Information Leakage Detection through Approximate Bayes-optimal Prediction

Pritha Gupta[a,*], Marcel Wever[b], Eyke Hüllermeier[c]

[a]*Paderborn University, Paderborn, Germany*
[b]*L3S Research Center, Leibniz University Hannover, Hannover, Germany*
[c]*MCML, LMU Munich, Munich, Germany*

Abstract

In today's data-driven world, the proliferation of publicly available information raises security concerns due to the information leakage (IL) problem. IL involves unintentionally exposing sensitive information to unauthorized parties via observable system information. Conventional statistical approaches rely on estimating mutual information (MI) between observable and secret information for detecting ILs, face challenges of the curse of dimensionality, convergence, computational complexity, and MI misestimation. Though effective, emerging supervised machine learning based approaches to detect ILs are limited to binary system sensitive information and lack a comprehensive framework. To address these limitations, we establish a theoretical framework using statistical learning theory and information theory to quantify and detect IL accurately. Using automated machine learning, we demonstrate that MI can be accurately estimated by approximating the typically unknown Bayes predictor's Log-Loss and accuracy. Based on this, we show how MI can

*Corresponding author
Email addresses: prithag@mail.upb.de (Pritha Gupta), marcel.wever@ai.uni-hannover.de (Marcel Wever), eyke@lmu.de (Eyke Hüllermeier)

effectively be estimated to detect ILs. Our method performs superior to state-of-the-art baselines in an empirical study considering synthetic and real-world OpenSSL TLS server datasets.
Keywords: Information Leakage Detection, Mutual Information, Bayes-optimal Predictor, AutoML, Statistical Tests, Privacy

1. Introduction

The rapid proliferation of publicly available data, coupled with the increasing use of Internet of Things (IoT) technologies in today's data-driven world, has magnified the challenge of information leakage (IL), posing substantial risks to system security and confidentiality (Shabtai et al., 2012; Kelsey, 2002). IL occurs when sensitive or confidential information is inadvertently exposed to unauthorized individuals through observable system information (Hettwer et al., 2019). This problem can lead to severe consequences, ranging from potential electrical blackouts to the theft of critical information like medical records and military secrets, making the efficient detection and quantification of IL of paramount importance (Hettwer et al., 2019; Shabtai et al., 2012).

According to information theory, quantifying IL typically involves estimating mutual information (MI) between observable and secret information (Chatzikokolakis et al., 2010). Despite being a pivotal measure, MI is difficult to compute for high-dimensional data, facing challenges such as the *curse of dimensionality*, convergence, and computational complexity (Gao et al., 2015; Maia Polo and Vicente, 2022). Traditional statistical estimation methods often struggle with all of these challenges (Gao et al., 2015; Maia Polo and Vicente, 2022), more recent robust non-parametric approaches with improved

convergence rates still find high-dimensional scenarios challenging (Moon et al., 2021).

In recent years, machine learning (ML) techniques have gained popularity in information leakage detection (ILD), particularly for performing side-channel attacks (SCAs) on cryptographic systems (Picek et al., 2023). These systems release the *observable information* via many modes called the side-channels, such as network messages, CPU caches, power consumption, or electromagnetic radiation, which are exploited by SCAs to reveal secret inputs (secret keys, plaintexts), potentially rendering cryptographic protections ineffective (Moos et al., 2021; Hettwer et al., 2019). Therefore, detecting the existence of a side-channel is equivalent to uncovering IL (Hettwer et al., 2019). In this field, the most relevant literature uses ML to perform SCAs rather than preventing side-channels through early detection of ILs (Hettwer et al., 2019). Current ML-based methods in this realm detect side-channels to prevent SCAs and protect the system on both algorithmic and hardware levels (Mushtaq et al., 2018; Moos et al., 2021). These approaches leverage observable information to classify systems as vulnerable (with IL) or non-vulnerable (without IL) (Perianin et al., 2020; Mushtaq et al., 2018). They extract observable information from secure systems, categorizing them as non-vulnerable (labeled 0), then introduce known ILs to categorize them as vulnerable (labeled 1), creating a classification dataset for the learning model. However, this approach is limited to domain-specific scenarios and cannot be easily transferred to detect other unknown leakages (Perianin et al., 2020).

Recent promising ML-based methods proposed for estimating MI within classification datasets grapple with challenges related to convergence and com-

putational complexity (Cristiani et al., 2020), and others may underestimate MI or miss specific subclasses of IL (Qin and Kim, 2019). Recent advancements have demonstrated the effectiveness of ML-based techniques in directly detecting IL by analyzing the accuracy of the supervised learning models on extracted system data (Moos et al., 2021). Yet, these methods exhibit limitations in handling imbalanced and noisy real-world datasets, commonly encountered in practical scenarios, and tend to miss ILs by producing false negatives (Zhang et al., 2020; Picek et al., 2018).

To address these limitations, in our prior work, we proposed utilizing binary classifiers integrated with Fisher's exact test (FET) and paired t-test (PTT) statistical tests to account for imbalance (Gupta et al., 2022). To mitigate noise, an ensemble of binary classifiers, including a deep multilayer perceptron (MLP), along with their derived results (p-values) from the statistical tests, is aggregated using Holm-Bonferroni correction to enhance ILD accuracy and confidence. Despite its merits, this approach is limited to binary classification tasks and needs a comprehensive theoretical framework.

Our Contributions.

- We establish a comprehensive theoretical framework leveraging the connection between MI and the performance of the Bayes predictor to quantify IL using leakage assessment score and formalize its existence conditions in a system.
- We propose two MI estimation approaches by approximating the Bayes predictor induced using automated machine learning (AutoML) and demonstrate its effectiveness through a rigorous empirical evaluation.

- Using a cut-off on estimated MI through a one-sample t-test (OTT), we devise a technique for ILD. Furthermore, we propose using the Holm-Bonferroni correction on multiple models' estimates to enhance IL detection confidence by making it robust against noise and variations in AutoML pipelines' quality.
- We conduct an extensive empirical study, comparing our ILD methods against state-of-the-art approaches for detecting timing side-channels to counter Bleichenbacher's attacks.

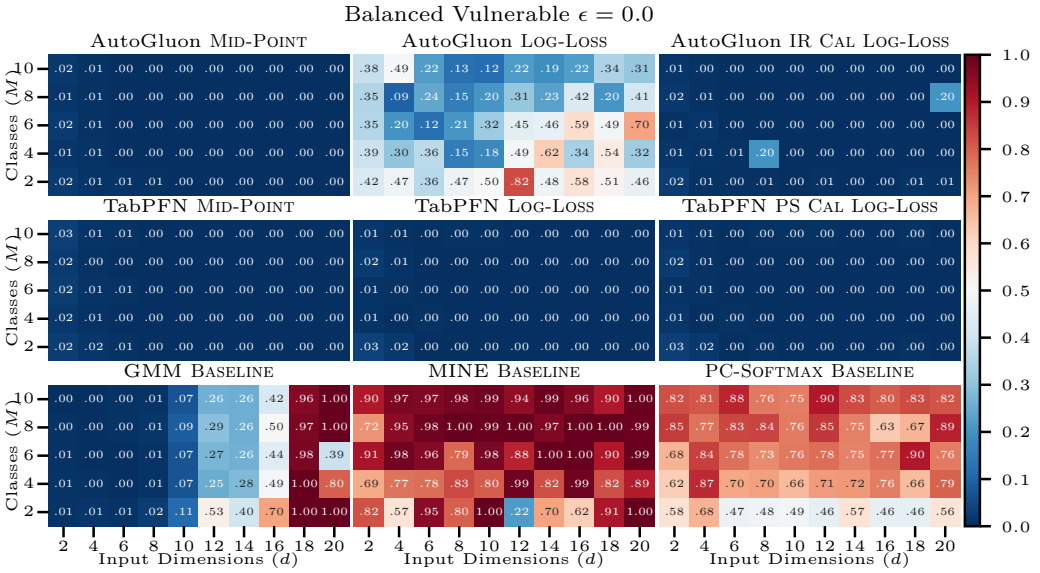
Appendix A. MI estimation methods

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. Appendix A.1), and across class imbalance (r) and noise level (ϵ) in binary and imbalanced multi-class datasets (c.f. Appendix A.2).

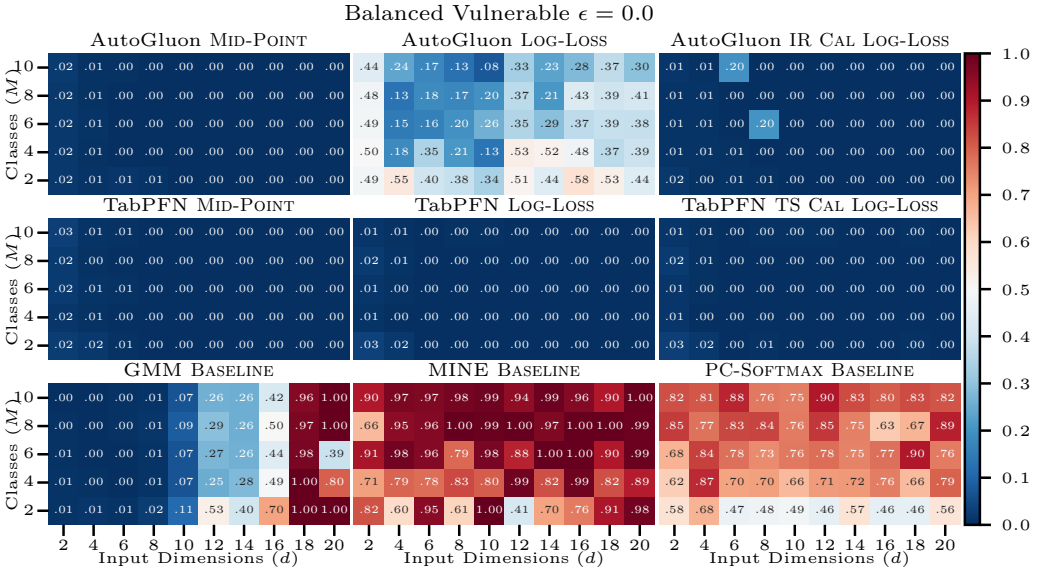
Appendix A.1. Number of Classes (C) and Input Dimensions (d)

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

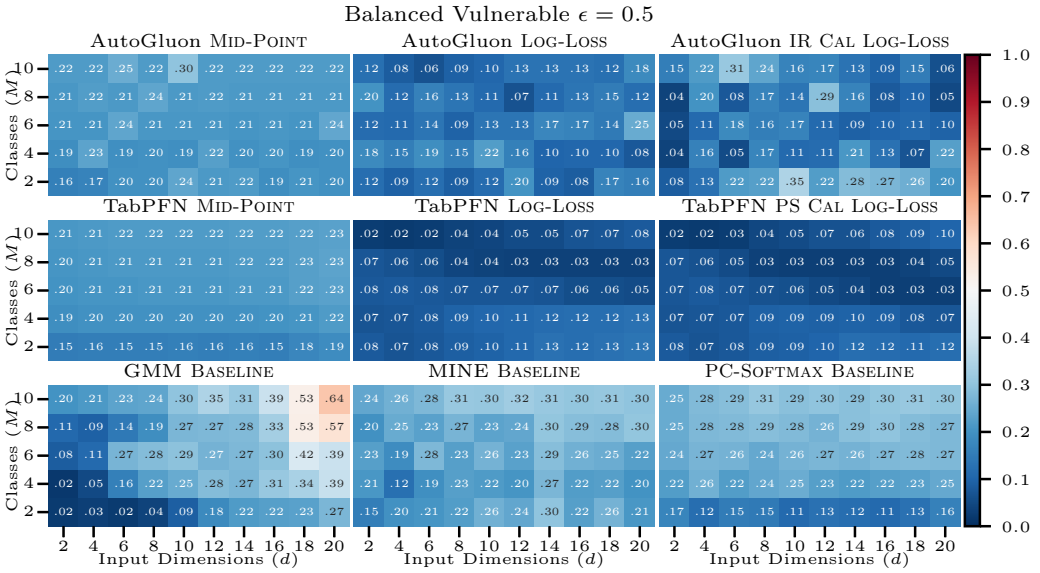


(a) Balanced multivariate normal (MVN) perturbation dataset with 0 % noise level

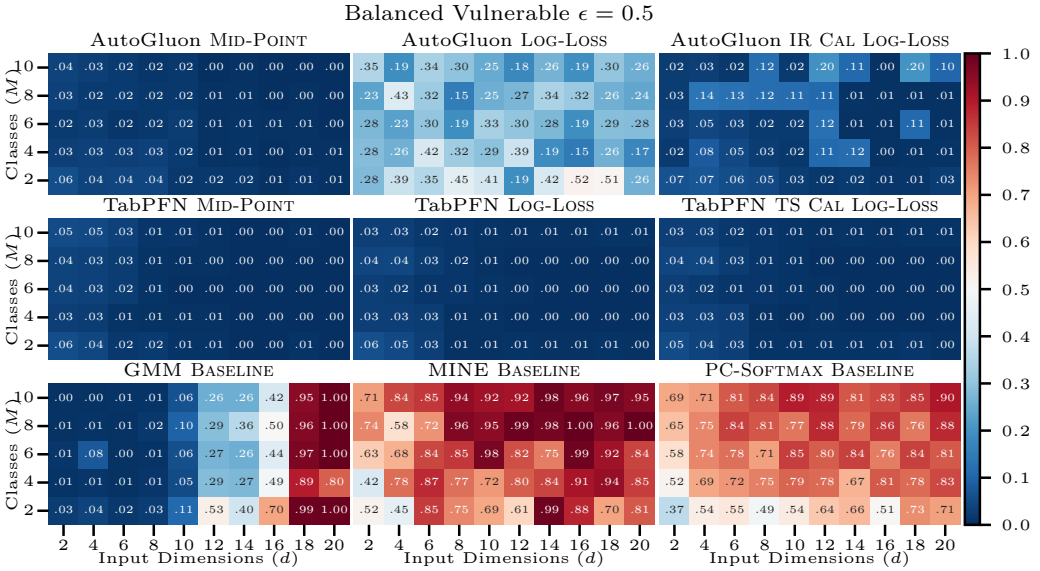


(b) Balanced MVN proximity dataset with 0 % noise level

Figure .1: Generalizability of MI estimation methods on noise-free vulnerable systems

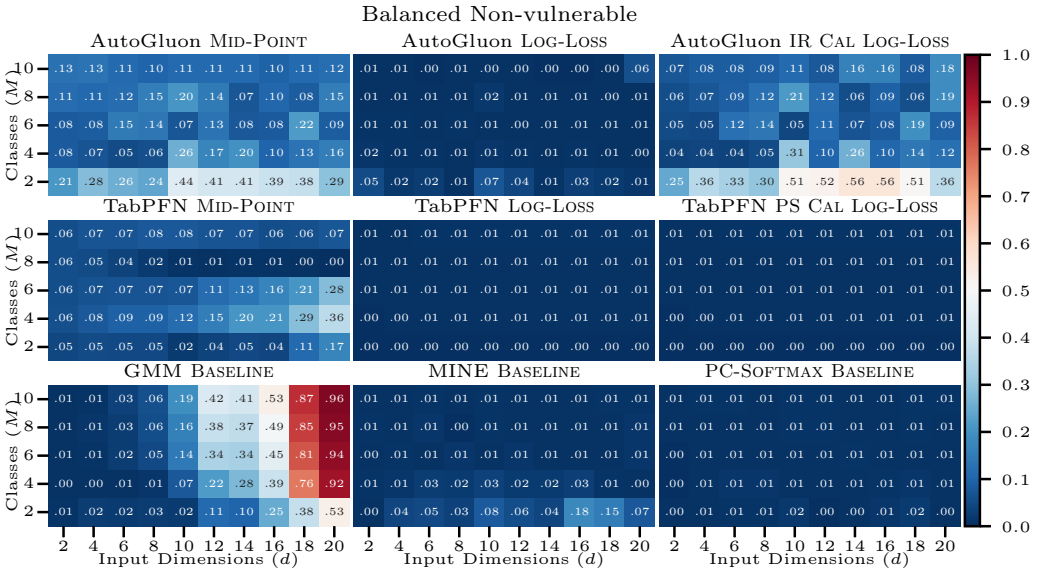


(a) Balanced MVN perturbation dataset with 50 % noise level

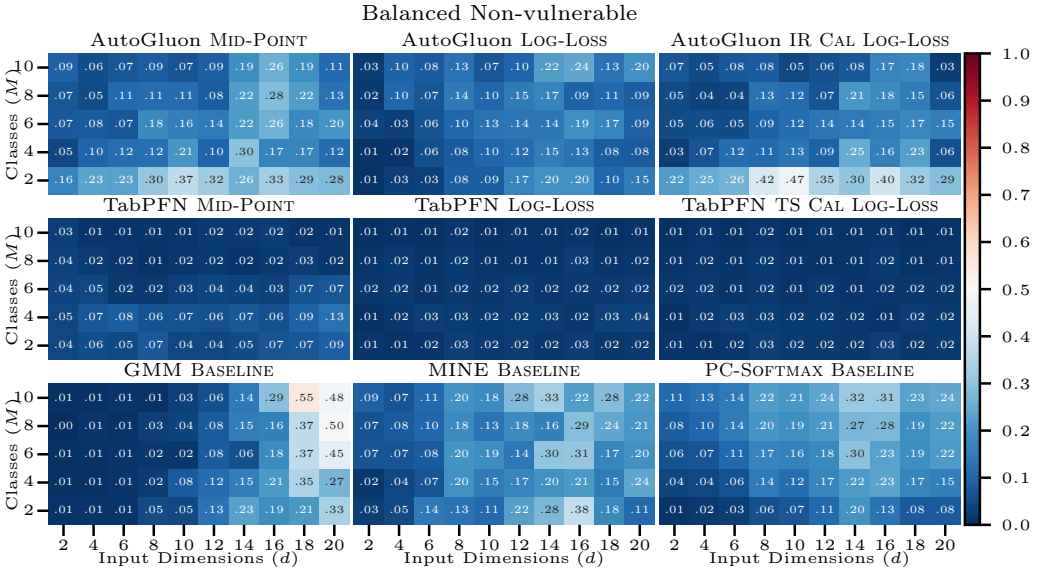


(b) Balanced MVN proximity dataset with 50 % noise level

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

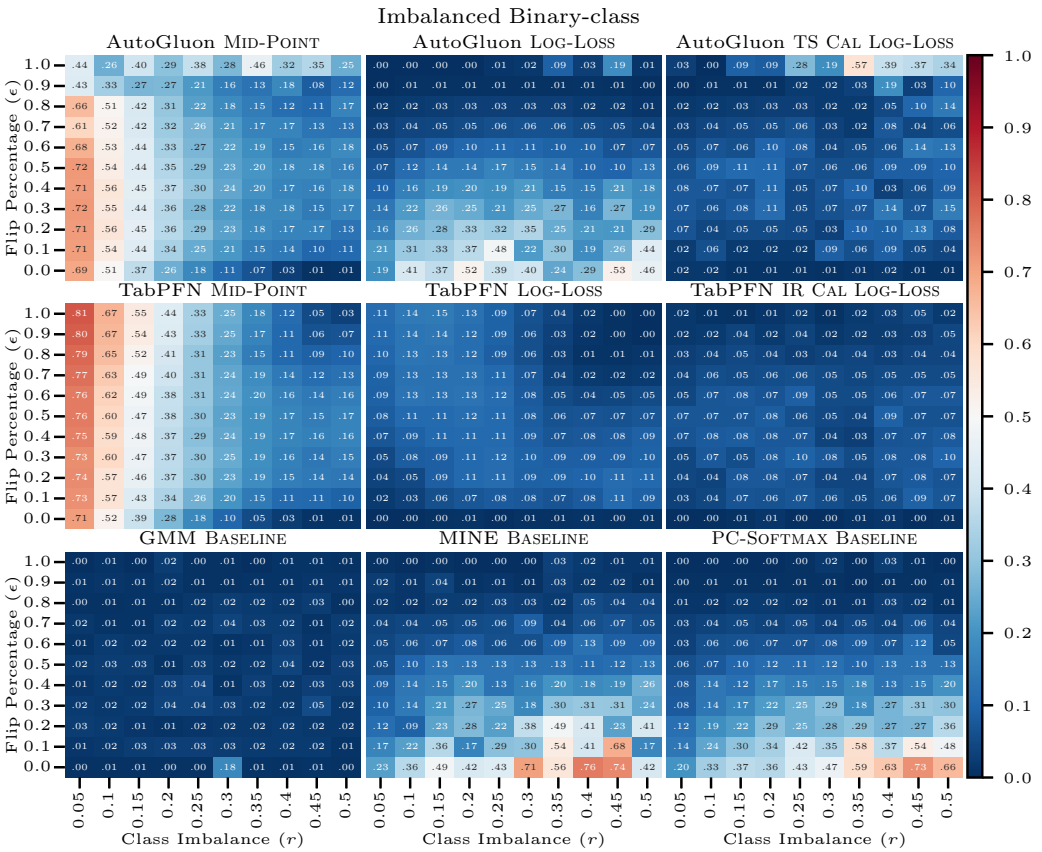


Figure A.4: Generalizability on MVN perturbation imbalanced binary-class datasets

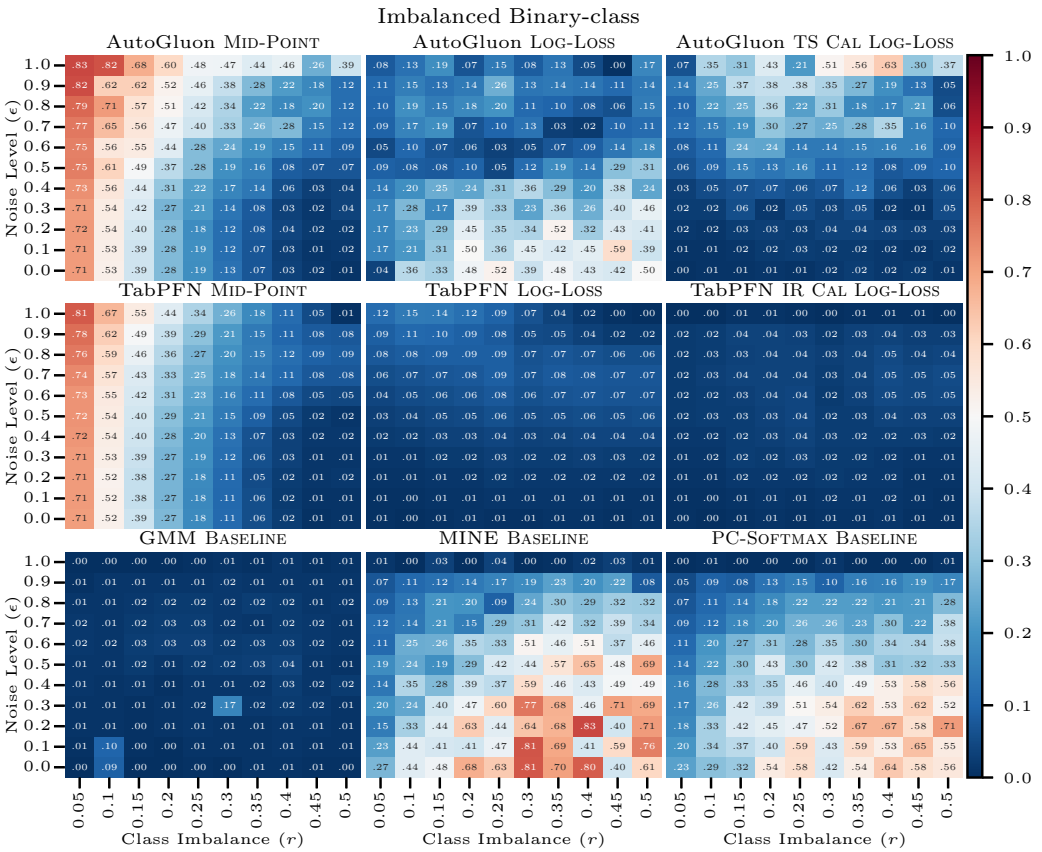


Figure A.5: Generalizability on MVN proximity imbalanced binary-class datasets

($d \geq 14$) vulnerable datasets, especially with MVN perturbations (NMAE ≈ 0.20).

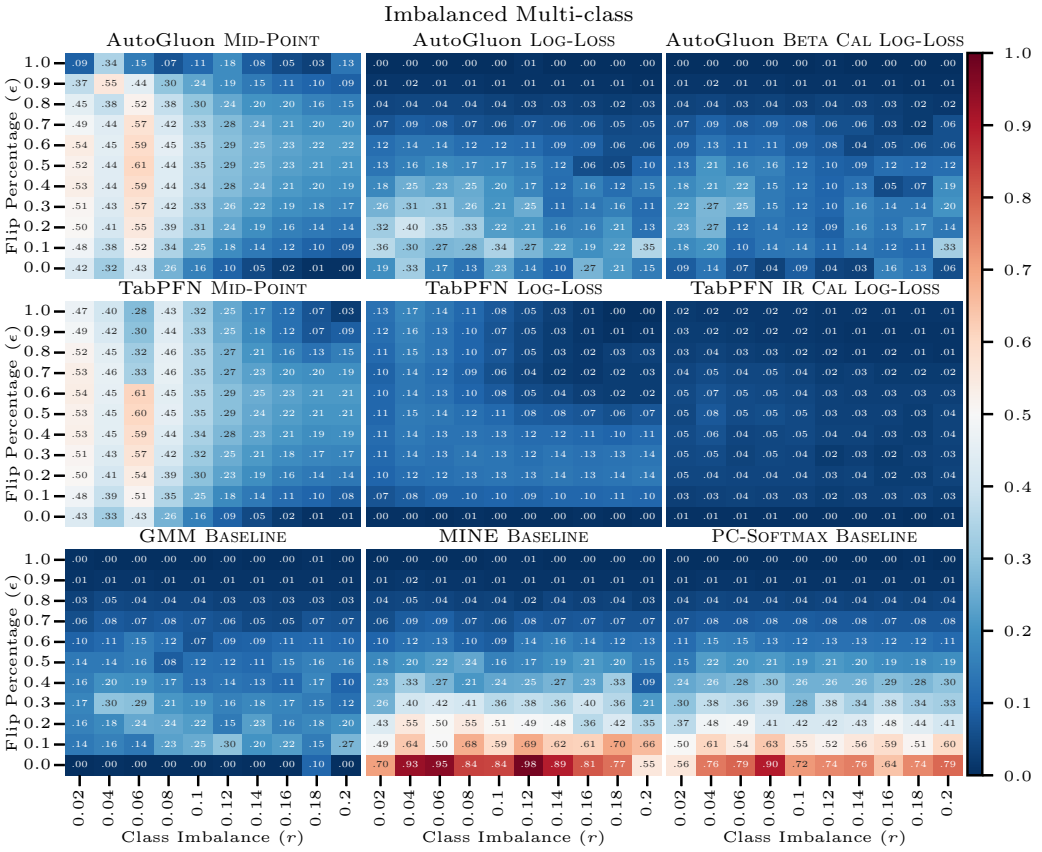


Figure A.6: Generalizability on MVN perturbation imbalanced multi-class datasets

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 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).

Appendix 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 A.4 and A.5 (binary) and Figures A.6 and A.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.

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.

Appendix 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.

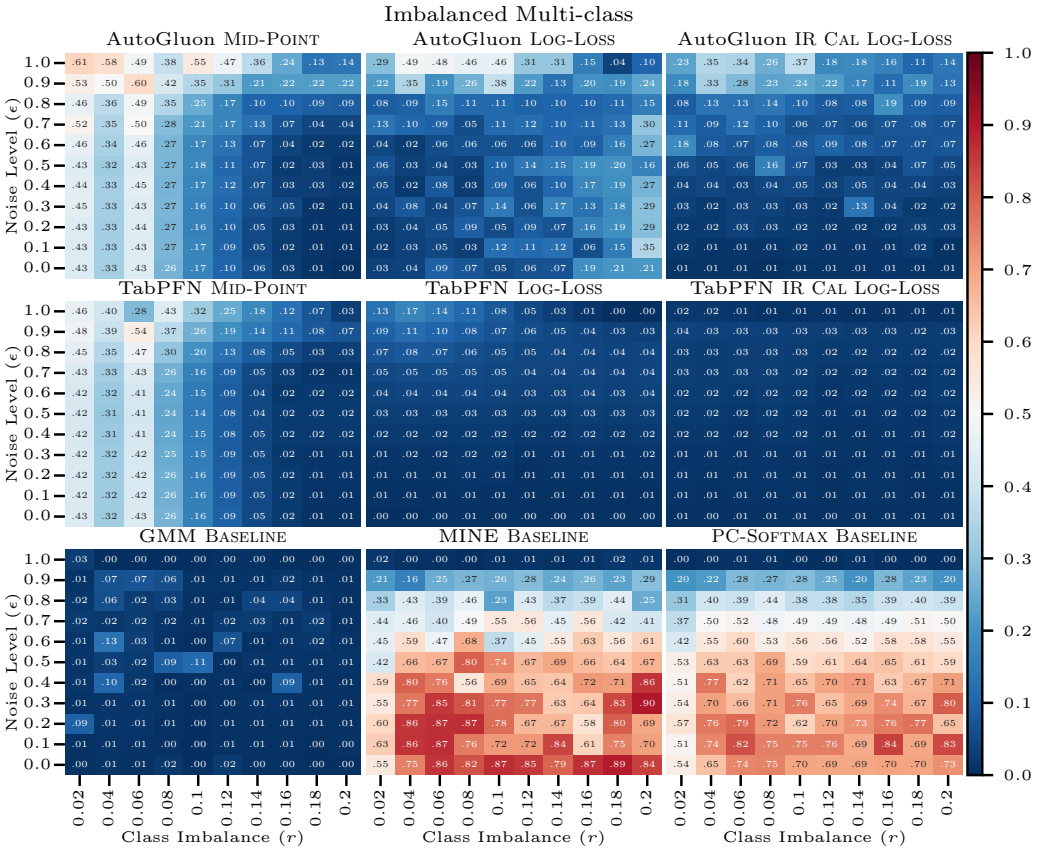


Figure A.7: Generalizability on MVN proximity imbalanced multi-class datasets

Appendix 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 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

Chatzikokolakis, K., Chothia, T., Guha, A., 2010. Statistical measurement of information leakage, in: Tools and Algorithms for the Construction and Analysis of Systems, Springer Berlin Heidelberg, Berlin, Heidelberg. pp. 390–404.

Cristiani, V., Lecomte, M., Maurine, P., 2020. Leakage assessment through neural estimation of the mutual information, in: Lecture Notes in Computer Science. Springer International Publishing, Berlin, Heidelberg, pp. 144–162. doi:10.1007/978-3-030-61638-0_9.

Gao, S., Ver Steeg, G., Galstyan, A., 2015. Efficient Estimation of Mutual Information for Strongly Dependent Variables, in: Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, PMLR, San Diego, California, USA. pp. 277–286.

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.

Gupta, P., Ramaswamy, A., Drees, J., Hüllermeier, E., Priesterjahn, C., Jager, T., 2022. Automated information leakage detection: A new method combining machine learning and hypothesis testing with an application

to side-channel detection in cryptographic protocols, in: Proceedings of the 14th International Conference on Agents and Artificial Intelligence, INSTICC. SCITEPRESS - Science and Technology Publications, Virtual Event. pp. 152–163. doi:10.5220/0010793000003116.

Hettwer, B., Gehrer, S., Güneysu, T., 2019. Applications of machine learning techniques in side-channel attacks: A survey. Journal of Cryptographic Engineering 10, 135–162. doi:10.1007/s13389-019-00212-8.

Kelsey, J., 2002. Compression and information leakage of plaintext, in: Fast Software Encryption, Springer Berlin Heidelberg, Berlin, Heidelberg. pp. 263–276.

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.

Moon, K.R., Sricharan, K., Hero, A.O., 2021. Ensemble estimation of generalized mutual information with applications to genomics. IEEE Transactions on Information Theory 67, 5963–5996. doi:10.1109/TIT.2021.3100108.

Moos, T., Wegener, F., Moradi, A., 2021. DL-LA: Deep learning leakage assessment. IACR Transactions on Cryptographic Hardware and Embedded Systems 2021, 552–598. doi:10.46586/tches.v2021.i3.552-598.

Mushtaq, M., Akram, A., Bhatti, M.K., Chaudhry, M., Lapotre, V., Gogniat, G., 2018. NIGHTs-WATCH: A cache-based side-channel intrusion detector using hardware performance counters, in: Proceedings of the 7th International Workshop on Hardware and Architectural Support for Security

and Privacy, Association for Computing Machinery, New York, NY, USA. doi:10.1145/3214292.3214293.

Perianin, T., Carré, S., Dyseryn, V., Facon, A., Guilley, S., 2020. End-to-end automated cache-timing attack driven by machine learning. Journal of Cryptographic Engineering 11, 135–146. doi:10.1007/s13389-020-00228-5.

Picek, S., Heuser, A., Jovic, A., Bhasin, S., Regazzoni, F., 2018. The curse of class imbalance and conflicting metrics with machine learning for side-channel evaluations. IACR Transactions on Cryptographic Hardware and Embedded Systems 2019, 209–237. doi:10.13154/tches.v2019.i1.209-237.

Picek, S., Perin, G., Mariot, L., Wu, L., Batina, L., 2023. Sok: Deep learning-based physical side-channel analysis. ACM Computing Surveys 55. doi:10.1145/3569577.

Qin, Z., Kim, D., 2019. Rethinking softmax with cross-entropy: Neural network classifier as mutual information estimator. CoRR abs/1911.10688. arXiv:1911.10688.

Shabtai, A., Elovici, Y., Rokach, L., 2012. A Survey of Data Leakage Detection and Prevention Solutions. 1 ed., Springer US, New York, NY. doi:10.1007/978-1-4614-2053-8.

Zhang, J., Zheng, M., Nan, J., Hu, H., Yu, N., 2020. A novel evaluation metric for deep learning-based side channel analysis and its extended application to imbalanced data. IACR Transactions on Cryptographic Hardware and Embedded Systems 2020, 73–96. doi:10.46586/tches.v2020.i3.73-96.