Tab-Distillation: Impacts of Dataset Distillation on Tabular Data For Outlier Detection

Dayananda Herurkar
German Research Center for Artificial Intelligence (DFKI)
Germany
RPTU Kaiserslautern-Landau
Germany
dayananda.herurkar@dfki.de

Federico Raue
German Research Center for Artificial Intelligence (DFKI)
Germany
federico.raue@dfki.de

Andreas Dengel
German Research Center for Artificial Intelligence (DFKI)
Germany
RPTU Kaiserslautern-Landau
Germany
andreas.dengel@dfki.de

Abstract

Dataset distillation aims to replace large training sets with significantly smaller synthetic sets while preserving essential information. This method reduces the training costs of advanced deep learning models and is widely used in the image domain. Among various distillation methods, "Dataset Condensation with Distribution Matching (DM)" stands out for its low synthesis cost and minimal hyperparameter tuning. Due to its computationally economical nature, DM is applicable to realistic scenarios, such as industries with large tabular datasets. However, its use in tabular data has not been extensively explored. In this study, we apply DM to tabular datasets for outlier detection. Our findings show that distillation effectively addresses class imbalance, a common issue in these datasets. The synthetic datasets offer better sample representation and class separation between inliers and outliers. They also maintain high feature correlation making them resilient against feature pruning. Classification models trained on these distilled datasets perform faster and better that will enhance outlier detection in industries that rely on tabular data.

CCS Concepts

• **Computing methodologies** → **Anomaly detection**; **Neural networks**.

Keywords

outlier detection, tabular data, neural networks, imbalanced dataset, feature correlation, dataset distillation

ACM Reference Format:
Dayananda Herurkar, Federico Raue, and Andreas Dengel. 2024. Tab-Distillation: Impacts of Dataset Distillation on Tabular Data For Outlier Detection. In *5th ACM International Conference on AI in Finance (ICAIF '24), November 14–16, 2024, Brooklyn, NY, USA*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3677052.3698660

ICAIF '24, November 14–16, 2024, Brooklyn, NY, USA

ACM ISBN 979-8-4007-1081-0/24/11
https://doi.org/10.1145/3677052.3698660

1 Introduction

Tabular data, characterized by its structured rows and columns, often encompasses mixed-type features, including numerical and categorical data. Despite the unique challenges posed by its heterogeneous nature, tabular data is critically important across various industries, supporting applications in finance, healthcare, manufacturing, and retail. The prevalence of tabular data in finance has made it a particularly promising area for the application of advanced machine learning methods. The recent escalation in fraudulent activities has heightened the focus on detecting financial fraud, making outlier detection a key area of research [15]. Outlier detection, crucial for identifying rare but significant outliers, has been an active research domain for decades, especially within the financial sector [2]. However, a significant challenge in this area is the common occurrence of class imbalances, where outliers are far less frequent than normal instances. This imbalance results in a dominance of majority classes, leading to models becoming biased towards these classes and reducing their ability to accurately detect classes [8]. Addressing class imbalance issues is essential for enhancing the performance and reliability of machine learning models in identifying outliers within tabular data.

To develop AI-based solutions for various industrial tasks, such as fraud detection, model training is essential. However, training these models can be expensive and resource-intensive, creating a strong demand for techniques that reduce the computational cost of training multiple models on the same dataset with minimal performance degradation. Traditionally, coreset selection has been employed to reduce training set sizes by picking samples deemed crucial for training using heuristic criteria. Examples include minimizing the distance between the centers of the coreset and the entire dataset [6], tracking the frequency of misclassifications [29], and enhancing the diversity of the chosen samples [3]. However, its efficiency is constrained by the information contained in the selected samples from the original dataset. Dataset Distillation has emerged as a superior alternative, overcoming the limitations of coreset approaches. Various methods for dataset distillation exist, such as those described in [30], [35], and [7]. Notably, DM [34] integrates the advantages of other distillation methods while circumventing their limitations. By avoiding costly bi-level optimization and extensive hyper-parameter tuning, DM is significantly faster and more applicable to realistic scenarios. Despite these advancements, distillation techniques have predominantly been explored within the image domain and have not been thoroughly investigated for tabular data.

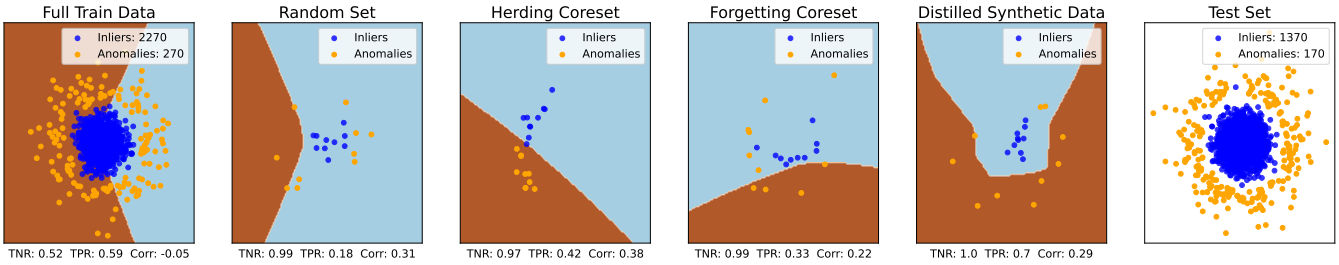

Figure 1: Visualization of 2D Toy Scenario. This figure illustrates four datasets generated using different methods: random selection, coreset selection, and distillation (as described in Section 4.2, Section 3). Separate Multilayer Perceptron (MLP) models, all with identical hyperparameters (two hidden layers with Relu activations), were trained from scratch on each dataset and evaluated on the test set of the Toy dataset. Each subplot displays the training samples and the decision boundary formed by the corresponding model. The x-axis of each subplot indicates the models' performance on the test set. Notably, the model trained on the distilled synthetic dataset exhibits superior outlier detection capabilities and achieves better separation between inlier and outlier classes compared to the others.

In this paper, we apply dataset condensation using distribution matching (DM) to tabular datasets used for outlier detection, aiming to bridge the gap in research and demonstrate its applicability beyond the image domain. By focusing on tabular data, we explore the effectiveness of DM in addressing the unique challenges posed by mixed-type features and class imbalances prevalent in outlier detection tasks. Our study shows three properties incorporated by the distillation process on the tabular data. Firstly, the distillation of highly imbalanced real outlier detection datasets brings class balance into synthetic datasets. Our experimental results indicate that class balance improves performance in distilled synthetic datasets compared to other approaches. Additionally, the synthetic data captures critical information from the real data and represents inliers and outliers in significantly fewer samples compared to real data. As a result, synthetic outliers are better represented and lead to better class separation between inliers and outliers. We demonstrate this property of synthetic datasets by conducting experiments on both toy and public datasets. Furthermore, generally independent features in the real dataset become highly correlated in the synthetic datasets. Due to highly correlated features these synthetic datasets become resilient against feature pruning. Hence, model performance does not degrade even after removing feature information from synthetic datasets that are lacking in real datasets. Due to such positive impacts of applying distillation, synthetic datasets with less than 5% of the samples of the real dataset provide better performance than real datasets and are better suited for outlier detection tasks.

In summary, we present the following contributions:

- We uncover the hidden properties of DM for tabular data, particularly in the context of outlier detection.
- By applying DM, we demonstrate that synthetic datasets can outperform full datasets in certain scenarios, owing to better outlier representation and class balancing.
- DM exhibits resilience against feature pruning, showcasing the robustness of distilled datasets even when features are reduced.

The remainder of this work is structured as follows: Section 2 provides a review of relevant literature, highlighting gaps in the domain. In Section 3, we detail our approach to tabular dataset distillation. The experimental setup, model architecture, datasets used, and evaluation measures are described in Section 4, while Section 5 presents the results and comparisons. We conclude in Section 6, summarizing our main findings and identifying opportunities for future research.

2 Related Work

Dataset Distillation in Images: Distilled Dataset is the task of compressing a dataset into a smaller size version with the condition that the performance of a model trained on both versions is as minimal as possible. This field is inspired by knowledge distillation, in which a teacher model transfers its knowledge to a student model that is usually a smaller version of the teacher. With this in mind, Wang *et al.*[30] proposed a bi-level optimization approach that transfers the real image dataset into a smaller version (i.e., less than 50 images per class). Several approaches have been developed based on matching gradients [35], distributions [34], and training trajectories [7]. Dataset Distillation is applied not only to images but also to multi-modal data [32], text [19], graph [17]. Medveded *et al.* presented the only work exploring distilled datasets on tabular data [20]. They used an artificial two-dimensional binary classification dataset to improve the generalization problem between different architectures. In this work, we are interested in a wider scenario that evaluates Distilled Datasets using outlier detection in tabular datasets. Distilled Dataset approaches showed unexplored attributes which are easier to analyze in these tabular datasets.

Outlier Detection on Tabular Data: Outlier detection has gathered extensive attention in research across diverse fields for several decades, with a heightened emphasis in the financial sector [22]. The use of tabular data in financial applications has recently flourished, offering promising opportunities for new methodologies [5]. There are reconstruction-based approaches to detect outliers that assume outliers are difficult to reconstruct from low-dimensional projections. DAGMM [36] is one such approach that integrates density estimation with both reduced representation and reconstruction error, aiming for a more holistic outlier detection approach that considers low-density regions in reduced spaces. Also, there

Table 1: Description of benchmark tabular datasets used for our experiments.

Dataset	Feature Type	Samples	Columns			Outliers (%)
			Categ.	Num.	Encoded	
Credit Default [33]	Mixed	30000	10	13	146	22.10
Credit Fraud [11]	Num	284807	-	29	29	0.17
Census Income [1]	Mixed	299285	33	8	511	6.21
Adult Data [4]	Mixed	48842	8	6	118	23.90
Bank Marketing [21]	Mixed	41188	10	10	63	11.20
IEEE Fraud [16]	Mixed	590540	31	400	3172	3.50

are clustering methods like Gaussian Mixture Models, and K-means are widely used for outlier detection but face challenges with high-dimensional data due to the curse of dimensionality. In addition, one-class classification methods are frequently employed in outlier detection. These algorithms, such as one-class SVM [25], learn a discriminative boundary around normal instances within the dataset. Furthermore, recent methodologies extend to tasks such as including outlier detection in accounting data [12],[27], interpretation of outliers within financial tabular datasets [26], transforming outliers across different tabular data formats [14], federated outlier detection on financial tabular data [13], modeling behavioral fraud patterns [31], and enhancing anti-money laundering efforts [18],[23].

3 Methodology

3.1 Preliminaries

In this work, we evaluate Distilled Datasets for Outlier Detection in Tabular Datasets. The goal is to find a synthetic dataset with the condition that the difference between the model performances trained on real and distilled datasets is similar. One solution is to express this problem as a bi-level optimization which is defined with the following equation:

$$\begin{aligned} \mathcal{S}^* &= arg\ min_{\mathcal{S}}\ \ \mathcal{L}^{\mathcal{T}}(\theta^{\mathcal{S}}(\mathcal{S})) \\ \text{subject to }\ \theta^{\mathcal{S}}(\mathcal{S}) &= arg\ min_{\theta}\ \ \mathcal{L}^{\mathcal{S}}(\theta), \end{aligned} \tag{1}$$

where $\mathcal{L}^{\mathcal{T}}$ is the training loss on the real data $\mathcal{T}$, $\theta^{\mathcal{S}}$ is a neural network expressed as a function of the synthetic data $\mathcal{S}$. One solution is the dataset condensation using distribution matching (DM) [34] approach that optimized the distance between the mean distribution of the real logits and synthetic logits.

$$\mathbb{E}_{v\sim P_v}\|\frac{1}{|\mathcal{T}|}\sum_{i=1}^{|\mathcal{T}|}\psi_v(\mathbf{x}_i) - \frac{1}{|\mathcal{S}|}\sum_{j=1}^{|\mathcal{S}|}\psi_v(\mathbf{s}_j)\|^2, \tag{2}$$

where $\psi_v(\mathbf{s}_j)$ is a neural network with parameters v and P_v is the distribution of network parameters.

3.2 Properties of DM in Tabular Datasets

DM was originally evaluated for image classification datasets that are balanced (i.e., each class has the same number of images) and contain high-dimensional input space. In contrast, this work shows empirical results regarding several properties of DM for Outlier Detection which is defined in the following equation:

$$\begin{aligned} \mathcal{T} &= \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{n} \\ where\ \mathbf{x}_i &= (x_i^1, \ldots, x_i^D) \\ x_i^d &= \begin{cases} x_i^{d_{\text{num}}} \in \mathbb{R} & \text{for numerical features} \\ x_i^{d_{\text{cat}}} \in \{1, ..., C\} & \text{for categorical features} \end{cases} \\ \mathbf{y}_i &\in \{inlier, outlier\} \\ \|inlier\| &>> \|outlier\|. \end{aligned} \tag{3}$$

Note that outlier detection has a mixed set of features (x_i^d) between continuous and discrete and the number of inliers is much greater than the number of outliers.

Property 1 - Class Balanced: DM inherits the option to learn a condensed version of the actual dataset, with a specific number of samples per class (e.g., 10, 50, or 100 samples). As a result, the imbalanced outlier detection problem is transformed into a more manageable balanced outlier detection task, with all features being continuous. The following equation shows the distilled dataset version of the outlier detection problem:

$$\begin{aligned} \mathcal{S}^* &= \{(\hat{\mathbf{x}}_j, \mathbf{y}_j)\}_{j=1}^{m} \\ where\ \hat{\mathbf{x}}_j &\in \mathbb{R}^D \\ n &>> m \\ \|inlier\| &= \|outlier\|. \end{aligned} \tag{4}$$

Property 2 - Outlier Representation: Another property is related to the outlier representation, which is a better representation than the real data samples. Fig 1 shows several empirical results that compare trained models in different conditions. We want to point out that the Distilled Synthetic Data reaches better results than the three coreset approaches (balanced dataset) and Full Datasets (imbalanced dataset). Additionally, the decision boundary of the distilled dataset looks quite similar to the Test Set, whereas, the decision boundaries of the other examples do not cover the outlier area.

Property 3 - Feature Correlation: The last property shows a correlation between features. It can be observed in Fig 1 that all small datasets (i.e., random, herding, forgetting, and distilled) reach a higher correlation than the full train data. We can infer that fewer samples show a higher feature correlation. The advantage of the distilled dataset is that it contains both higher feature correlation and

better representation. This property can be exploited for pruning the feature space.

4 Experimental Setup

This section includes descriptions of the datasets and data preprocessing procedures, alongside the baseline methods used for comparison, and different evaluation metrics.

4.1 Datasets

We evaluated the DM technique using six standard financial tabular datasets for outlier detection. Five out of six datasets contain mixed-type features and one dataset contains only numerical features. During data preprocessing, all categorical attributes were encoded using the one-hot encoding method, and numerical attributes were standardized to have a mean of 0 and a standard deviation of 1. The one-hot encoded categorical attributes were then combined with the standardized numerical attributes. Hence, the total number of encoded attributes is the concatenation of categorical and numerical attributes. Table 1 summarized the attribute features of several datasets.

4.2 Baseline Methods

To benchmark the performance of the Tab-Distillation method, we conducted a comparative analysis against four baselines.

- **Full Dataset:** This involves utilizing the entire dataset for training and evaluation purposes. It serves as a reference point to assess the performance of other methods by providing a comprehensive view of the data, ensuring that no information is omitted. This approach is particularly useful for establishing an upper bound on performance, as it leverages all available data without any sampling or selection bias.
- **Random Selection:** In this baseline method, a random subset of the dataset is selected for training. This technique involves reducing the size of the dataset by randomly selecting a subset of instances from the original dataset. The primary goal is to create a smaller, more manageable dataset for training, which can help mitigate the computational burden while still capturing the essential characteristics of the data. During the selection, class balance is maintained to make a fair comparison with the model performance on the distilled dataset.
- **Herding [6], [10], [24]:** This method iteratively selects data points that minimize the maximum discrepancy between the empirical distribution of the subset and the target distribution, ensuring that the selected subset preserves the essential statistical properties of the full data.
- **Forgetting [29]:** The forgetting coreset method tracks how often each training sample is learned and subsequently forgotten during network training. Samples that are less frequently forgotten are deemed less informative and can be excluded from the coreset.

We have also compared the Tab-Distillation to two standard resampling methods listed below

- **SMOTE [9]:** Synthetic Minority Over-sampling Technique is a data augmentation method used to address class imbalance in datasets. It generates synthetic examples of the minority class (outliers) by interpolating between existing samples, effectively increasing its representation in the dataset. This technique helps prevent overfitting and improves classifier performance by promoting a more balanced decision boundary.
- **TomekLinks [28]:** It is an undersampling technique aimed at cleaning class boundaries by identifying and removing ambiguous instances. A Tomek Link exists between two samples of opposite classes if they are each other's nearest neighbors; removing the majority class (inliers) instance in such pairs reduces class overlap, improving class separability. This method enhances classifier performance by refining the decision boundary and mitigating noise.

4.3 Hyperparameters

In this study, we employed a Multilayer Perceptron (MLP) to learn synthetic sets across all datasets. For each original dataset, optimal hyperparameters were determined through an exhaustive search and then a specific architecture was selected, characterized by a unique combination of the number of neurons and layers. The exhaustive search includes the following ranges: number of neurons [4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048], number of hidden layers [2, 3, 5], activation function [Sigmoid, Relu, Leaky_Relu], batch size [32, 64, 128, 512] and optimizers [SGD, Adam]. The model architecture for each dataset is as follows: Credit Default (64,32), Credit Fraud (32, 16), Census Income (256,128,64), Adult Data (64,32), Bank Marketing (32,16), and IEEE Fraud (2048,1024,512,256,128). Sigmoid activation functions were utilized, and model parameters were optimized using the Stochastic Gradient Descent (SGD) optimizer with a momentum factor set to 0.9. A fixed learning rate of 1.0 was used for optimizing synthetic samples for learning scenarios with 10, 50, and 100 samples per class (SPC) across all datasets. The synthetic samples were trained for the following number of iterations: 7000 for Credit Default, 8000 for Credit Fraud, 10000 for Census Income, 6000 for Adult Data, 14000 for Bank Marketing, and 20000 for IEEE Fraud datasets. Synthetic samples were initialized using random real samples with corresponding labels. For the evaluation of synthetic data, we used the same MLP hyperparameters as those used during distillation, except for a learning rate of 0.01 and a training duration of 300 epochs.

4.4 Evaluation Metrics

To evaluate the quality of the DM technique, we employed four distinct metrics to measure the detection rate. Given those datasets for outlier detection often exhibit highly imbalanced class ratios, dominated by inliers, we used Mean Accuracy across classes. This metric provides a balanced measure of performance across both the classes, ensuring it is not overly biased towards the dominant class in particular outlier detection cases. Additionally, to quantitatively assess the outlier detection performance of the models, we utilized the 'F1-Score', the area under the precision-recall curve ('PR-AUC'), 'TNR' and 'TPR' all of which are standard metrics in the domain of outlier detection.

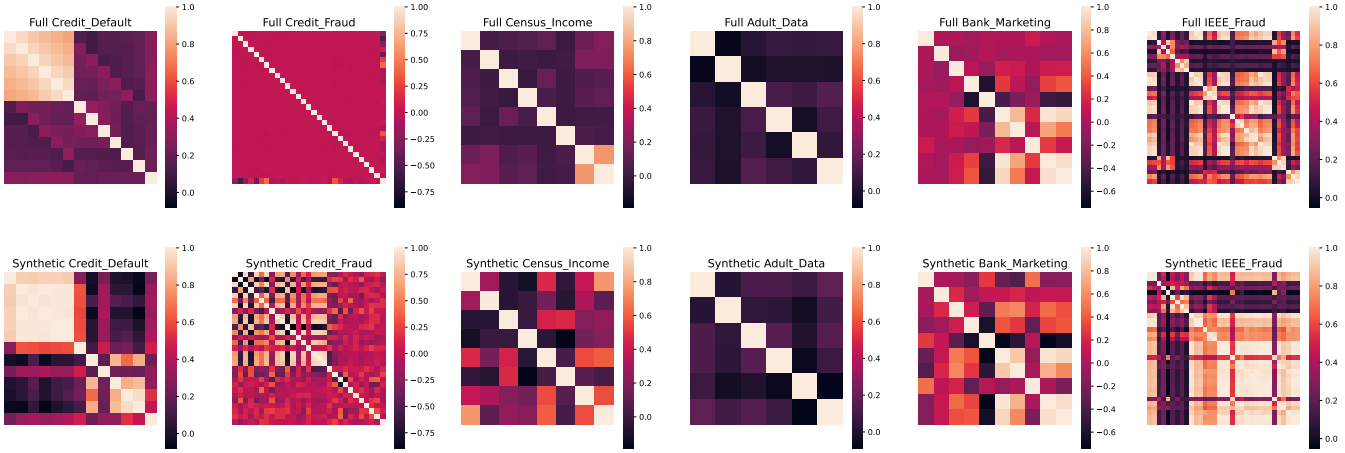

Figure 2: Comparison of Feature Correlation Matrix between Synthetic and Full Datasets: Each subplot represents the correlation matrix of features in the dataset. The matrix of the distilled synthetic dataset contains higher feature correlation values compared to full datasets across all six benchmark tabular datasets for outlier detection.

Table 2: Comparison of model performance trained on random set (R), herding coreset (H), forgetting coreset (F), Distilled Synthetic Set (DM), and full dataset (Full): DM outperforms other baselines such as R, H, F and performs as good as Full across six datasets and different metrics. In a few cases when the SPC is 100 (marked by underline), DM outperforms the Full.

Datasets	SPC	Mean Accuracy					F1-Score					PR-AUC				
		R	H	F	DM	Full	R	H	F	DM	Full	R	H	F	DM	Full
Credit Default	10	0.535	0.498	0.506	**0.634**		0.324	0.355	0.349	**0.423**		0.274	0.238	0.276	**0.425**	
	50	0.524	0.525	0.540	**0.621**	0.664	0.367	0.349	0.386	**0.411**	0.453	0.288	0.283	0.309	**0.459**	0.455
	100	0.539	0.528	0.547	**0.671**		0.381	0.351	0.367	**0.466**		0.365	0.329	0.283	**0.504**	
Credit Fraud	10	0.929	0.296	0.338	**0.944**		0.277	0.003	0.002	**0.464**		0.728	0.040	0.006	**0.781**	
	50	0.913	0.913	0.912	**0.938**	0.865	0.217	0.354	0.254	**0.383**	0.313	0.747	0.728	0.206	**0.754**	0.799
	100	0.888	0.913	0.917	**0.939**		0.308	0.410	0.119	**0.415**		0.745	0.699	0.255	**0.779**	
Census Income	10	0.680	0.515	0.517	**0.735**		0.207	0.057	0.127	**0.330**		0.156	0.125	0.079	**0.337**	
	50	0.586	0.484	0.701	**0.759**	0.761	0.104	0.088	0.245	**0.242**	0.385	0.164	0.074	0.292	**0.342**	0.416
	100	0.604	0.516	0.574	**0.791**		0.127	0.106	0.144	**0.398**		0.260	0.123	0.093	**0.402**	
Adult Data	10	0.686	0.512	0.469	**0.780**		0.051	0.040	0.295	**0.614**		0.478	0.496	0.257	**0.662**	
	50	0.665	0.544	0.441	**0.800**	0.790	0.151	0.304	0.583	**0.629**	0.637	0.460	0.568	0.263	**0.658**	0.701
	100	0.667	0.504	0.473	**0.794**		0.158	0.235	0.225	**0.631**		0.476	0.544	0.265	**0.693**	
Bank Marketing	10	0.699	0.663	0.575	**0.802**		0.356	0.281	0.236	**0.470**		0.395	0.466	0.347	**0.536**	
	50	0.722	0.485	0.775	**0.813**	0.810	0.334	0.176	0.256	**0.494**	0.499	0.488	0.299	0.526	**0.586**	0.618
	100	0.673	0.597	0.734	**0.853**		0.363	0.257	0.290	**0.535**		0.371	0.337	0.377	**0.621**	
IEEE Fraud	10	0.479	0.479	0.622	**0.632**		0.064	0.246	0.161	**0.260**		0.116	0.248	0.087	**0.222**	
	50	0.499	0.499	0.651	**0.675**	0.501	0.067	0.067	0.181	**0.201**	0.1	0.040	0.242	0.088	**0.249**	0.408
	100	0.506	0.496	0.499	**0.657**		0.068	0.067	0.067	**0.209**		0.173	0.247	0.121	**0.217**	

$$\text{TNR} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}} \tag{5}$$

$$\text{TPR} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{6}$$

5 Experiments and Results

This section presents a comprehensive overview of the experiments conducted. It details the results of each experiment, providing insights into the effectiveness of our approach through different evaluation measures.

5.1 Performance Comparison

In this scenario, we first distill several synthetic sets with 10, 50, and 100 samples per class for all datasets using the MLP model

Table 3: Comparison of model performance trained on oversampling method SMOTE, undersampling method TomekLinks, Distilled Synthetic Set (DM) with SPC=100, and full dataset (Full): DM outperforms other resampling methods on Mean Accuracy metric.

Datasets	Methods	Inliers	Outliers	Mean Accuracy	F1-Score	PR-AUC
Credit Default	Full	16355	4645	0.664	0.453	0.455
	SMOTE	16355	16355	0.620	0.408	0.379
	TomekLinks	15131	4645	0.634	0.431	0.418
	DM	100	100	**0.671**	**0.466**	**0.504**
Credit Fraud	Full	199020	344	0.865	0.313	0.799
	SMOTE	199020	199020	0.871	**0.519**	0.752
	TomekLinks	199002	344	0.888	0.495	**0.803**
	DM	100	100	**0.939**	0.415	0.779
Census Income	Full	196501	12998	0.761	0.385	0.416
	SMOTE	196501	196501	0.727	0.498	0.503
	TomekLinks	194139	12998	0.742	**0.519**	**0.526**
	DM	100	100	**0.791**	0.398	0.402
Adult Data	Full	26008	8181	0.790	0.637	0.701
	SMOTE	26008	26008	0.788	**0.647**	0.699
	TomekLinks	24460	8181	0.762	0.641	**0.722**
	DM	100	100	**0.794**	0.631	0.693
Bank Marketing	Full	25583	3248	0.810	0.499	0.618
	SMOTE	25583	25583	0.813	0.533	0.547
	TomekLinks	24852	3248	0.767	0.530	0.598
	DM	100	100	**0.853**	**0.535**	**0.621**
IEEE Fraud	Full	398914	14464	0.501	0.1	0.408
	SMOTE	398914	398914	0.524	0.198	0.415
	TomekLinks	397151	14464	0.528	**0.237**	**0.426**
	DM	100	100	**0.657**	0.209	0.217

described in Section 4.3. These synthetic sets were then utilized to train randomly initialized MLP models from scratch, which were subsequently evaluated on real test data. This experiment was repeated five times, and the average performance of the models across all five runs was reported. Our method was compared against the full dataset and three standard coreset selection methods as detailed in Section 4.2. For these comparisons, the MLP models were trained using the training sets of the baseline methods and tested on real test data.

The results of this experiment are presented in Table 2. Models trained on the synthetic datasets demonstrated a clear superiority over those trained with coreset selection methods. Across all datasets and three different evaluation metrics, the models trained with synthetic datasets consistently outperformed the baselines in outlier classification. This superior performance can be attributed to the fact that synthetic training data have balanced class ratios and are not confined to real sample sets, whereas baseline methods use subsets of real samples. Notably, models trained with synthetic datasets comprising 100 samples per class often outperformed those trained on the full training set. Models trained on the full datasets suffer from class imbalance, where dominant classes overshadow others. This issue is mitigated in models trained on synthetic datasets. We also compared the performance of the model trained on DM with SPC=100 against standard resampling methods like SMOTE [9] and TomekLinks [28]. Models trained using DM outperformed SMOTE and TomekLinks on mean accuracy metric across all six datasets.

5.2 Improved Class Separation Between Inliers and Outliers

To analyze the performance enhancement of models trained on synthetic data compared to those trained on the full dataset (discussed in Section 5.1), we investigated the representations of synthetic inlier and outlier samples separately. Initially, we calculated the density of the inlier and outlier samples separately. The density of a cluster was defined as the average distance of each point in the cluster from its center. This metric provides insight into the spread and compactness of the samples within each class. Subsequently, we computed the True Positive Rate (TPR) of the models on a real test set. Here, TPR represents the proportion of correctly predicted outliers (positive class). We conducted a similar analysis on the full dataset for comparison.

The results of this experiment are presented in Table 4. Our findings indicate that outlier samples in the synthetic dataset are more sparsely distributed compared to those in the full dataset. During the distillation process, the synthetic dataset effectively captures

Table 4: Lower the density measure output, more denser the samples. Outliers in distilled synthetic datasets are sparser than outliers in the full dataset. Also better TPR in the distilled dataset than full set. The distilled set with fewer samples represents outliers better than the Full dataset and also better decision boundary between inliers and outliers resulted in better TPR. Showing results for SPC=100.

Dataset		Density in Outliers	TPR
Credit Default	Full	3.570	0.684
	Herding	3.250	0.698
	Forgetting	2.917	0.673
	DM	3.752	**0.757**
Credit Fraud	Full	16.509	0.754
	Herding	16.700	0.832
	Forgetting	16.414	0.845
	DM	16.965	**0.867**
Census Income	Full	5.564	0.743
	Herding	4.385	0.231
	Forgetting	5.168	0.614
	DM	6.043	**0.893**
Adult Data	Full	3.098	0.757
	Herding	2.291	0.323
	Forgetting	2.597	0.493
	DM	3.961	**0.822**
Bank Marketing	Full	4.771	0.794
	Herding	3.206	0.713
	Forgetting	3.860	0.535
	DM	4.863	**0.838**
IEEE Fraud	Full	2.172	0.011
	Herding	0.769	0.159
	Forgetting	2.345	0.246
	DM	2.885	**0.349**

the critical information from the outlier samples, representing a broader range of the outlier space. This results in a better representation of outliers and a clearer separation between inliers and outliers in the synthetic dataset compared to the full dataset. We can see the same outcome from the toy dataset visualization shown in Figure 1. Consequently, models trained on the synthetic dataset exhibit a higher TPR than those trained on the full dataset shown in Table 4, demonstrating superior performance in identifying outliers. This improved performance underscores the efficacy of using synthetic data for training models in scenarios where distinguishing between inliers and outliers is crucial.

5.3 Pruning

In this experiment, we evaluate the resiliency of the synthetic data for feature pruning. Resiliency here refers to the ability of the synthetic data to maintain performance despite the pruning of part of its feature information. To assess this property, we begin by selectively removing information from specific features in the synthetic dataset. This involves randomly selecting columns and replacing their original values with zeros. The modified dataset is then used to train an MLP model from scratch, which is subsequently tested on the real test set (without pruning). We conduct this process for varying percentages of column removal—0%, 10%, 25%, 50%, and 75%—to observe how performance degrades as more information is removed. At each interval, the model's performance is evaluated using the F1-Score, the harmonic mean of precision and recall, and is particularly suitable for outlier detection tasks. To provide a comprehensive comparison, we also check the resiliency of the original full dataset by applying the same procedure.

Figure 3 illustrates the results, with the x-axis representing the percentage of pruned columns and the y-axis showing the F1-Score. Across all six datasets, synthetic datasets exhibit greater resiliency against column-wise information removal compared to full datasets. This holds true for different samples per class (10/50/100), where the performance drop from 0% to 75% column pruning is minimal for synthetic datasets compared to full datasets. In some cases, such as 'Credit Default', 'IEEE Fraud', and 'Census Income', removing just 10-25% of columns from the full dataset results in the model's F1-Score dropping to zero. The high correlation between features in synthetic datasets, a result of the distillation process in tabular datasets, is a key factor contributing to this behavior. To verify that we also computed the correlation between features of synthetic datasets and then compared it against feature correlation of full datasets. We can visualize the comparison of the feature correlation matrix in Figure 2. By comparing the feature correlation matrix of synthetic and full sets, the features in the synthetic datasets are more correlated than in the full datasets for all six datasets. These results complement the outcome in Figure 3. This property suggests a potential for privacy preservation in synthetic datasets while not losing their utility.

5.4 Cross-Model Generalization

One potential practical application of data distillation is the accelerated training of multiple models with varied initializations to an acceptable level of performance. Consequently, the synthetic data must generalize well across different algorithms. To investigate this aspect, we conducted experiments by training models of various algorithms on each distilled dataset. In this study, we implement a more rigorous cross-model experiment, utilizing 100 samples per class across all six datasets. As shown in Table 5, the synthetic data are initially learned with an MLP and subsequently evaluated on different models by training them from scratch and testing them on real test data. We evaluate the performance using several standard algorithms, including Random Forest (RF), Decision Tree (DT), Logistic Regression (LR), Gradient Boosting (GB), and Naive Bayes (NB).

Table 5 demonstrates that the highest performance is achieved when the synthetic dataset is both learned and evaluated using the same model i.e., MLP. In three out of six datasets, the MLP attains the best PR-AUC score for outlier detection. However, in the remaining three datasets, different algorithms outperform the MLP in detecting outliers. Notably, in the 'Census Income' dataset, the top three algorithms are not MLPs. These findings suggest

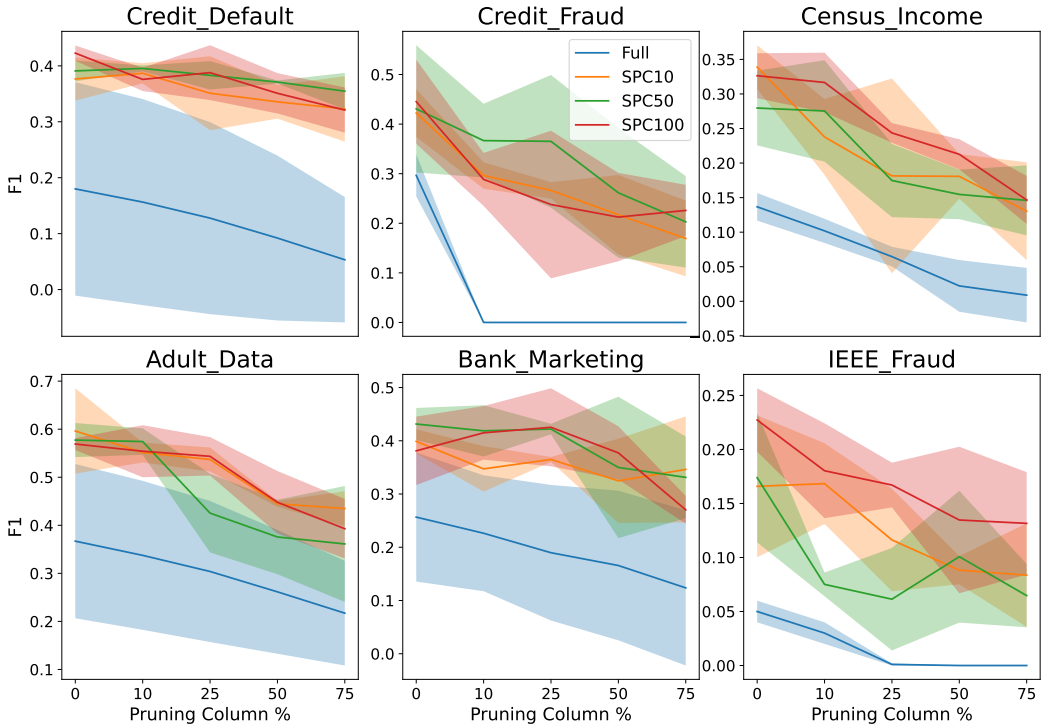

Figure 3: Robustness of Synthetic vs. Full Datasets. This figure illustrates the robustness of synthetic datasets under varying degrees of column-wise information removal, measured by the F1-Score. The x-axis shows the percentage of columns removed (0%, 10%, 25%, 50%, 75%), and the y-axis represents the F1-Score. Synthetic datasets maintain higher F1-Scores compared to full datasets, which exhibit significant performance drops with even minimal column removal. This improved robustness is due to higher feature correlation in synthetic datasets, as shown in Figure 2, indicating their effectiveness in preserving performance despite data degradation.

that synthetic samples generated through distillation exhibit good generalization performance across various models.

Table 5: Cross-Model Generalization of Synthetic Set (SPC=100): Outlier Detection Results (PR-AUC) using Other Supervised Models. RF: Random Forest, DT: Decision Tree, LR: Logistic Regression, GB: Gradient Boosting, NB: Naive Bayes. MLP the model that is been used for learning and evaluating the synthetic dataset archives the best results in three out of six datasets.

Dataset	MLP	RF	DT	LR	GB	NB
Credit Default	0.504	0.322	0.358	**0.523**	0.469	0.592
Credit Fraud	**0.779**	0.691	0.464	0.429	0.269	0.422
Census Income	0.402	0.132	**0.531**	0.462	0.143	0.500
Adult Data	0.693	0.388	0.619	**0.712**	0.279	0.599
Bank Marketing	**0.621**	0.229	0.140	0.560	0.140	0.543
IEEE Fraud	**0.217**	0.061	0.212	0.202	0.063	0.141

6 Conclusion

This study demonstrates that dataset condensation using distribution matching (DM) applied to tabular data significantly enhances outlier detection by addressing class imbalance, improving class separation, pruning resiliency, and cross-model generalization. Synthetic datasets generated through DM consistently outperformed traditional coreset selection methods and full datasets by offering balanced class ratios, better representation, and clearer separation of inliers and outliers. Furthermore, synthetic datasets exhibited remarkable resiliency to feature-wise information removal, maintaining high-performance levels even with significant data reduction. Additionally, these synthetic datasets generalized well across various machine learning models, demonstrating versatility and practical applicability for different algorithms. The study's findings underscore the effectiveness of dataset distillation in creating efficient and reliable models for outlier detection in industries reliant on tabular data, thereby extending the benefits of this technique beyond the image domain and paving the way for broader adoption in various industrial applications.

7 Future Work

Future research will explore other distillation methods for tabular data to further enhance model performance and efficiency. Investigating alternative approaches could uncover new techniques that offer even greater benefits for outlier detection. Additionally, leveraging the feature pruning capability of synthetic datasets presents a promising avenue for privacy preservation. By ensuring that models perform well even with reduced data fidelity, we can develop methods to protect sensitive information without compromising utility. Future studies will also focus on applying these findings to a broader range of industrial applications, ensuring the scalability and versatility of dataset distillation techniques in real-world settings.

Acknowledgments

This work was supported by the BMBF project Albatross (Grant 01IW24002).

References

[1] 2000. Census-Income (KDD). UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5N30T.

[2] Mohiuddin Ahmed, Abdun Naser Mahmood, and Md. Rafiqul Islam. 2016. A Survey of Anomaly Detection Techniques in Financial Domain. *Future Gener. Comput. Syst.* 55, C (feb 2016), 278–288. https://doi.org/10.1016/j.future.2015.01.001

[3] Rahaf Aljundi, Min Lin, Baptiste Goujaud, and Yoshua Bengio. 2019. *Gradient based sample selection for online continual learning.* Curran Associates Inc., Red Hook, NY, USA.

[4] Barry Becker and Ronny Kohavi. 1996. Adult. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20.

[5] Vadim Borisov, Tobias Leemann, Kathrin Sessler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. 2022. Deep Neural Networks and Tabular Data: A Survey. *IEEE Transactions on Neural Networks and Learning Systems* (2022), 1–21. https://doi.org/10.1109/tnnls.2022.3229161

[6] Francisco M. Castro, Manuel J. Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari. 2018. End-to-End Incremental Learning. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part XII* (Munich, Germany). Springer-Verlag, Berlin, Heidelberg, 241–257. https://doi.org/10.1007/978-3-030-01258-8_15

[7] George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A Efros, and Jun-Yan Zhu. 2022. Dataset distillation by matching training trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 4750–4759.

[8] Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly Detection: A Survey. *ACM Comput. Surv.* 41, 3, Article 15 (July 2009), 58 pages. https://doi.org/10.1145/1541880.1541882

[9] Nitesh Chawla, Kevin Bowyer, Lawrence Hall, and W. Kegelmeyer. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res. (JAIR)* 16 (06 2002), 321–357. https://doi.org/10.1613/jair.953

[10] Yutian Chen, Max Welling, and Alex Smola. 2010. Super-samples from kernel herding. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence* (Catalina Island, CA) *(UAI'10).* AUAI Press, Arlington, Virginia, USA, 109–116.

[11] Andrea Dal Pozzolo, Olivier Caelen, Yann-Aël Le Borgne, Serge Waterschoot, and Gianluca Bontempi. 2014. Learned lessons in credit card fraud detection from a practitioner perspective. *Expert Systems with Applications* 41 (08 2014), 4915–4928. https://doi.org/10.1016/j.eswa.2014.02.026

[12] Dayananda Herurkar, Mario Meier, and Jörn Hees. 2023. RECol: Reconstruction Error Columns for Outlier Detection. In *KI 2023: Advances in Artificial Intelligence: 46th German Conference on AI, Berlin, Germany, September 26–29, 2023, Proceedings* (Berlin, Germany). Springer-Verlag, Berlin, Heidelberg, 60–74. https://doi.org/10.1007/978-3-031-42608-7_6

[13] Dayananda Herurkar, Sebastian Palacio, Ahmed Anwar, Joern Hees, and Andreas Dengel. 2024. Fin-Fed-OD: Federated Outlier Detection on Financial Tabular Data. arXiv:2404.14933 [cs.LG] https://arxiv.org/abs/2404.14933

[14] Dayananda Herurkar, Timur Sattarov, Jörn Hees, Sebastian Palacio, Federico Raue, and Andreas Dengel. 2023. Cross-Domain Transformation for Outlier Detection on Tabular Datasets. In *International Joint Conference on Neural Networks, IJCNN 2023, Gold Coast, Australia, June 18-23, 2023.* IEEE, 1–8. https://doi.org/10.1109/IJCNN54540.2023.10191326

[15] Waleed Hilal, S. Andrew Gadsden, and John Yawney. 2022. Financial Fraud: A Review of Anomaly Detection Techniques and Recent Advances. *Expert Systems with Applications* 193 (2022), 116429. https://doi.org/10.1016/j.eswa.2021.116429

[16] Addison Howard, Bernadette Bouchon Meunier, IEEE CIS inversion, John Lei, Lynn Vesta, Marcus2010, and Prof. Hussein Abbass. 2019. IEEE-CIS Fraud Detection. https://kaggle.com/competitions/ieee-fraud-detection

[17] Wei Jin, Lingxiao Zhao, Shichang Zhang, Yozen Liu, Jiliang Tang, and Neil Shah. 2021. Graph condensation for graph neural networks. *arXiv preprint arXiv:2110.07580* (2021).

[18] Zahra Kazemi and Houman Zarrabi. 2017. Using deep networks for fraud detection in the credit card transactions. In *2017 IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI).* 0630–0633. https://doi.org/10.1109/KBEI.2017.8324876

[19] Yongqi Li and Wenjie Li. 2021. Data distillation for text classification. *arXiv preprint arXiv:2104.08448* (2021).

[20] Dmitry Medvedev and Alexander D'yakonov. 2021. New properties of the data distillation method when working with tabular data. In *Analysis of Images, Social Networks and Texts: 9th International Conference, AIST 2020, Skolkovo, Moscow, Russia, October 15–16, 2020, Revised Selected Papers 9.* Springer, 379–390.

[21] S. Moro, P. Rita, and P. Cortez. 2012. Bank Marketing. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5K306.

[22] Jack Nicholls, Aditya Kuppa, and Nhien-An Le-Khac. 2021. Financial Cybercrime: A Comprehensive Survey of Deep Learning Approaches to Tackle the Evolving Financial Crime Landscape. *IEEE Access* 9 (2021), 163965–163986. https://doi.org/10.1109/ACCESS.2021.3134076

[23] Ebberth L. Paula, Marcelo Ladeira, Rommel N. Carvalho, and Thiago Marzagão. 2016. Deep Learning Anomaly Detection as Support Fraud Investigation in Brazilian Exports and Anti-Money Laundering. In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA).* 954–960. https://doi.org/10.1109/ICMLA.2016.0172

[24] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, G. Sperl, and Christoph H. Lampert. 2016. iCaRL: Incremental Classifier and Representation Learning. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), 5533–5542. https://api.semanticscholar.org/CorpusID:206596260

[25] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. 2018. Deep One-Class Classification. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80),* Jennifer Dy and Andreas Krause (Eds.). PMLR, 4393–4402. https://proceedings.mlr.press/v80/ruff18a.html

[26] Timur Sattarov, Dayananda Herurkar, and Jörn Hees. 2022. Explaining Anomalies using Denoising Autoencoders for Financial Tabular Data. *CoRR* abs/2209.10658 (2022). https://doi.org/10.48550/ARXIV.2209.10658 arXiv:2209.10658

[27] Marco Schreyer, Timur Sattarov, Damian Borth, Andreas Dengel, and Bernd Reimer. 2017. Detection of Anomalies in Large Scale Accounting Data using Deep Autoencoder Networks. https://doi.org/10.48550/ARXIV.1709.05254

[28] Ivan Tomek. 1976. Two modifications of CNN. *IEEE Transactions on Systems, Man, and Cybernetics* SMC-6, 11 (1976), 769–772. https://doi.org/10.1109/TSMC.1976.4309452

[29] Mariya Toneva, Alessandro Sordoni, Rémi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. 2018. An Empirical Study of Example Forgetting during Deep Neural Network Learning. *ArXiv* abs/1812.05159 (2018). https://api.semanticscholar.org/CorpusID:55481903

[30] Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A Efros. 2018. Dataset distillation. *arXiv preprint arXiv:1811.10959* (2018).

[31] Roy Wedge, James Max Kanter, Santiago Moral Rubio, Sergio Iglesias Perez, and Kalyan Veeramachaneni. 2017. Solving the "false positives" problem in fraud prediction. arXiv:1710.07709 [cs.AI]

[32] Xindi Wu, Byron Zhang, Zhiwei Deng, and Olga Russakovsky. 2023. Vision-language dataset distillation. (2023).

[33] I-Cheng Yeh. 2016. Default of Credit Card Clients. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C55S3H.

[34] Bo Zhao and Hakan Bilen. 2023. Dataset condensation with distribution matching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision.* 6514–6523.

[35] Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. 2020. Dataset condensation with gradient matching. *arXiv preprint arXiv:2006.05929* (2020).

[36] Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Dae ki Cho, and Haifeng Chen. 2018. Deep Autoencoding Gaussian Mixture Model for Unsupervised Anomaly Detection. In *International Conference on Learning Representations.* https://api.semanticscholar.org/CorpusID:51805340