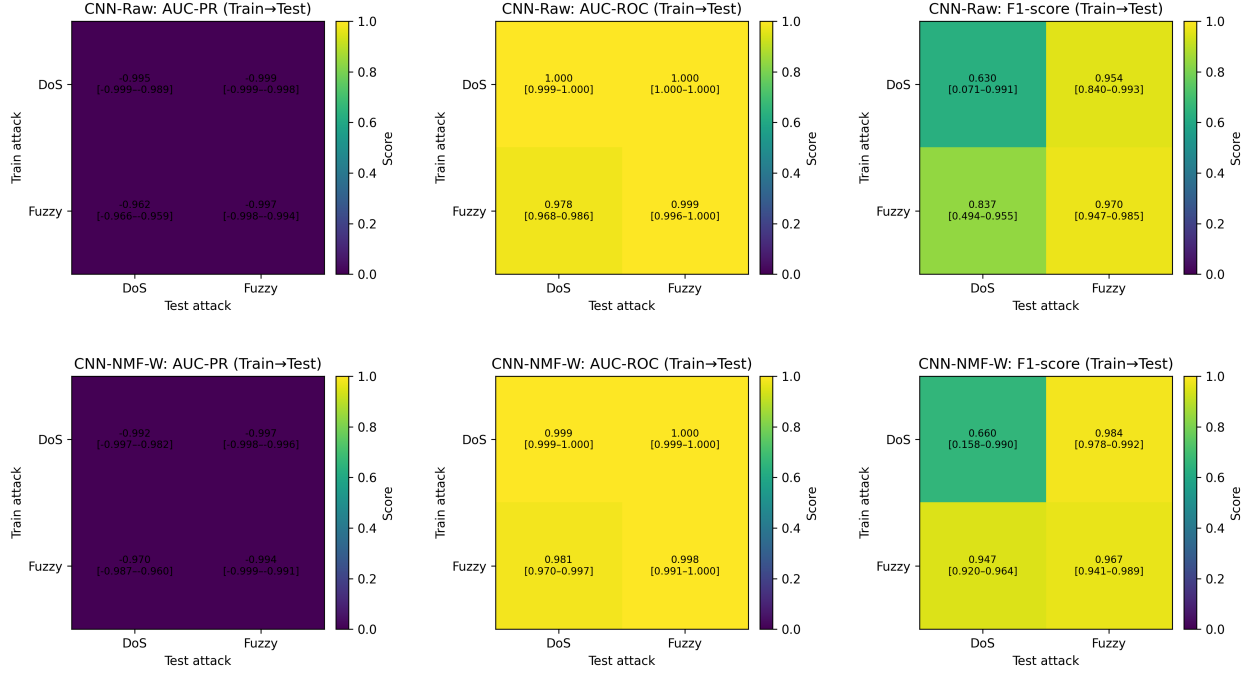


Figure S3: Cross-attack generalization of the CNN-based sequence model on DoS and Fuzzy attacks, using Raw (top row) and NMF-W (bottom row) inputs.

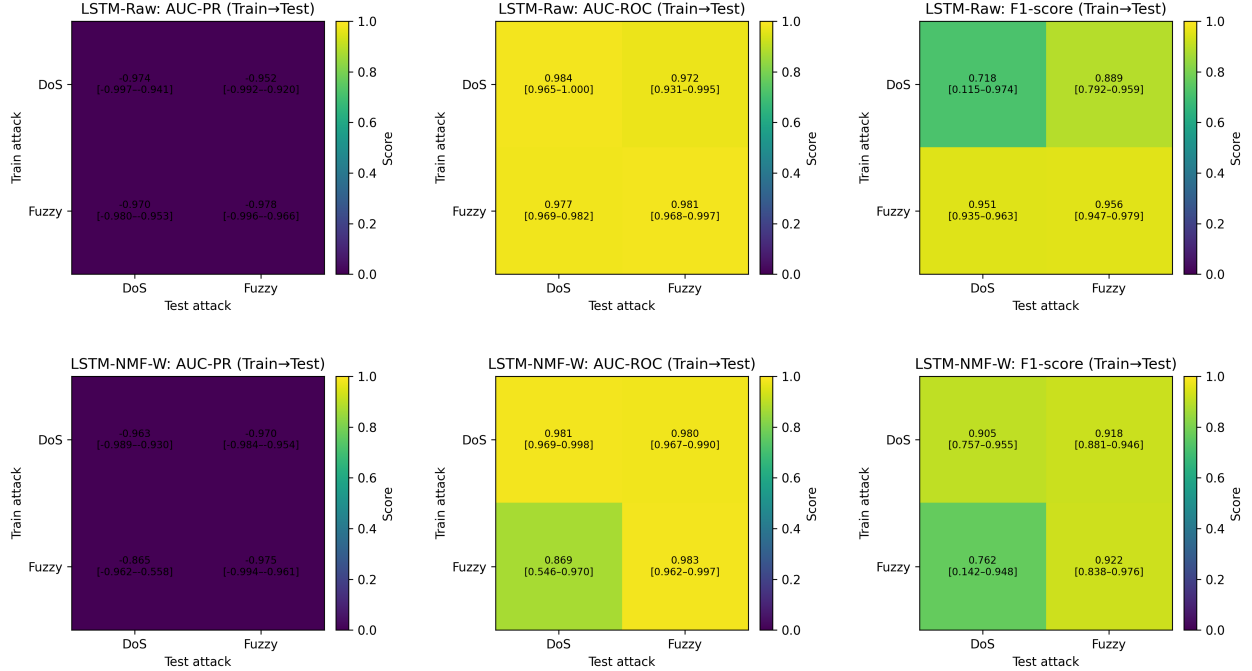


Figure S4: Cross-attack generalization of the LSTM-based sequence model on DoS and Fuzzy attacks for Raw (top row) and NMF-W (bottom row) representations.

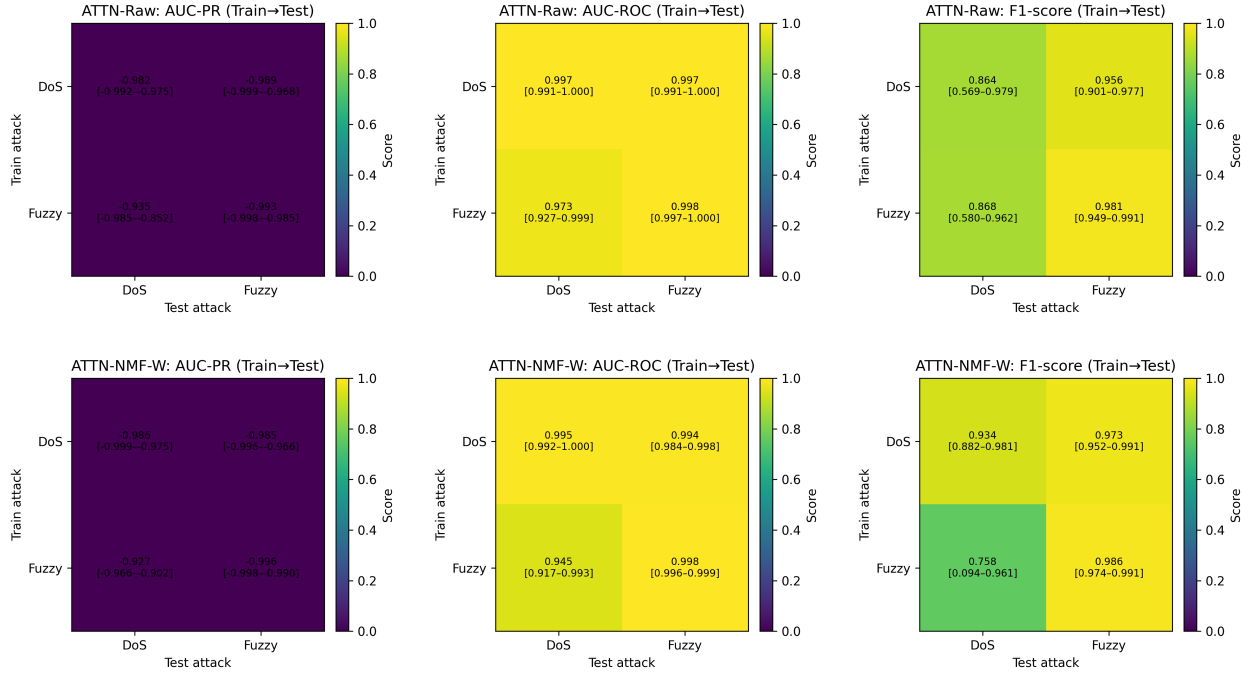


Figure S5: Cross-attack generalization of the Attention-based sequence model on DoS and Fuzzy attacks, comparing Raw (top row) and NMF-W (bottom row) inputs.

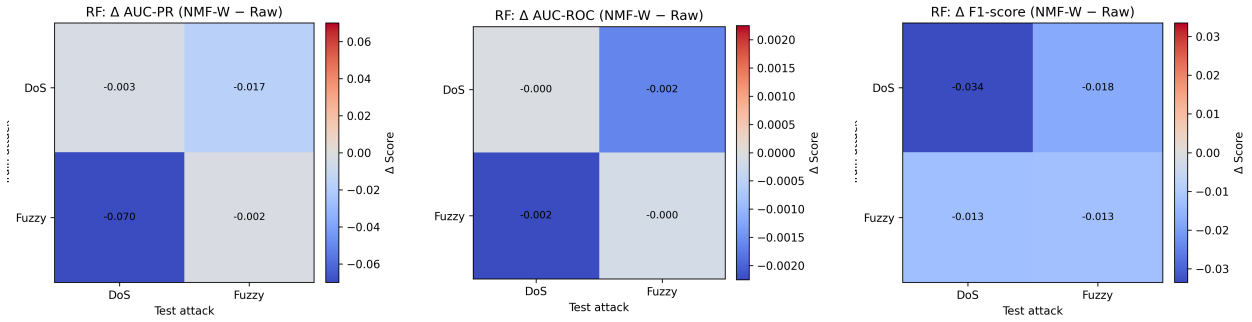


Figure S6: Difference heatmaps for RF, showing the change in performance when switching from Raw features to NMF-W representations (NMF-W - Raw) for AUC-PR, AUC-ROC, and F1-score.

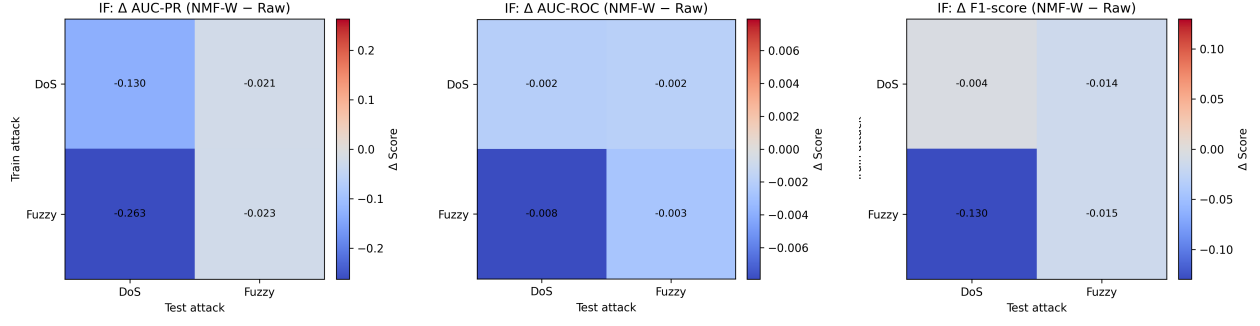


Figure S7: Difference heatmaps for IF, showing the change in performance when switching from Raw features to NMF-W representations (NMF-W – Raw) for AUC-PR, AUC-ROC, and F1-score.

2 Supplementary Interpretability Plots

In addition to the main interpretability analysis, this section provides violin plots showing the distribution of per-component activations (\mathbf{W}) across different attack phases.

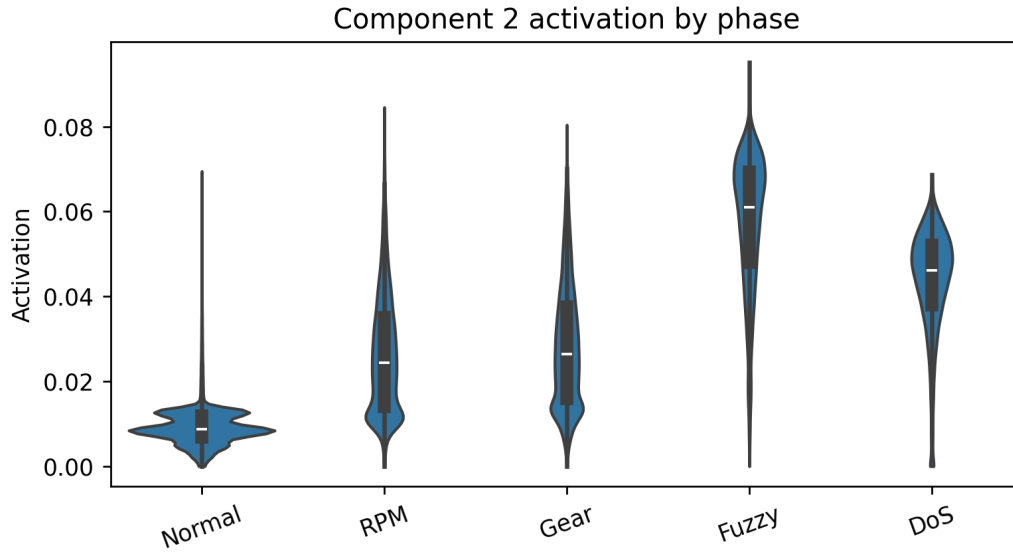


Figure S8: Violin plot of NMF component (C2) activations by attack type. A distinct separation between Normal and attack traffic is visible.

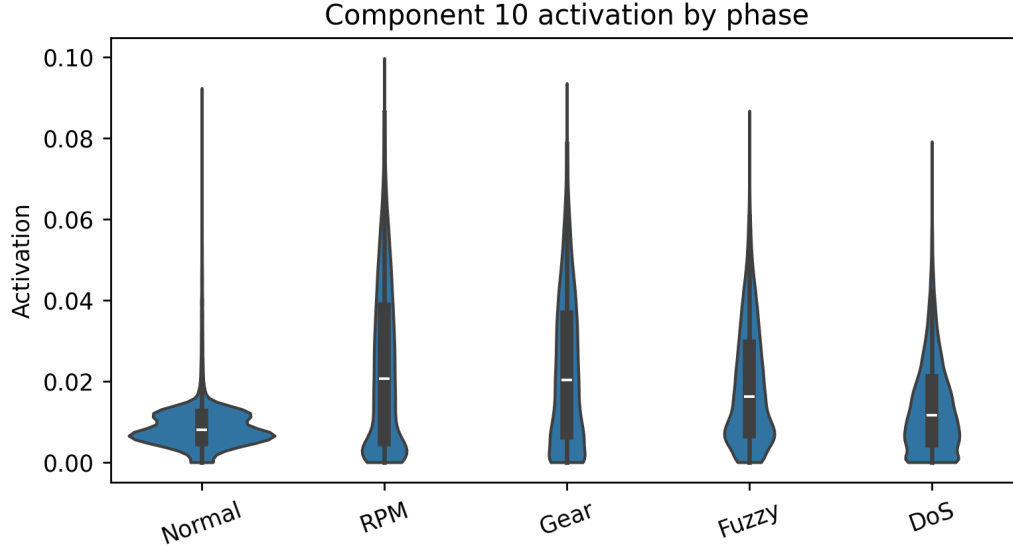


Figure S9: Violin plot of NMF component (C10) activations by attack type. A distinct separation between Normal and attack traffic is visible.

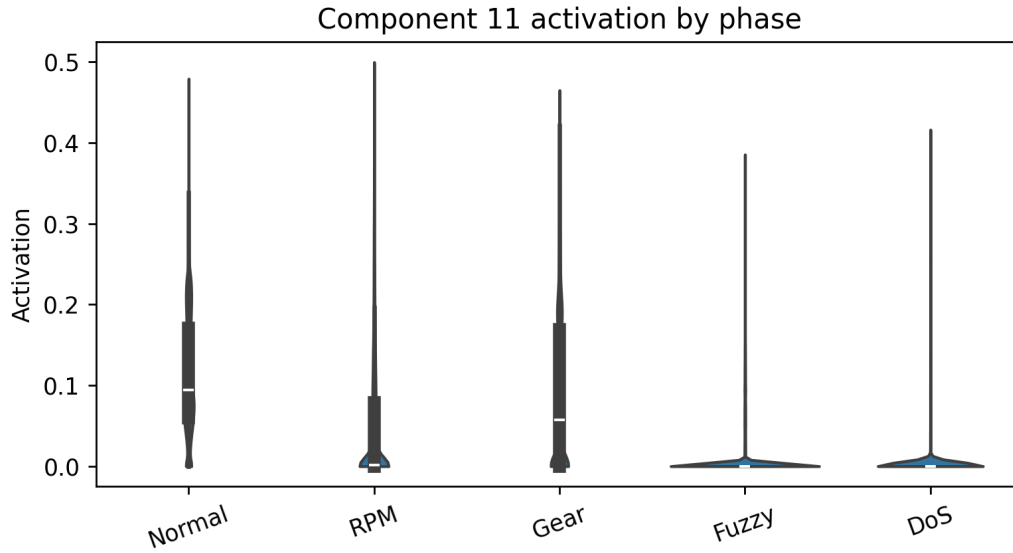


Figure S10: Violin plot of NMF component (C11) activations by attack type. A distinct separation between Normal and attack traffic is visible.

These plots highlight the variability of NMF component activations under Normal, DoS, Fuzzy, Gear spoofing, and RPM spoofing traffic, complementing the averaged prototypes presented in the main text.

3 Model Feature Importances

Tables S1 and S2 list the complete top-ranked features and NMF components selected by RF, XGB, and LR.

Table S1: Full top-ranked raw features across RF, XGB, and LR.

Rank	RF (Gini)	XGB (Gain)	LR (—coef—)
1	dlc_hist_2	idfreq_1349	new_id_count
2	idfreq_1349	msg_count	idfreq_399
3	idfreq_1088	idfreq_1088	idfreq_880
4	idfreq_2	new_id_ratio	idfreq_1088
5	idfreq_809	new_id_count	idfreq_1072
6	idfreq_848	msg_rate	idfreq_1087
7	topID_dominance	dlc_hist_2	idfreq_790
8	idfreq_304	idfreq_339	idfreq_704
9	idfreq_305	idfreq_848	idfreq_608
10	idfreq_339	topID_dominance	idfreq_1264
11	idfreq_608	per_id_repeat_ratio_mean	idfreq_848
12	idfreq_320	idfreq_2	idfreq_2
13	id_entropy	id_entropy	idfreq_339
14	idfreq_880	idfreq_304	idfreq_809
15	idfreq_704	idfreq_880	idfreq_672
16	interarrival_mean	idfreq_399	idfreq_304
17	msg_rate	idfreq_790	idfreq_688
18	idfreq_399	idfreq_672	idfreq_305
19	msg_count	idfreq_1087	idfreq_1201
20	dlc_mean	dlc_hist_8	idfreq_1349

Table S2: Full top-ranked NMF components across RF, XGB, and LR.

Rank	RF (Gini)	XGB (Gain)	LR (—coef—)
1	C2	C11	C8
2	C3	C1	C6
3	C10	C8	C2
4	C7	C14	C10
5	C4	C12	C1
6	C8	C5	C12
7	C6	C2	C14
8	C9	C15	C7
9	C13	C0	C4
10	C1	C13	C3
11	C15	C6	C15
12	C11	C4	C13
13	C0	C3	C11
14	C14	C9	C5
15	C12	C7	C9
16	C5	C10	C0

These detailed rankings highlight the strong emphasis on specific CAN IDs in the raw feature space, contrasted with the distributed latent factors leveraged in the NMF-W feature space.