THESIS SUMMARY

Balancing Utility and Privacy: Evaluating the Privacy-Preserving Capabilities of Synthetic Data from Generative Models

Julia L. Wang[a], **Yuri Lawryshyn**[b], **Lucy Liu**[c]
[a]University of Toronto, Faculty of Applied Science and Engineering, Department of Engineering Science
[b]University of Toronto, Faculty of Applied Science and Engineering, Department of Chemical Engineering
[c]Royal Bank of Canada

Abstract. This thesis investigates the privacy-preserving properties of synthetic financial datasets generated by advanced generative AI models. In the age of big data, where financial institutions increasingly rely on machine learning for decision-making, protecting personal information is a significant consideration. This work assesses four prominent generative models: CTGAN, TVAE, DoppelGANger, and Banksformer. The research methodology includes comprehensive privacy evaluation metrics, such as re-identification risk, attribute disclosure risk, and susceptibility to membership inference attacks. The results reveal the strengths and limitations of each model in terms of privacy protection and demonstrate that while some models maintain higher fidelity to the original data distributions, others offer stronger privacy guarantees by deviating more significantly from the real datasets. Notably, Banksformer and DoppelGANger reveal the best performance concerning utility, however, there are higher risks of privacy breaches. Concluding that the current generative models exhibit varying levels of privacy preservation, the thesis advocates for future research to establish a quantifiable balance between data utility and privacy.

Keywords: privacy, generative AI, tabular data, time-series data, machine learning.

1 Introduction

Recent advancements in deep learning and machine learning (ML) have highlighted the need for privacy-preserving mechanisms in the processing of sensitive personal information (PI). As data-driven methodologies become increasingly integral to decision-making across various sectors, the importance of implementing ethical data practices to protect PI is underscored. This need is particularly vital in the financial sector, where the privacy and confidentiality of customer data are crucial. Techniques such as Conditional Tabular Generative Adversarial Networks (CTGAN) [1], Tabular Variational Autoencoder (TVAE) [2], and others represent current SOTA generative models for synthesizing tabular series data, facilitating the creation of diverse and realistic datasets. Particularly noteworthy are models like Banksformer [3] and DoppelGANger [4], which excel in generating synthetic financial time-series data, catering to the unique requirements of the financial sector. Financial institutions stand to gain immensely from harnessing the potential of such techniques, but this potential is contingent upon the implementation of measures to protect the sensitive data underpinning these models. The importance of this research lies in its potential to advance the field of privacy-preserving machine learning and to provide financial institutions with the means to securely employ advanced techniques, reconciling the objectives of maximizing data utility and upholding privacy standards. As such, this thesis investigates how SOTA generative models manage and mask sensitive financial information within synthetic datasets. We evaluate the effectiveness of these models in striking a balance between maintaining data utility for ML applications and ensuring the anonymity of underlying PI. This research sheds light on the nuanced interplay between data utility and privacy preservation, offering insights crucial for navigating the ethical and practical dimensions of data-driven decision-making in the financial realm.

2 Previous Work

In response to evolving privacy concerns and rapid advancements in ML, recent research has focused on leveraging generative AI models to generate synthetic datasets which protect sensitive data. However, there are diverse strategies used by malicious actors to compromise data confidentiality of these new datasets including identification, inference, and linkage attacks [5]. Recent research in privacy-preserving methods have been employed in combination with ML models, such as homomorphic encryption (e.g. [6, 7, 8, 9]), and differential privacy (e.g. [10, 11, 12, 13]), offer frameworks for maintaining data privacy while balancing utility. Within the realm of generative models, several state-of-the-art techniques have emerged for generating synthetic data which may be utilized against privacy attacks.

Specifically for tabular data applications, CTGAN [1] tackles challenges in generating tabular data, such as imbalanced data and mixed types, by employing conditional generation and mode-specific normalization techniques. Its architecture, featuring fully connected networks and Wasserstein GAN with gradient penalty for stability, enables the generation of synthetic data under specific conditions, making it suitable for diverse datasets. Additionally, TVAE [2] enhances the traditional VAE framework with deep metric learning, utilizing triplet loss to capture finer-grained information in latent space for generating tabular data. By optimizing latent vectors to be more informative and semantically structured, TVAE ensures the preservation of semantic similarities and detailed relationships within the synthetic datasets it generates.

Specific to time-series and tabular synthetic data generation are models DoppelGANger [4] and Banksformer [3]. DoppelGANger focuses on generating networked time series data, leveraging GANs to produce realistic sequences with metadata while addressing challenges such as mode collapse. Its dual-generator and dual-discriminator architecture, complemented by techniques like batched RNN generation and auto-normalization, ensure the generation of high-fidelity synthetic data that preserves the intricate relationships and dynamics within the original dataset. Banksformer utilizes transformer models within a GAN framework to generate synthetic financial time-series data, effectively capturing complex temporal relationships and market dynamics. Its unique preprocessing step translates timestamps into multiple features, enhancing the model's capability to learn and replicate date-based transaction patterns, making it suitable for synthesizing financial data while maintaining data utility and privacy.

3 Methodology

This study investigates the privacy preservation capabilities of synthetic datasets generated by deep learning models, focusing on a real anonymized transaction dataset from Czech banks [14]. The dataset comprises over 1 million transactions from 1993 to 1998, with key features utilized being account ID, date, type, operation, amount, and k-symbol. Models [1, 2, 3, 4] were utilised to generate 4 synthetic datasets which were evaluated with the metrics described in the following sections.

Statistical similarity metrics were used compare the distributions of features between real and synthetic datasets to assess their similarity. They include Jensen-Shannon divergence [15] for categorical columns, entropy differences [16] for categorical columns, and Wasserstein distance [17] for numerical columns.

Privacy evaluation metrics were similarly used as a baseline to evaluate privacy preservation including K-Anonymity, L-Diversity, and T-Closeness. K-Anonymity [18] measures the level of de-identification in a dataset by ensuring that each record is indistinguishable from at least $k-1$ others based on quasi-identifiers like transaction types or operations. L-Diversity [19] extends this concept by ensuring diversity in sensitive attributes within indistinguishable groups, protecting against attribute disclosure. T-Closeness [20] further ensures that the distribution of sensitive attributes within groups remains close to the overall dataset distribution, minimizing the risk of inference attacks based on statistical analysis.

Risk metrics were utilized, including re-identification risk which measures the probability that an individual's record in a dataset can be traced back to them, compromising their privacy. It involves comparing frequency distributions of categorical variables and using distance metrics including the K-S Test [21] for numerical variables between real and synthetic datasets. Attribute disclosure risk similarly assesses the likelihood of inferring sensitive information about individuals from a dataset. Predictive modeling with Random Forest [22] was utilized to evaluate how well sensitive attributes can be predicted from other attributes in the dataset, indicating the degree of privacy protection.

Membership inference attacks aim to determine if a particular data record was used in the training set of a model, compromising individual privacy. This metric involves training binary classifiers (logistic regression, Naive Bayes, Random Forest, XGBOOST, GBM, and a feedforward neural network) to differentiate between real and synthetic records by simulating a labelled dataset

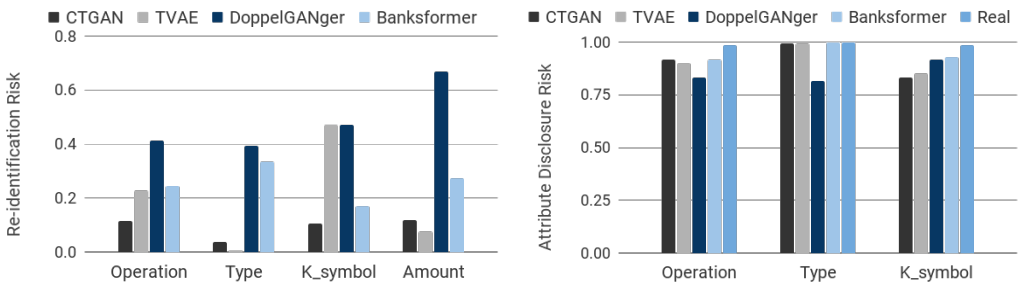

Figure 1: (a) Re-identification risk assessment and (b) attribute disclosure risk

a malicious attacker may have access to with n data points. Various n were evaluated on balanced accuracy and f1 scores with 5-fold cross-validation. Note that high success rates of classification here indicate a higher risk of re-identification and privacy breaches.

4 Results

In the statistical analysis distributions, CTGAN and TVAE demonstrated minimal differences in distribution metrics, indicating close statistical similarity to the real data while maintaining a lower risk of privacy breaches. Conversely, Banksformer and DoppelGANger exhibited significantly higher values in these metrics, suggesting a departure from the real data distribution. Note that when comparing the models on their time-series data generation, Banksformer was the best model at synthesizing the relationship between temporal and non-temporal transactional features.

Privacy analysis metrics indicated variations in privacy preservation capabilities across models. While CTGAN and TVAE showed lower K-Anonymity and L-Diversity compared to the real dataset, Banksformer demonstrated a more balanced approach with moderate values. DoppelGANger exhibited the lowest T-Closeness, indicating a significant departure from the real data distribution.

The risk assessments in Figure 1 highlighted significant differences in the models' ability to mimic the original data distribution while preserving privacy. For re-identification risk, CTGAN showed relatively minor differences in both categorical and numerical distributions, potentially mitigating re-identification risks. DoppelGANger presented the highest distribution differences, indicating stronger obfuscation of dataset characteristics but potentially impacting data utility. Attribute disclosure risk assessments revealed a reduction in prediction accuracy across all the generated synthetic data compared to real data, indicating improved privacy preservation moving from real to synthetic data. However, predictability for certain attributes persisted across models, suggesting the need for further refinement in obfuscating sensitive information.

The membership inference attack results demonstrated that Banksformer and DoppelGANger exhibited lower levels of privacy preservation. As the number of data points available to the attacker (n) increased, the accuracies of all classifiers converged, particularly around $n = 200$ for each synthetic dataset. This convergence indicated that beyond this threshold, additional data did not significantly improve the attacker's ability to distinguish between real and synthetic data, highlighting a potential limit in the usefulness of additional data for membership inference attacks. The results displayed in Figure 2 are the final balanced accuracies for which each model converged to.

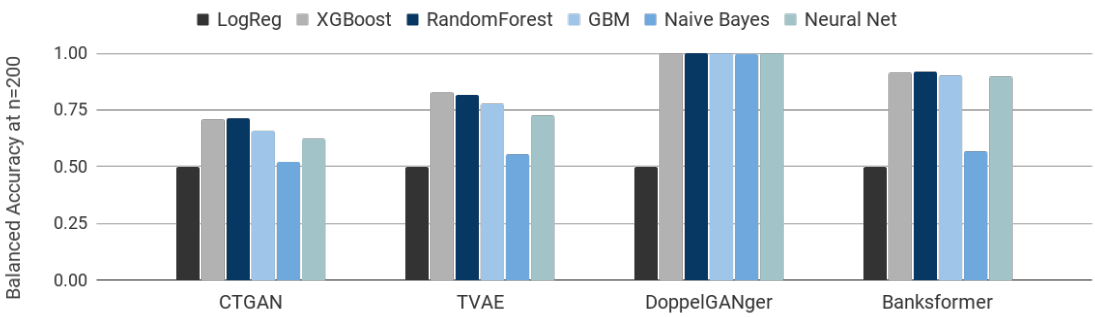

Figure 2: Membership inference attack balanced accuracy convergence

Note that logistic regression tends to overfit and predict a single class, as indicated by a consistent 50% balanced accuracy across all synthetic datasets. Aside from logistic regression, Banksformer and DoppelGANger exhibited high classification accuracies of 90% and 100% respectively across almost all classifiers, indicating significant privacy concerns. The rapid convergence towards high accuracy suggested that these models retained too many characteristics from the training data, making it easier for an attacker to identify the synthetic data as originating from real data. Such high accuracies far exceed the 50% threshold that would indicate a good balance between data utility and privacy, and are indicative of a heightened risk of re-identification in the synthetic data generated by Banksformer and DoppelGANger. In contrast, CTGAN and TVAE showed more favorable results for privacy preservation, with their metrics indicating better resistance to attacks. The balanced accuracies for CTGAN and TVAE converged within the range of 50% to 75%, suggesting a stronger performance in terms of privacy preservation.

5 Conclusions and Recommendations

This thesis comprehensively evaluated the privacy-preserving characteristics of synthetic data generated by four distinct models: CTGAN, TVAE, DoppelGANger, and Banksformer. Through analyses employing various differential privacy metrics, re-identification risk assessments, attribute disclosure risk evaluations, and membership inference attack simulations, the research illuminated the varying degrees of privacy preservation these models afford. CTGAN and TVAE generally demonstrated higher fidelity to the original data's distribution while suggesting a lower risk of privacy breaches. Notably, membership inference attacks highlighted the susceptibility of Banksformer and DoppelGANger to privacy compromises, as evidenced by their tendency towards high classification accuracies. These findings illustrate the need for balancing the utility of synthetic data with individuals' privacy concerns. Establishing a quantitative threshold defining this balance is a key avenue for future research, guiding the development of generative models towards outputs that are both analytically useful and privacy-compliant. Future research directions could also explore integrating differential privacy mechanisms directly into the generative process of these models, potentially enhancing the capability to generate data representative for analysis while simultaneously protecting against re-identification. In conclusion, continued research and development are needed to refine generative models for synthetic data generation, effectively balancing the dual imperatives of data utility and privacy.

References

[1] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional gan," in *Advances in Neural Information Processing Systems*, 2019.

[2] H. Ishfaq, A. Hoogi, and D. Rubin, "Tvae: Triplet-based variational autoencoder using metric learning," 2023.

[3] K. Nickerson, T. Tricco, A. Kolokolova, F. Shoeleh, C. Robertson, J. Hawkin, and T. Hu, "Banksformer: A deep generative model for synthetic transaction sequences," in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2022, Grenoble, France, September 19–23, 2022, Proceedings, Part VI*. Berlin, Heidelberg: Springer-Verlag, 2023, p. 121–136.

[4] Z. Lin, A. Jain, C. Wang, G. Fanti, and V. Sekar, "Generating high-fidelity, synthetic time series datasets with doppelganger," *CoRR*, vol. abs/1909.13403, 2019. [Online]. Available: http://arxiv.org/abs/1909.13403

[5] B. Liu, M. Ding, S. Shaham, W. Rahayu, F. Farokhi, and Z. Lin, "When machine learning meets privacy: A survey and outlook," *ACM Computing Surveys*, vol. 5, no. 2, Mar. 2021, doi: 10.1145/3436755.

[6] F.-J. González-Serrano, Á. Navia-Vázquez, and A. Amor-Martín, "Training support vector machines with privacy-protected data," *Pattern Recognition*, vol. 72, pp. 93–107, 2017.

[7] M. A. Almaiah, A. Ali, F. Hajjej, M. F. Pasha, and M. A. Alohali, "A lightweight hybrid deep learning privacy preserving model for fc-based industrial internet of medical things," *Sensors*, vol. 22, no. 6, p. 2112, 2022.

[8] J. Ma, S.-A. Naas, S. Sigg, and X. Lyu, "Privacy-preserving federated learning based on multi-key homomorphic encryption," *International Journal of Intelligent Systems*, vol. 37, no. 9, pp. 5880–5901, 2022.

[9] A. Fu, X. Zhang, N. Xiong, Y. Gao, H. Wang, and J. Zhang, "Vfl: A verifiable federated learning with privacy-preserving for big data in industrial iot," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 5, pp. 3316–3326, 2020.

[10] J. Wang, J. Zhang, W. Bao, X. Zhu, B. Cao, and P. S. Yu, "Not just privacy: Improving performance of private deep learning in mobile cloud," in *Proceedings of the 24th ACM SIGKDD international conference*, 2018, pp. 2407–2416.

[11] T. Zhang, Z. He, and R. B. Lee, "Privacy-preserving machine learning through data obfuscation," *arXiv preprint arXiv:1807.01860*, 2018.

[12] X. Lu, Y. Liao, P. Lio, and P. Hui, "Privacy-preserving asynchronous federated learning mechanism for edge network computing," *IEEE Access*, vol. 8, pp. 48 970–48 981, 2020.

[13] P. Kairouz, B. McMahan, S. Song, O. Thakkar, A. Thakurta, and Z. Xu, "Practical and private (deep) learning without sampling or shuffling," 2021.

[14] L. Petrocelli, "Czech financial dataset: Real anonymized transactions," https://data.world/lpetrocelli/czech-financial-dataset-real-anonymized-transactions.

[15] B. Fuglede and F. Topsoe, "Jensen-shannon divergence and hilbert space embedding," in *International symposium onInformation theory, 2004. ISIT 2004. Proceedings.* IEEE, 2004, p. 31.

[16] T. Li, S. Ma, and M. Ogihara, "Entropy-based criterion in categorical clustering," in *Proceedings of the twenty-first international conference on Machine learning*, 2004, p. 68.

[17] V. M. Panaretos and Y. Zemel, "Statistical aspects of wasserstein distances," *Annual review of statistics and its application*, vol. 6, pp. 405–431, 2019.

[18] L. Sweeney, "k-anonymity: A model for protecting privacy," *International journal of uncertainty, fuzziness and knowledge-based systems*, vol. 10, no. 05, pp. 557–570, 2002.

[19] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," *Acm transactions on knowledge discovery from data (tkdd)*, vol. 1, no. 1, pp. 3–es, 2007.

[20] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *2007 IEEE 23rd international conference on data engineering*. IEEE, 2007, pp. 106–115.

[21] D. S. Dimitrova, V. K. Kaishev, and S. Tan, "Computing the kolmogorov-smirnov distribution when the underlying cdf is purely discrete, mixed, or continuous," *Journal of Statistical Software*, vol. 95, no. 10, p. 1–42, 2020. [Online]. Available: https://www.jstatsoft.org/index.php/jss/article/view/v095i10

[22] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, p. 5–32, oct 2001. [Online]. Available: https://doi.org/10.1023/A:1010933404324