Reconstructability of Training Data Enhances with Increasing Spatial Correlation

Arda Hüseyinoglu[1], Nicole Martin[1], and Erencan Uysal[1]

Technical University of Munich, Arcisstraße 21, 80333 München, Germany

Abstract. Ensuring the privacy and security of machine learning tools, particularly concerning the confidentiality of their training data, is of most importance. Current research reveals the potential for reconstructing training data from trained neural networks, raising concerns about inversion attacks on sensitive datasets, including genetic data. This paper contributes by exploring the applicability of novel gray-box approaches for reconstructing training data from trained cancer predictors. We present a pipeline designed for inversion attacks on trained cancer predictors, demonstrating that reconstructability enhances with increasing spatial correlation, as observed in medical images rather than genetic data.

Keywords: Inversion Attacks · Genetic Data · Medical Images · Spatial Correlation · Deep Learning

1 Introduction

With rapid advances in machine learning and the increasing availability of data, privacy and security concerns have emerged as significant challenges. One such concern are model inversion attacks, a class of privacy violations in which attackers attempt to reconstruct sensitive training data from the parameters of a trained model. While model inversion attacks have been extensively studied in areas such as natural language processing and image recognition, their impact in the field of genetic data has been relatively understudied [4].

In this article, we test the privacy of genetic machine learning models, focusing particularly on model inversion attacks. Genetic data is highly personal and sensitive information, which holds potential for medical research, diagnosis, and personalized treatments. However, the utilization of machine learning models on this data introduces novel privacy risks, raising critical questions about the security of individuals' genetic information.

2 Related Work

Reconstructing Training Data: The association of overparameterized neural networks, with associative memory capabilities, as explored by Radhakrishnan et al. [6], sheds light on the inherent memorization properties of such networks. This work suggests that overparameterized auto-encoders can store training samples

as attractors, offering a theoretical foundation for understanding the memorization and retrieval mechanisms in neural networks. The feasibility of reconstructing training data from model gradients, as presented by Wang et al., introduces a provable inversion attack method that poses severe privacy threats, particularly in learning scenarios. Their study provides a mathematical foundation for understanding the conditions under which training data can be reconstructed from gradients [7]. Wang et al. further explore model inversion attacks across multiple models, proposing an ensemble inversion technique that enhances the quality of reconstructed training data [8].

Building on the mathematical foundations laid out by Lyu and Li [2] and Ji and Telgarsky [1], Haim et al. demonstrated that a significant portion of training data could be reconstructed from just the parameters of a trained neural network classifier, leveraging the implicit bias in training neural networks with gradient-based methods [9]. The mathematical foundation assumes a homogenous ReLU network which is trained using gradient flow for binary classification. If at some point the loss function is close to zero, then the gradient flow converges in direction to a first-order stationary point (KKT point) of the optimization problem. This means, that there is an implicit bias of the gradient flow towards such KKT points. The reconstruction hence is a nonconvex optimization method to find the training data and coefficients that fulfill stationary and dual feasibility conditions and minimize the reconstruction loss. Further advancing this study, Buzaglo et al. showed that reconstructing training samples from multiclass neural networks is feasible, with even higher quality than binary classification reconstructions [10].

Our work builds upon these foundations, employing the reconstruction scheme from Haim et al.[9] to genetic data, thus navigating the novel terrain of high-dimensional data reconstruction with just a gray-box access to the model, meaning that we only have access to the model's weights and not the training procedure, such as the gradients during training. This reconstructionability is highly reliant on the performance of the prediction model, in our case trained on genetic data.

Cancer Detection using Genetic Data: In the area of cancer detection and classification using gene expression data, numerous innovative methods have been employed to address the challenges posed by high dimensionality and complexity of the data.

Danaee et al. employ a Stacked Denoising Autoencoder (SDAE) as well as differentially expressed genes (DEG) to extract meaningful features from gene expression data, highlighting the performance of subsequently trained cancer predictors [11]. Khan et al. introduce a transformer-based model that leverages the self-attention mechanism, offering a promising approach for classifying cancer subtypes without preliminary feature selection [14]. Rukhsar et al. show the application of Convolutional Neural Networks (CNNs) after converting RNA-Seq data into 2D images showcasing deep learning's effectiveness in classifying multiple cancer types [13]. Alharbi et al. provide a thorough overview of machine learning techniques, emphasizing the shift towards deep learning for its abil-

ity to discern distinctive gene patterns relevant to cancer classification [12]. At last, Khalsan et al. combine fuzzy logic with feature selection to enhance cancer classification, significantly improving accuracy by effectively reducing the dimensionality of gene expression data [15].

Our approach draws on these advancements, particularly the dimensionality reduction techniques from the [11], to enhance the accuracy and efficiency of classifying samples while enabling the reconstruction of genetic data.

3 Experiments

MedMNIST Data As the previous work from Haim et al. [9] shows the effectiveness of inversion attacks using a gray-box approach on images from the CIFAR-10 and MNIST dataset, we initially reproduced their pipeline for medical images using the MedMNIST dataset proposed by Yang et al. [3]. We hereby trained a classifier to differentiate between breast cancer (BreastMNIST) and organ images (OrganCMNIST) and aimed at reconstructing the training images.

Genetic Data Gene expression values indicate the frequency at which the gene is synthesized in that sample. It can be measured with RNA sequencing, which is a technique that involves isolating and reading the RNA inside cells. Gene expression values carry important information because changes in the levels of gene expression are associated with the development and progression of diseases, including cancer. For instance, the over or under-expression of certain genes can indicate the presence of tumor cells and their characteristics, making gene expression values valuable for cancer classification and diagnosis [11].

Building up on the work of Danaee et al. [11], our dataset originates from the Genomic Data Commons (GDC) Data Portal [16], specifically from the TCGA-BRCA (The Cancer Genome Atlas Breast Invasive Carcinoma) project. This dataset contains gene expression values for 60.660 different genes across 282 healthy and 1.211 cancer samples. The gene expression values in our dataset have been normalized across all samples to ensure comparability and are quantified using TPM (Transcripts Per Million) unstranded values. TPM is a normalization method for RNA-seq data that accounts for gene length and sequencing depth, allowing for the comparison of gene expression levels within and across samples. We utilize undersampling to address the issue of class imbalance between cancer and healthy patients. This approach aligns with findings from Haim et al. [9], suggesting that using a smaller dataset for training a neural network can lead to more effective inversion attacks. The resulting dataset is organized in a tabular format, where each row corresponds to one sample (patient), and each column represents the gene expression value for a specific gene. We trained neural networks to differentiate between cancer and healthy patients and aimed at reconstructing the training data.

Experimental Setup We aimed to demonstrate the reconstructablity of the medical images and genetic data from trained neural networks using the gray

box approach proposed by Haim et al. [9]. Most importantly, we do not assume any knowledge about the internal configuration of the model such as the gradients of the weights. As the medical images from MedMNIST have the same dimensionality as the CIFAR-10 and MNIST datasets, we were able to apply the pipeline proposed by Haim et al. directly.

Given the high-dimensional nature of the genetic data (60.660 features), we applied several dimensionality reduction techniques to extract meaningful information without losing significant data. For this we utilized the dimensionality reduction methods proposed by Danaee et al. [11]: SDAE, DEG, Principal Component Analysis (PCA), Kernel PCA (KPCA). These methods reduce the feature space from 60.660 to 784 features, subsequently using the most important features to reduce complexity for classification and reconstruction while preserving the essence of the original high-dimensional data.

After training a cancer predictor to differentiate between cancer and healthy patients using the transformed training data, the approach by Haim et al. is used to reconstruct the input data. The reconstructions are compared to the training data to quantify the effectiveness of the inversion attack.

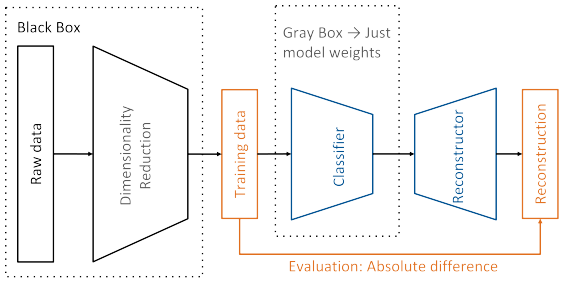


Fig. 1. Setup for Reconstructing Genetic Data from Trained Cancer Predictors

4 Results

4.1 MedMNIST

Using MedMNIST data of equivalent dimensionality to the CIFAR-10 and MNIST datasets used by Haim et al. [9], our goal was to gather insights into the applicability of the reconstruction process for medical data. For our experiments, we trained a classifier to differentiate between breast cancer (BreastMNIST) and organ images (OrganCMNIST) and performed an inversion attack on the trained model. When visually analyzing the reconstructions, it becomes apparent that the breast cancer data was reconstructed to resemble the mean of the dataset, whereas the reconstructions of the organs retain their semantic meaning. This observation hints at the significance of spatial correlation for the reconstruction

process. Additionally, our investigation of other MedMNIST classifiers revealed that datasets exhibiting high inter-class variance provided more successful reconstruction results.

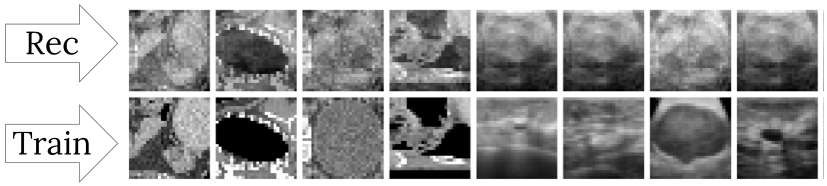


Fig. 2. Reconstructed images and their nearest neighbour out of the training data calculated based on L2 distance. Classifier trained to differentiate between breast cancer (BreastMNIST) and organ images (OrganCMNIST).

Furthermore, these experiments revealed that while the reconstruction process exhibits a noisy optimization under L2 distance evaluation, it showed a more stable optimization when evaluated using the structural dissimilarity metric DSSIM (see figure 3). This further indicates that reconstructability might be linked to spatial correlation within the training data.

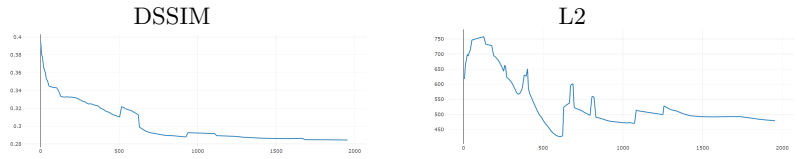


Fig. 3. Development of DSSIM and L2 distance between reconstructions and their closest training data over various reconstruction epochs.

4.2 Genetic Data

After successfully demonstrating the reconstructability of medical images from trained classifiers, we moved on to genetic data. We hereby trained a classifier to differentiate between cancer and healthy patients using gene expression data after dimensionality reduction as an input. We employed different patient selection strategies to investigate whether increasing the inter or intra-class variance (measured with the euclidean distance) increases the performance of the classifier. This is important as Haim et al. [9] showed that the lower the loss of the model the higher the reconstructability. Our result was that increasing the inter-class variance by selecting the most dissimilar pairs of patients between

cancer and healthy samples leads to the models with lower training and test loss. The best reconstruction was generated from a classifier trained on the most dissimilar pairs of patients where the features have been reduced using PCA. The train error of this model was 0.019 and the test error 0.408. The mean absolute difference between the reconstructions and their nearest neighbour over all reconstructions was 0.85 ± 0.00 over all genetic features. While this was the best reconstruction we achieved, the reconstruction quality is still poor, as the standardized training data takes on values between $[0, 1]$.

To investigate whether the classifiers performance is limited by the constrains set by the mathematical assumptions, we trained a simple ResNet18 architecture to differentiate between cancer and healthy patients (see table 1). Using this approach, the highest test accuracy was 0.9667 compared to 0.592. Hence, we showed that the performance of the model is limited by the constrains posed by Haim et al. on the gray-box inversion attack.

Reduction	DEG	DEG	DEG	DEG	PCA	PCA
Patient Selection	Random	Dissimilar	Random	Dissimilar	Random	Dissimilar
Test Accuracy	63.33	93.33	**55.00**	**96.67**	73.33	90.00

Table 1. ResNET18 test accuracy values across experiments with different setups. Selecting dissimilar patients resulted well compared to random selection with all combinations of parameters. The best result was achieved with using DEG.

4.3 Spatial Correlation Experiments

The results of the previous work and experiments with the MedMNIST datasets showed higher reconstruction quality for images compared to genetic data. One hypothesis for this notable difference is that spatial correlation within the training data, as observed in images, enhances the reconstructability. Spatial correlation refers to the degree to which each pixel's value is influenced by its neighboring pixels. Images often exhibit high spatial correlations, as patterns or clusters, often referred to as blobs, can be observed. Conversely, low spatial correlation is characterized by the independent distribution of pixel values, lacking observable patterns. This characteristic is often observed in tabular data, such as genetic information.

To test our hypothesis we investigated the relationship between the reconstruction error and spatial correlation for artificially generated images with varying degrees of spatial correlation. Therefore, we designed a series of experiments with two classes of images: One control group and one experimental group each consisting of grids with dimensionalities of 28x28 pixels. The control group served as the baseline with low spatial correlation with 100 randomly placed black pixels. The experimental group is designed to exhibit varying degrees of spatial

correlation, achieved through the formation of blobs. This class initially featured isolated black pixels to represent low spatial correlation. Progressively, we increase the adjacency of black pixels in subsequent experiments, forming larger blobs to symbolize higher spatial correlation. The total black pixel count remains consistent throughout all experiments to eliminate bias related to pixel density. This methodological decision ensured that any observed differences in reconstruction quality could be attributed to the degree of spatial correlation.

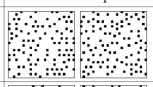
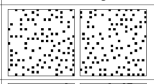
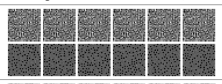
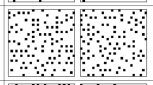
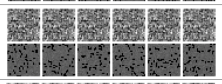
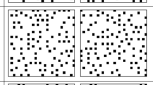
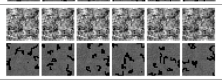
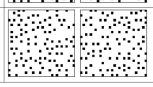
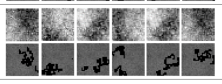
Experiments	Class 1 Control Group Samples	Class 2 Experimental Group Samples	Best Reconstructions (bottom is original data, top is reconstructed data)	Mean Error of Reconstruction Score (the lower the better)
Exp1				1263.55
Exp2				1280.26
Exp3				1105.53
Exp4				962.66

Fig. 4. This table presents the experimental setup and corresponding qualitative and quantitative reconstruction results where the blobs in the experimental group are generated via a random walk.

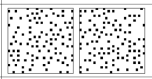
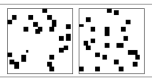
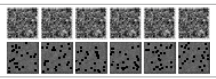
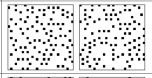
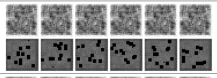
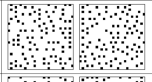
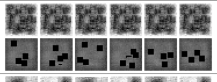
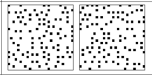
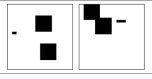
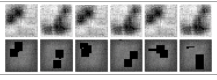
Experiments	Class 1 Control Group Samples	Class 2 Experimental Group Samples	Best Reconstructions (bottom is original data, top is reconstructed data)	Mean Error of Reconstruction Score (the lower the better)
Exp1				1273.13
Exp2				1233.77
Exp3				1160.03
Exp4				880.47

Fig. 5. This table presents the experimental setup and corresponding qualitative and quantitative reconstruction results where the blobs in the experimental group are generated using rectangular shapes.

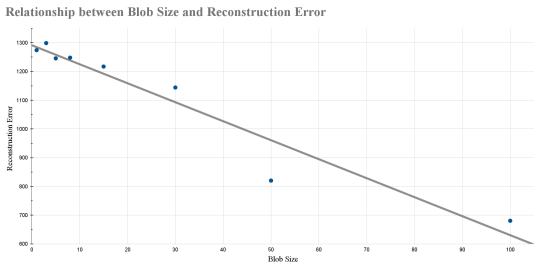


Fig. 6. Scatter plot demonstrating the inverse relationship between blob size and reconstruction error in experiments with rectangular, regular-shaped blobs as samples for Class 2. As the blob size increases, a consistent decrease in reconstruction error is observed.

We employed two methods for creating these blobs in the experimental group: A random walk method for irregularly shaped blobs (see figure 4) and rectangular, more regular-shaped clusters (see figure 5). This variation allowed us to explore how different patterns of spatial correlation affect reconstruction quality.

The results for the experiments are the same for both blob generation methods: With increasing spatial correlation, the reconstruction error decreases. This linear relationship is further visually represented in figure 6, where the reconstruction error is plotted against the blob size. These findings underscore the importance of spatial correlation in the context of reconstructing training data from trained neural networks using the proposed gray-box approach. Conversely, they underscore the challenge of reconstructing genetic data, which typically lacks structural dependency: The absence of spatial correlation in such datasets may prevent comparable reconstruction quality using the proposed gray-box approach.

5 Conclusion

In this paper, we investigated the applicability of an inversion attack of binary classifiers while only assuming a gray-box access to the model. When using medical images as training data, results comparable to the current state-of-the-art proposed by Haim et al. [9] are achieved. When training a cancer predictor on genetic data in a tabular format, the same reconstruction quality can not be reproduced. We showed that complying with the constraints set out for the gray-box reconstructability limits the performance of the trained cancer predictors and subsequently the inversion attack. Using a simple ResNet architecture improved the quality of the classifier, but no further gray-box attack on the training data is possible due to the violation of the mathematical constraints. Using experimental data we showed that reconstruction quality is enhanced with increasing spatial correlation, hence giving reason for the poor reconstruction quality of the tabular genetic data.

References

1. Ji Z., Telgarsky M.: Directional convergence and alignment in deep learning. In Advances in Neural Information Processing Systems 33 (2020): 17176-17186.
2. Lyu K., Li J.: Gradient descent maximizes the margin of homogeneous neural networks. In arXiv preprint arXiv:1906.05890 (2019).
3. Yang J., Shi R., Wei D., Liu Z., Zhao L., Ke B., Pfister H., Ni B.: MedMNIST v2-A large-scale lightweight benchmark for 2D and 3D biomedical image classification. In Scientific Data 10.1 (2023): 41.
4. Fredrikson, M., Jha, S., Ristenpart, T.: Model inversion attacks that exploit confidence information and basic countermeasures. In Proceedings of the 22nd ACM SIGSAC conference on computer and communications security, 1322-1333 (2015)
5. Shokri, R., Stronati, M., Song, C., Shmatikov, V. : Membership inference attacks against machine learning models. In 2017 IEEE symposium on security and privacy (SP), 3-18 (2017)
6. Radhakrishnan, A., Belkin, M., Uhler, C.: Overparameterized neural networks implement associative memory. Proceedings of the National Academy of Sciences **117**(44), 27162-27170 (2020)
7. Wang, Z., Lee, J., Lei, Q.: Reconstructing Training Data from Model Gradient, Provably. In International Conference on Artificial Intelligence and Statistics, 6595-6612 (2023)
8. Wang, Q., Kurz, D.: Reconstructing training data from diverse ml models by ensemble inversion. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2909-2917 (2022)
9. Haim, N., Vardi, G., Yehudai, G., Shamir, O., Irani, M. : Reconstructing training data from trained neural networks. Advances in Neural Information Processing Systems **35**, 22911-22924 (2022)
10. Buzaglo, G., Haim, N., Yehudai, G., Vardi, G., Irani, M.: Reconstructing Training Data from Multiclass Neural Networks. arXiv preprint arXiv:2305.03350. (2023)
11. Danaee, P., Ghaeini, R., Hendrix, D. A.: A deep learning approach for cancer detection and relevant gene identification. In Pacific symposium on biocomputing 2017, 219-229 (2017)
12. Alharbi, F., Vakanski, A. : Machine learning methods for cancer classification using gene expression data: A review. Bioengineering **10**(2), 173 (2023)
13. Rukhsar, L., Bangyal, W. H., Ali Khan, M. S., Ag Ibrahim, A. A., Nisar, K., Rawat, D. B. : Analyzing RNA-seq gene expression data using deep learning approaches for cancer classification. Applied Sciences **12**(4), 1322-1333 (2022)
14. Khan, A., Lee, B.: Gene transformer: Transformers for the gene expression-based classification of lung cancer subtypes. arXiv preprint arXiv:2108.11833. (2021)
15. Khalsan, M., Mu, M., Al-Shamery, E. S., Machado, L., Ajit, S., Agyeman, M. O.: Fuzzy Gene Selection and Cancer Classification Based on Deep Learning Model. arXiv preprint arXiv:2305.04883. (2023)
16. National Cancer Institute GDC Data Portal, `https://portal.gdc.cancer.gov/`.