

Detecting AI-Generated Content Workshop – COLING 2025

Cross-table Synthetic Tabular Data Detection

G. Charbel N. Kindji ^{1,2}, Lina M. Rojas-Barahona ¹, Elisa Fromont ², Tanguy Urvoy ¹



1

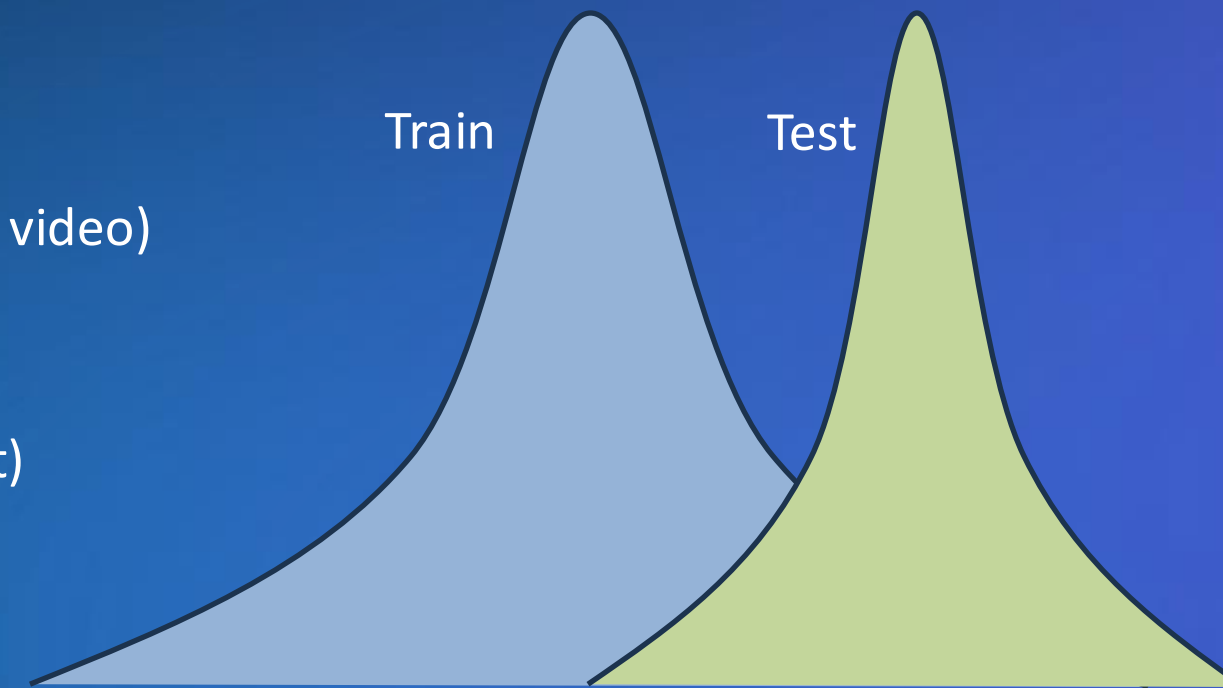


2



Introduction. (1/2)

- Generative AI: more effective data generation models (text, image, audio, video)
 - Risk: data forgery
 - Important to develop detectors
 - Challenging task → Distribution shift (detectors struggle with new content)
-
- Tabular Data Generation → Hot topic
 - High quality tabular data generators (TabSyn ¹, TabDDPM ², TVAE ³, CTGAN ³)
 - Data forgery → E.g. fake accounting results
 - Specific table issue → structure shift



1- Zhang et al., *Mixed-Type Tabular Data Synthesis with Score-based Diffusion in Latent Space*, ICLR 2024

2- Kotelnikov et al., *TabDDPM: Modelling Tabular Data with Diffusion Models*, ICML 2023

3- Xu et al., *Modeling Tabular data using Conditional GAN*, NeurIPS 2019

Introduction. (2/2)

- Synthetic tabular data detection → Classification problem
- Can use the same table structure:

| Product ID | Price | Rating | Label |
|------------|--------|--------|-----------|
| P001 | 19.99 | 4.5 | Real |
| P265 | 29.99 | 3.0 | Real |
| P4565 | 199.99 | 5.0 | Synthetic |
| P018 | 39.99 | 4.2 | Real |
| P107 | 100.00 | 8.5 | Synthetic |

- We focus on **cross-table** (cross-structure) and **table-agnostic** detectors:

| Table Rows | | | | | Source |
|------------|------------|--------|--------|---------|-----------|
| Product ID | Price | Rating | | | Real |
| P001 | 19.99 | 4.5 | | | |
| Fruit | Quantity | | | | Synthetic |
| Apple | 54,80 | | | | |
| Employee | Department | Salary | Level | | Real |
| E2535 | Sales | 75000 | Senior | | |
| Model | Brand | Year | RAM | Price | Real |
| XPS 13 | Dell | 2022 | 16GB | 1299.99 | |
| Name | Major | GPA | | | Synthetic |
| Alice S. | Biology | 15.5 | | | |
| Fruit | Quantity | | | | Real |
| Banana | 15 | | | | |

Synthetic Tabular Data Detection – No Shift.

Training Data

| Table Rows | | | | | Source |
|------------|------------|--------|--------|---------|--------|
| Product ID | Price | Rating | | | Real |
| P001 | 19.99 | 4.5 | | | |
| Fruit | Quantity | | | | TVAE |
| Apple | 54,80 | | | | |
| Employee | Department | Salary | Level | | Real |
| E2535 | Sales | 75000 | Senior | | |
| Model | Brand | Year | RAM | Price | Real |
| XPS 13 | Dell | 2022 | 16GB | 1299.99 | |
| Name | Major | GPA | | | TabSyn |
| Alice S. | Biology | 15.5 | | | |

- Training and testing on the same set of tables
- Two approaches:
 - Single Generator vs Real
 - All Generators vs Real

Testing Data

| Table Rows | | | | | Source |
|-------------|------------|--------|--------|---------|---------|
| Employee | Department | Salary | Level | | TabSyn |
| E0458 | Marketing | 35000 | Junior | | |
| Name | Major | GPA | | | TabDDPM |
| John D. | History | 3.9 | | | |
| Fruit | Quantity | | | | Real |
| Mango | 3,14 | | | | |
| Product ID | Price | Rating | | | Real |
| P6659 | 100.00 | 4.5 | | | |
| Model | Brand | Year | RAM | Price | Real |
| ThinkPad X1 | Lenovo | 2023 | 16GB | 2265,98 | |

Synthetic Tabular Data Detection – Cross-table Shift.

Training Data

Training and testing on different table structures (cross-structure detectors)

| Table Rows | | | | | Source |
|------------|------------|--------|--------|---------|--------|
| Product ID | Price | Rating | | | Real |
| P001 | 19.99 | 4.5 | | | |
| Fruit | Quantity | | | | TVAE |
| Apple | 54,80 | | | | |
| Employee | Department | Salary | Level | | Real |
| E2535 | Sales | 75000 | Senior | | |
| Model | Brand | Year | RAM | Price | Real |
| XPS 13 | Dell | 2022 | 16GB | 1299.99 | |
| Name | Major | GPA | | | TabSyn |
| Alice S. | Biology | 15.5 | | | |

Testing Data

| Table Rows | | | | | Source |
|------------|------------|---------|-----------|-----------|--------|
| Country | Population | | | | CTGAN |
| Canada | 409,19 | | | | |
| Event ID | Name | Date | Location | Attendees | Real |
| 001 | COLING | 2025 | Abu Dhabi | 14678 | |
| Course ID | Instructor | Credits | | | TabSyn |
| CS4A A | Jack S. | -75 | | | |
| Brand | Model | Year | | | TVAE |
| Toyota | Camry | 1256 | | | |
| Month | Sales | Region | Growth | | Real |
| January | 450000 | South | 15% | | |

Text-based Encodings.

- Row linearization:

| Table Rows | | | | | Source |
|------------|------------|--------|--------|---------|-----------|
| Product ID | Price | Rating | | | Real |
| P001 | 19.99 | 4.5 | | | |
| Fruit | Quantity | | | | Synthetic |
| Apple | 54,80 | | | | |
| Employee | Department | Salary | Level | | Real |
| E2535 | Sales | 75000 | Senior | | |
| Model | Brand | Year | RAM | Price | Real |
| XPS 13 | Dell | 2022 | 16GB | 1299.99 | |
| Name | Major | GPA | | | Synthetic |
| Alice S. | Biology | 15.5 | | | |



| Table Rows | Source |
|---|-----------|
| "Product ID:P001,Price:19.99,Rating:4.5" | Real |
| "Fruit:Apple,Quantity:54,80" | Synthetic |
| "Employee:E2535,Department:Sales,Salary:75000,Level:Senior" | Real |
| "Model:XPS 13,Brand:Dell,Year:2022,RAM:16GB,Price:1299.99" | Real |
| "Name:Alice S.,Major:BiologY,GPA:15.5" | Synthetic |

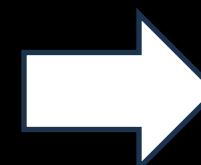
- Logistic Regression → characters trigrams:
 - "Pro", "duc", "t l", "D:P"
- Text-based Transformer → sequence of character embedded + positionnal embedding

Table-based Encoding.

- Data is kept in its tabular form

Step 1: Separate numerical and categorical features

| Num. Features | | Cat. Features | | | Source |
|---------------|---------|---------------|------------|--------|-----------|
| Price | Rating | Product ID | | | Real |
| 19.99 | 4.5 | P001 | | | |
| Quantity | | Fruit | | | Synthetic |
| 54,80 | | Apple | | | |
| Salary | | Employee | Department | Level | Real |
| 75000 | | E2535 | Sales | Senior | |
| Year | Price | Model | Brand | RAM | Real |
| 2022 | 1299.99 | XPS 13 | Dell | 16GB | |
| GPA | | Name | Major | | Synthetic |
| 15.5 | | Alice S. | Biology | | |



Step 2: Apply padding

| Num. Features | | Cat. Features | | | Source |
|---------------|---------|---------------|------------|--------|-----------|
| Price | Rating | Product ID | X | X | Real |
| 19.99 | 4.5 | P001 | X | X | |
| Quantity | X | Fruit | X | X | Synthetic |
| 54,80 | X | Apple | X | X | |
| Salary | X | Employee | Department | Level | Real |
| 75000 | X | E2535 | Sales | Senior | |
| Year | Price | Model | Brand | RAM | Real |
| 2022 | 1299.99 | XPS 13 | Dell | 16GB | |
| GPA | X | Name | Major | X | Synthetic |
| 15.5 | X | Alice S. | Biology | X | |

Experiments.

14 Tables

| Name | Size | #Num | #Cat |
|-----------------|--------|------|------|
| Abalone | 4177 | 7 | 2 |
| Adult | 48842 | 6 | 9 |
| Bank Marketing | 45211 | 7 | 10 |
| Black Friday | 166821 | 6 | 4 |
| Bike Sharing | 17379 | 9 | 4 |
| Cardio | 70000 | 11 | 1 |
| Churn Modelling | 4999 | 8 | 4 |
| Diamonds | 26970 | 7 | 3 |
| HELOC | 5229 | 23 | 1 |
| Higgs | 98050 | 28 | 1 |
| House 16H | 22784 | 17 | 0 |
| Insurance | 1338 | 4 | 3 |
| King | 21613 | 19 | 1 |
| MiniBooNE | 130064 | 50 | 1 |

4 Generators ¹

Modeling Tabular Data using Conditional GAN

Lei Xu
MIT LIDS
Cambridge, MA
lei@mit.edu

Maria Skoularidou
MRC-BSU, University of Cambridge
Cambridge, UK
ms2407@cam.ac.uk

TABDDPM: MODELLING TABULAR DATA WITH DIFFUSION MODELS

Akzim Kotelnikov
HSE, Yandex
ya@akotelnikov.ru

Dmitry Baranchuk
Yandex

Ivan Rubachev
HSE, Yandex

Artem Babenko
Yandex

MIXED-TYPE TABULAR DATA SYNTHESIS WITH SCORE-BASED DIFFUSION IN LATENT SPACE

Hengrui Zhang^{1*} Jiani Zhang^{2,1} Balasubramaniam Srinivasan² Zhengyuan Shen²
Xiao Qin² Christos Faloutsos² Huzefa Rangwala^{2,3} George Karypis²
¹Computer Science Department, University of Illinois at Chicago ²Amazon Web Services
³Computer Science, George Mason University
hszhan55@uic.edu {zhajiani, srbsalas, donshen}@amazon.com
{dxrqin, faloutsos, chusefa, gkarypis}@amazon.com

ABSTRACT

Recent advances in tabular data generation have greatly enhanced synthetic data quality. However, extending diffusion models to tabular data is challenging due to the intricately varied distributions and a blend of data types of tabular data. This paper introduces TABSYN, a methodology that synthesizes tabular data by leveraging a diffusion model within a variational autoencoder (VAE) crafted latent space. The key advantages of the proposed TABSYN include (1) **Generality**: the ability to handle a broad spectrum of data types by converting them into a single unified space and explicitly capture inter-column relations. (2) **Quality**: optimizing the distribution of latent embeddings to enhance the subsequent training of diffusion models, which helps generate high-quality synthetic data. (3) **Speed**: much fewer number of reverse steps and faster synthesis speed than existing diffusion-based methods. Extensive experiments on six datasets with five metrics demonstrate that TABSYN outperforms existing methods. Specifically, it reduces the error rates by 86% and 67% for column-wise distribution and pair-wise column correlation estimations compared with the most competitive baselines. Code has been made available at <https://github.com/amazon-science/tabsyn>.

TVAE &
CTGAN ²

TabDDPM ³

TabSyn ⁴

3 Detectors

- Logistic Regression on character trigrams
- Two Transformer Architectures:
 - text-based and
 - table-based

3 Setups

- No-shift setup: training and testing on the same set of tables
 - Single Generator
 - All Generators
- Cross-table shift: testing on a distinct set of tables

1- Kindji et al., *Under the Hood of Tabular Data Generation Models: Benchmarks with Extensive Tuning*, ArXiv preprint 2024

2- Xu et al., *Modeling Tabular data using Conditional GAN*, NeurIPS 2019

3- Kotelnikov et al., *TabDDPM: Modelling Tabular Data with Diffusion Models*, ICML 2023

4- Zhang et al., *Mixed-Type Tabular Data Synthesis with Score-based Diffusion in Latent Space*, ICLR 2024

Results – No Shift Setup.

Single Generator (TVAE) vs Real

- Decent performance of detectors (both AUC and Accuracy)
- AUC = 0.92 with table-based Transformer
- Kindji et al. ¹ → Avg. AUC for table-wise TVAЕ's sample detection = 0.81

All Generators vs Real

- Little drop of performance
- Transformer-based approaches delivers best results

| Cross-structure Setup | Model | Metrics | |
|------------------------------------|---------------|--------------------|--------------------|
| | | AUC | Accuracy |
| Single Generator (TVAE) vs Real | 3grm-LReg. | 0.72 ± 0.00 | 0.65 ± 0.00 |
| | Text-Transf. | 0.76 ± 0.01 | 0.67 ± 0.01 |
| | Table-Transf. | 0.92 ± 0.00 | 0.83 ± 0.00 |
| All Generators vs Real | 3grm-LReg. | 0.64 ± 0.00 | 0.59 ± 0.00 |
| | Text-Transf. | 0.73 ± 0.04 | 0.66 ± 0.05 |
| | Table-Transf. | 0.77 ± 0.00 | 0.69 ± 0.00 |

Results – Cross-table Shift.

- Cross-table shift is a challenging task
- Table-based model struggle to generalize (AUC=0.51)
- Text-based approaches seems to generalize better (AUC=0.60)

| Model | Metrics | |
|---------------|--------------------|--------------------|
| | AUC | Accuracy |
| 3grm-LReg. | 0.60 ± 0.05 | 0.52 ± 0.03 |
| Text-Transf. | 0.60 ± 0.07 | 0.52 ± 0.01 |
| Table-Transf. | 0.51 ± 0.00 | 0.50 ± 0.00 |

Conclusion and Future Work.

- No-shift → reasonable performance
- Cross-table synthetic tabular data detection → challenging task
- Data encoding is key for detector's performance (table-based vs text-based Transformer)
- Further investigation on Transformer-based detectors
 - Incorporating metadata in the model (column names / description, ...)
 - Using pretrained models like TaBERT ¹
 - Explore additional distribution shift setups (cross-generator shift)



Cross-Table Synthetic Tabular Data Detection

G. Charbel N. Kindji, Lina M. Rojas-Barahona, Elisa Fromont, Tanguy Urvoy

Thank you

charbel.kindji@orange.com

charbel.kindji.orange@gmail.com

