A Generalizability of MI estimation methods

We assessed the generalization of our MI estimation methods with top-performing calibration (CAL LOG-LOSS), against baseline methods, using the normalized mean absolute error (NMAE) metric [Gupta(2025), chap. 4]. The heatmaps illustrate generalization across the number of classes (C) and input dimension (d) in balanced datasets (c.f. A.1), and across class imbalance (r) and noise level (ϵ) in binary and imbalanced multi-class datasets (c.f. A.2).

A.1 Number of Classes (C) and Input Dimensions (d)

To understand the generalization capabilities of different MI estimation methods, with respect to the number of classes (C) and input dimensions (d), and evaluate their performance on *vulnerable* systems at 0 % and 50 % noise levels in Figures 1 and 2, respectively, and on *non-vulnerable* synthetic systems at 100 % noise level in Figure 3. In the heatmaps, the Y-axis represents the number of classes (2 to 10), and the X-axis represents the input dimensions (2 to 20).

TabPFN Overall, MI estimation methods using TabPFN generalize well with the number of classes (C) and input dimensions (d). With few exceptions in vulnerable systems using MVN perturbation, TabPFN LOG-LOSS and CAL LOG-LOSS achieve an NMAE around 0.01 across most cases, demonstrating robustness in high-dimensional, multi-class settings. This also reaffirms that calibration does not improve TabPFN LOG-LOSS precision [Gupta(2025), chap. 4]. In contrast, MID-POINT performs poorly on high-dimensional ($d \geq 14$) vulnerable datasets, especially with MVN perturbations (NMAE $\approxeq 0.20$).

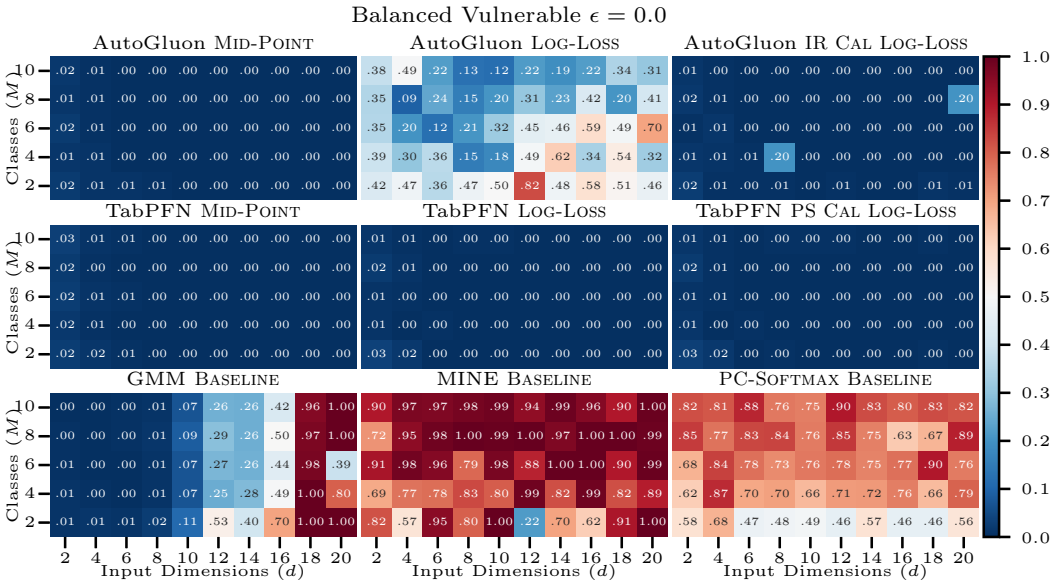
AutoGluon Overall, MI estimation methods using AutoGluon perform well across varying class counts (C) and input dimensions (d), with some exceptions noted in [Gupta(2025), chap. 4]. MID-POINT and CAL LOG-LOSS tend to overestimate MI in non-vulnerable binary-class settings due to overfitting, a pattern that worsens with increasing C and d in vulnerable datasets. Calibration often degrades AutoGluon LOG-LOSS performance in non-vulnerable systems, while improving precision in vulnerable settings [Gupta(2025), chap. 4], highlighting its role in falsely detecting non-existent information leakages (ILs) [Gupta(2025), chap. 5].

Baselines All baseline methods show limited generalization in both vulnerable and non-vulnerable systems, with performance declining as the number of classes (C) and input dimensions (d) increase. As expected, Gaussian mixture model (GMM) struggles in high-dimensional settings, degrading beyond 10 features [Maia Polo and Vicente(2022)]. It generally outperforms PC-SOFTMAX and mutual information neural estimation (MINE), except in non-vulnerable systems simulated using MVN perturbation technique [Gupta(2025)].

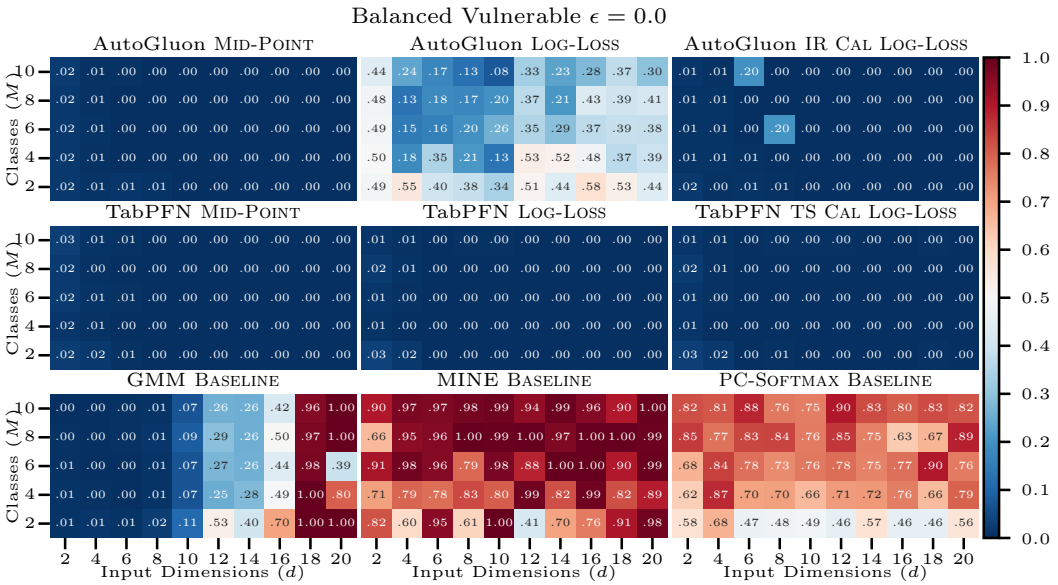
A.2 Class Imbalance (r) and Noise Level (ϵ)

We assess the generalization of MI estimation methods with respect to class imbalance (r) and noise level (ϵ) using NMAE on binary and imbalanced multi-class datasets generated via MVN perturbation and proximity techniques. Heatmaps in Figures 4 and 5 (binary) and Figures 6 and 7 (multi-class) illustrate the results.

The Y-axis represents noise levels ($\epsilon \in [0.0, 1.0]$), and the X-axis represents class imbalance: $r \in [0.05, 0.5]$ for binary and $r \in [0.02, 0.2]$ for multi-class datasets.

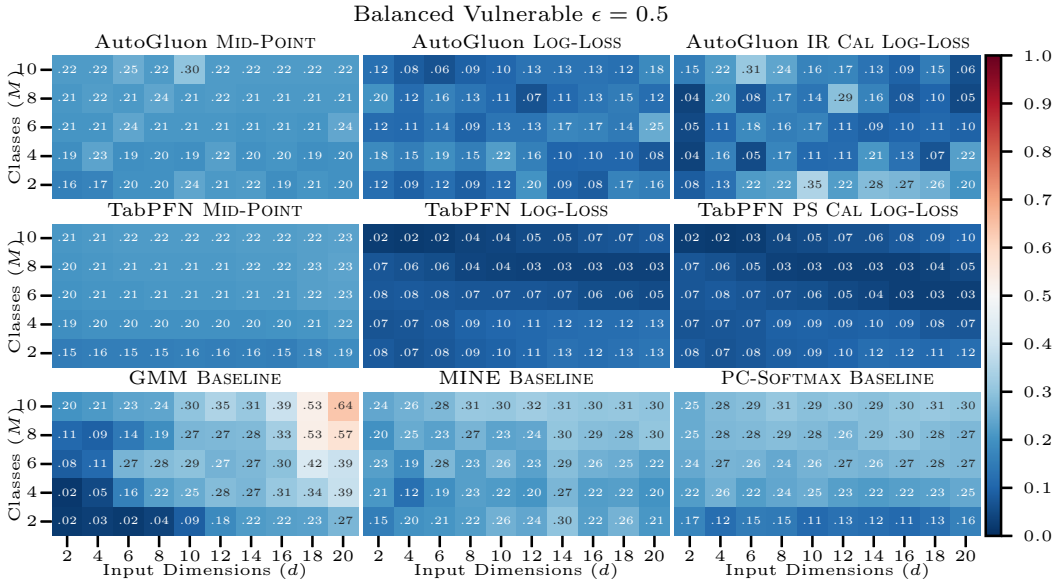


(a) Balanced multivariate normal (MVN) perturbation dataset with 0 % noise level

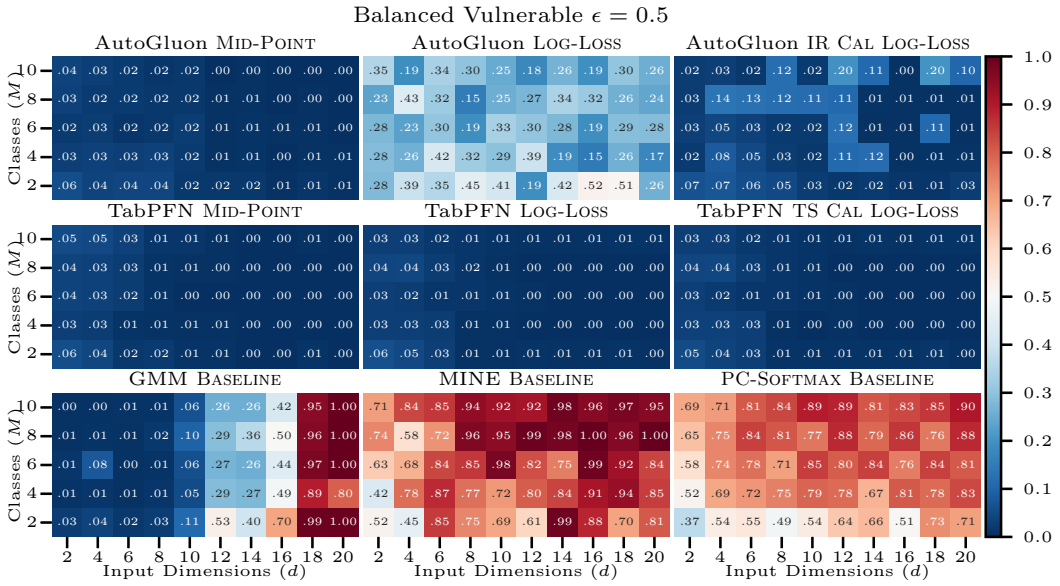


(b) Balanced MVN proximity dataset with 0 % noise level

Figure 1: Generalizability of mutual information (MI) estimation methods on noise-free vulnerable systems

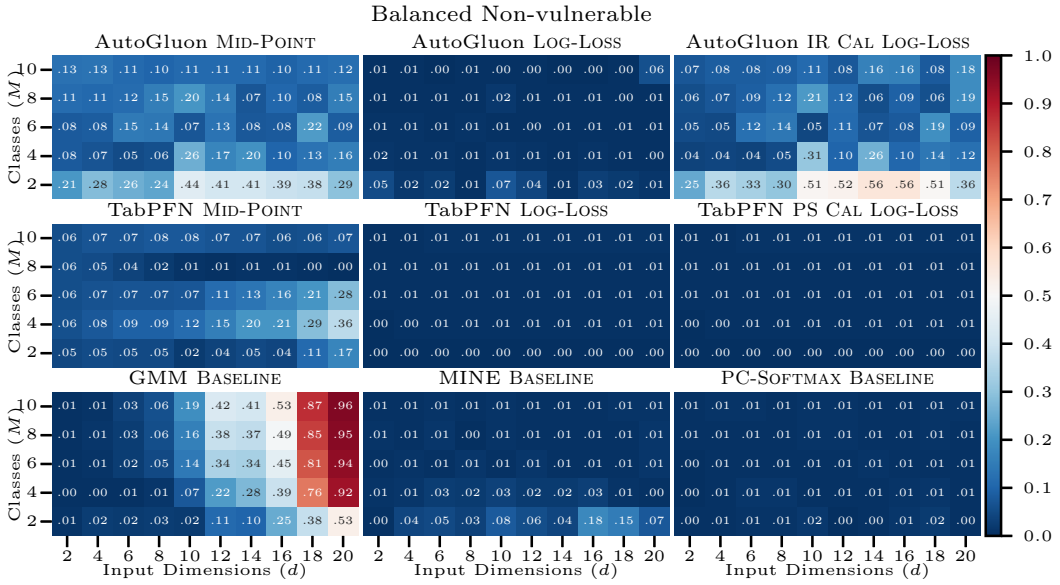


(a) Balanced MVN perturbation dataset with 50 % noise level

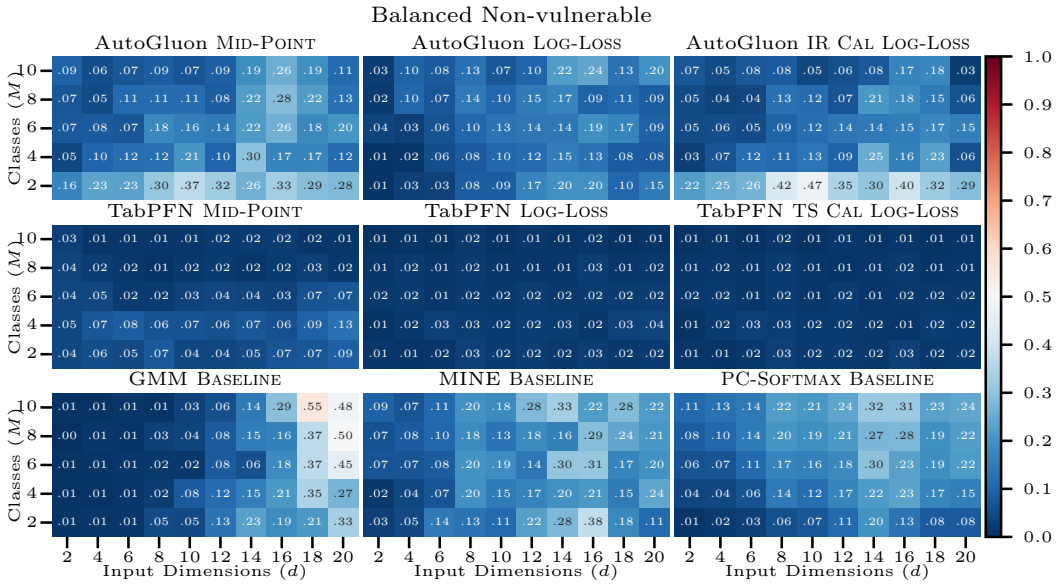


(b) Balanced MVN proximity dataset with 50 % noise level

Figure 2: Generalizability of MI estimation methods on noisy vulnerable systems



(a) Balanced MVN perturbation dataset with 100 % noise level



(b) Balanced MVN proximity dataset with 100 % noise level

Figure 3: Generalizability of MI estimation methods on non-vulnerable systems

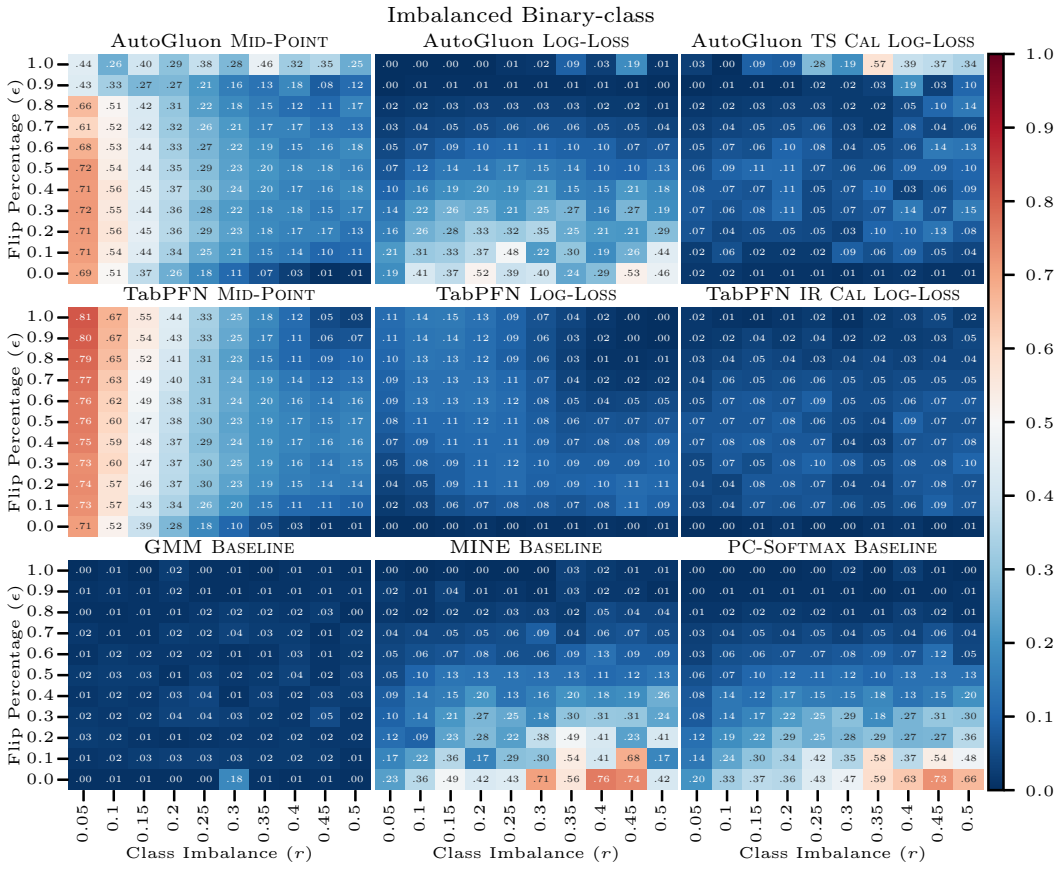


Figure 4: Generalizability on MVN perturbation imbalanced binary-class datasets

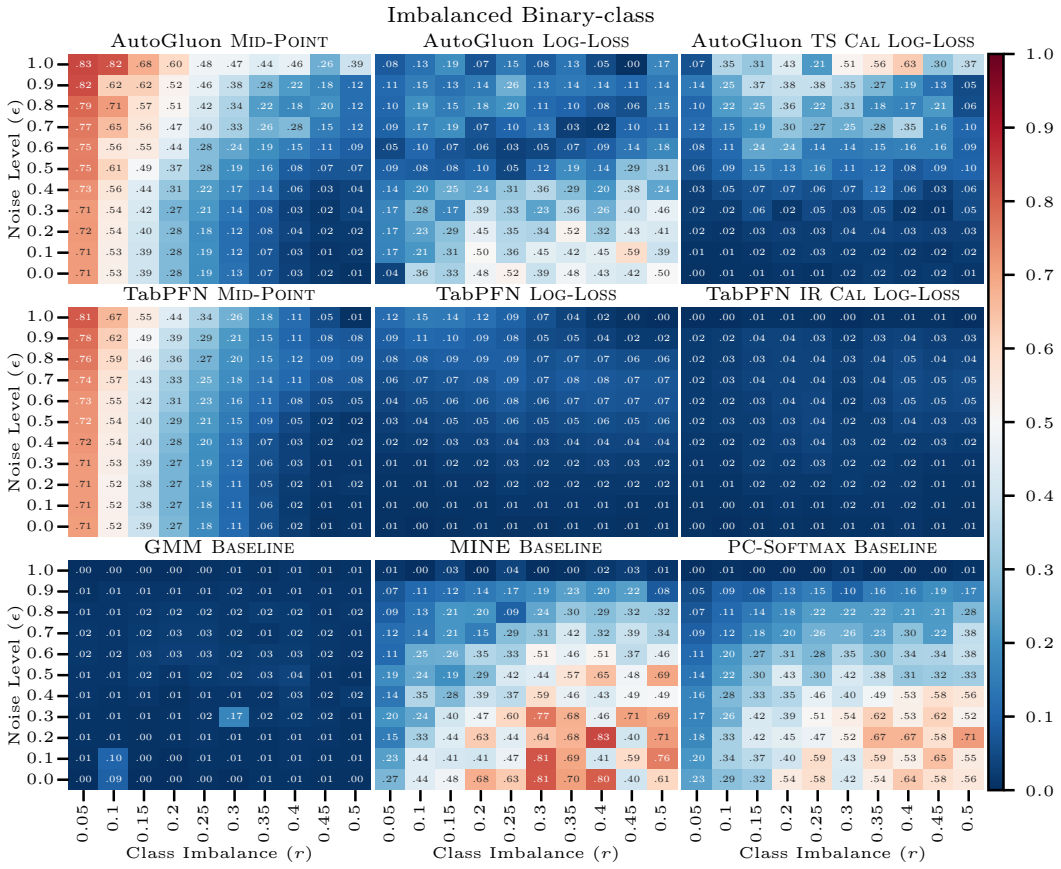


Figure 5: Generalizability on MVN proximity imbalanced binary-class datasets

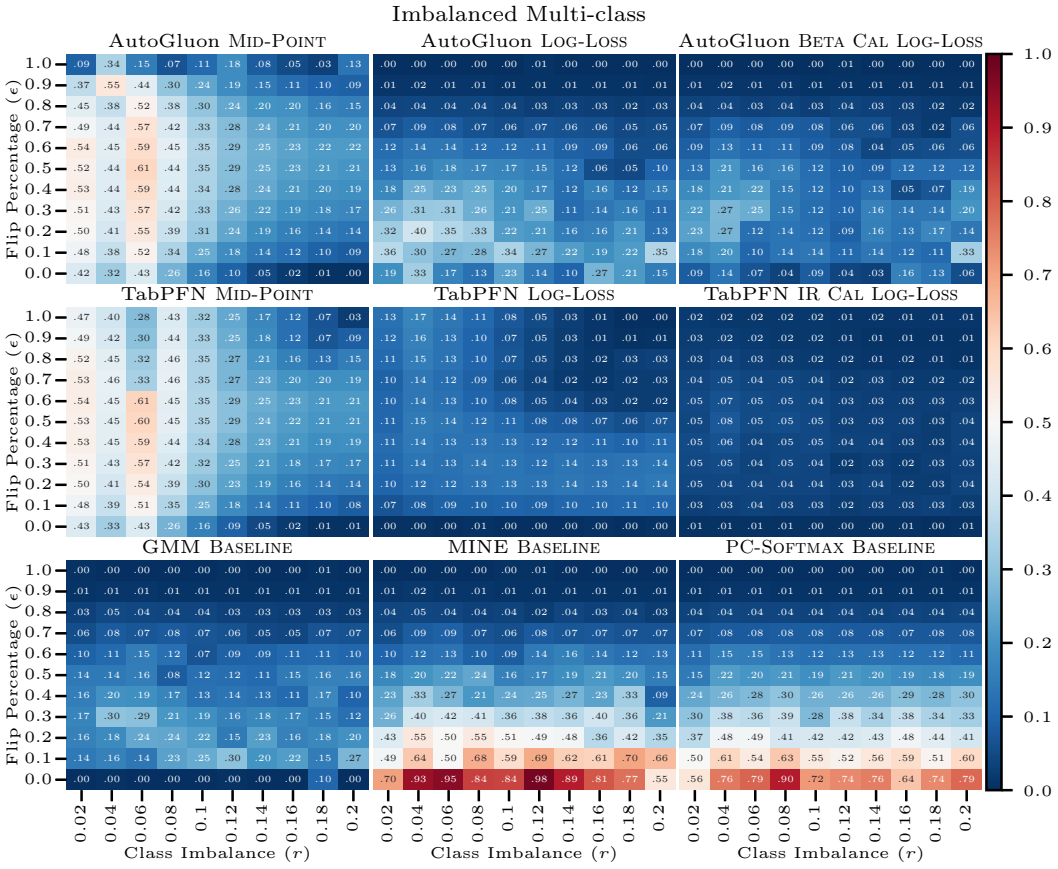


Figure 6: Generalizability on MVN perturbation imbalanced multi-class datasets

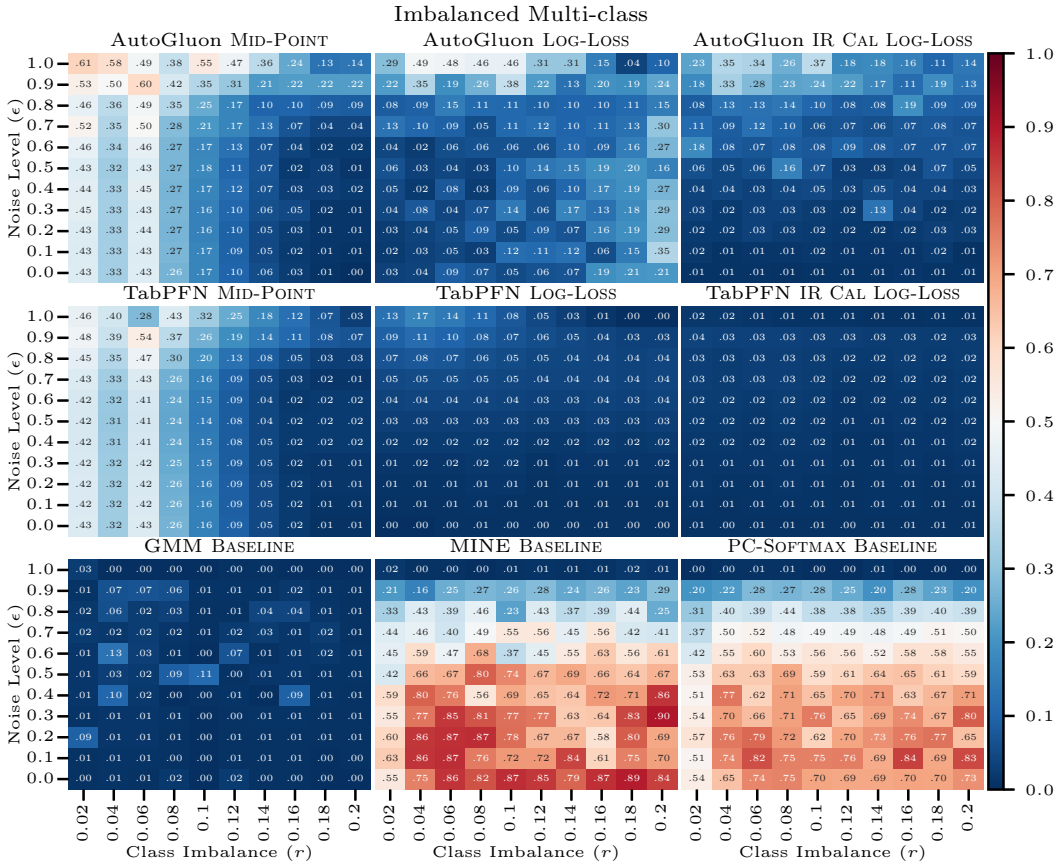


Figure 7: Generalizability on MVN proximity imbalanced multi-class datasets

TabPFN The Log-Loss and Cal Log-Loss methods using TabPFN show strong generalization across varying class imbalance and noise levels in both binary and imbalanced multi-class datasets. While TabPFN Mid-Point performs well under noise, it underperforms in highly imbalanced settings due to its tendency to overestimate MI [Gupta(2025), chap. 3]. Calibration (Cal Log-Loss) notably improves Log-Loss precision, especially in multi-class systems simulated using the perturbation technique.

AutoGluon MI estimation with AutoGluon shows mixed generalization in imbalanced datasets. Mid-Point performs worst under high imbalance ($r \leq 0.2$ binary, $r \leq 0.08$ multi-class) and elevated noise, often overestimating MI [Gupta(2025), chap. 3]. Log-Loss and Cal Log-Loss also overfit under certain imbalances and low-noise conditions, particularly with proximity-generated data, though less severe in multi-class settings. AutoGluon is notably sensitive to noise variation. Cal Log-Loss performs well at moderate noise levels in multi-class datasets, highlighting the benefits of calibration, but degrades estimation precision in non-vulnerable, imbalanced cases.

A.3 Baselines

The MINE and PC-Softmax baselines perform poorly on imbalanced binary or multi-class datasets, with generalization deteriorating as the imbalance and noise decrease. However, both methods estimate MI accurately in non-vulnerable synthetic datasets [Gupta(2025)]. GMM handles class imbalance and noise better, particularly in multi-class datasets generated via MVN perturbation, and slightly outperforms TabPFN in imbalanced binary-class cases. Its strong generalization

likely stems from the low dimensionality ($d = 5$) of the imbalanced datasets, highlighting that its limitations are primarily in high-dimensional scenarios, with minimal impact from noise and class imbalance.

A.4 Summary

TabPFN CAL LOG-LOSS consistently shows robust generalization in MI estimation across both perturbation and proximity datasets, with CAL LOG-LOSS often improving LOG-LOSS accuracy. In contrast, AutoGluon CAL LOG-LOSS improves MI estimates in vulnerable settings but overfits in non-vulnerable ones—explaining the high false-positive rate of AutoGluon CAL LOG-LOSS information leakage detection (ILD) approach, while TabPFN CAL LOG-LOSS ILD outperforms all other approaches [Gupta(2025)]. Baselines consistently struggle with high-dimensional, imbalanced, and noisy datasets, confirming their limitations [Gupta(2025)].

References

[Gupta(2025)] Gupta, P., 2025. Advanced Machine Learning Methods for Information Leakage Detection in Cryptographic Systems. Ph.D. thesis. Paderborn. URL: https://nbn-resolving.org/urn:nbn:de:hbz:466:2-54956. tag der Verteidigung: 09.05.2025.

[Maia Polo and Vicente(2022)] Maia Polo, F., Vicente, R., 2022. Effective sample size, dimensionality, and generalization in covariate shift adaptation. Neural Computing and Applications 35, 18187–18199. doi:10.1007/s00521-021-06615-1.