Gretel.ai: Open-Source Artificial Intelligence Tool To Generate New Synthetic Data

A.H. Noruzman[1-2], N.A. Ghani[1-2], N.S.A. Zulkifli[1],

[1]Faculty of Computing, University Malaysia Pahang,26600 Pekan, Pahang, Malaysia.
[2]Centre for Software Development and Integrated Computing, University Malaysia Pahang, Lebuhraya Tun Razak, 26300 Gambang, Pahang, Malaysia.

Corresponding Author's Email: [1]ainiehayatinoruzman@gmail.com

Article History: Received 200421; Revised 280621; Accepted 020921;

ABSTRACT – Nowadays, machine learning is widely employed to solve real-world problems, particularly in medical and diagnostics fields. However, to trained a machine learning model required a massive amount of data making it challenging due to the demand access of medical data is rigid since the data is kept confidential, secure, and difficult to obtain. Therefore, this paper introduces Gretel.ai, an artificial intelligence tool for generating synthetic data that related to Autism Spectrum Disorder (ASD). The results demonstrate the proposed framework for generating synthetic data using the Autism Quotient 10 (AQ10) screening instrument growth from 1054 original records to 5000 synthetic records while preserving the primary characteristics of the originals dataset's features. Additionally, the Gretel.ai also provide a quick synthetic report that quantifies the utility on exploratory data analysis by presenting data summary statistics comparing the synthesized and the original training data comparisons.

KEYWORDS: Synthetic data, Machine Learning, Synthetic Report, ASD dataset, Synthetic tool

1.0 INTRODUCTION

Artificial Intelligence (A.I.) and machine learning solve real-world problems, particularly in the medical industry and diagnostics. This is employed to access an enormous amount of high-quality data to develop successful A.I. and machine learning models, especially for medical data. However, gathering such data is difficult due to the sensitive nature of the data, and Personal Health Information (PHI) cannot be obtained easily without permission [1]. To solve this issue, a new technology known as Synthetic Data was implemented. Synthetic data are used in many A.I. projects and machine learning models. According to [2] on the Gartner blog, by 2024, 60% of data utilized for developing A.I. and analytical projects will be synthetically generated, and consumption for A.I. projects will increase by 2030, whereby synthetic data would eventually supplant actual data in A.I. models as shown in Figure 1.

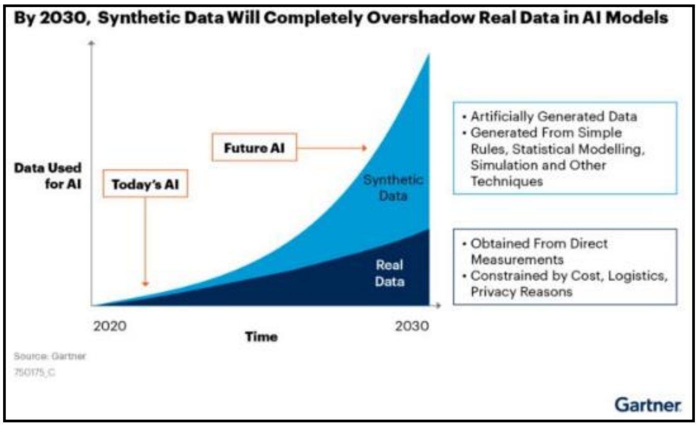


Figure 1. Synthetic Data will replace actual data in A.I. models by 2030 [2]

Synthetic data are generated artificially by A.I. algorithms rather than by actual events [3]. The synthetic data preserve the inclusive properties and characteristics of the original data [4]. Thus, the synthetic data are easily shared while preserving the privacy and security of the actual data. In 2021, more than one synthetic data vendor had advanced because of the demand for synthetic data and offering services through their platforms and APIs [5]. Figure 2 shows the list of data vendors divided into companies that offer synthetic data for structured data and others for unstructured data. In addition, more than one open-source tools are available online to produce synthetic data, for example, GenerateData [6], Synth [7], Sdv [8], Mostly A.I. [9], and Wakefield [10]. However, utilizing an open-source project may demand additional technical skills to install and utilize the library.

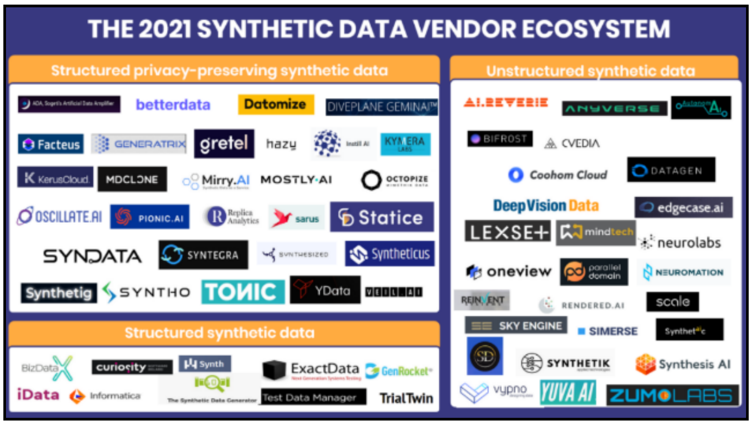


Figure 2. The 2021 Synthetic Data Vendor Ecosystem [5]

Machine learning has become a trend to solve real-world problems, especially in the medical field and disease diagnosis. Machine learning requires a massive amount of data to train a machine learning model, especially in medical data; however, the main challenges in the medical domain are how to cope with the small datasets and the limited amount of annotated records[11]. Even though there is an online data repository with many medical datasets hence, there are still small and only valid for some medical issues. In order to meet the machine learning requirements of a large dataset, the small dataset must be transformed into a massive amount dataset so that the process can be applied and transformed into useful and ageable output.

Therefore, this paper focused on Autism Spectrum Disorder (ASD) medical dataset using the Autism Quotient 10 screening tool (AQ10). The AQ10, on the other hand, has only 1054 records, and synthetic application tools can help generate and optimize datasets while keeping the original dataset features and properties unchanged. This study also shows how to use the Gretel [12], an open-source synthetic program that can conduct and generate both an original and synthetic report.

2.0 METHODOLOGY

The framework of the methodology is illustrated in Figure 3. It consists of the original input dataset with 1054 records, synthetic open-source applications, and the synthetic dataset's output. The synthetic application processes the original input data using applications algorithms and validates the training data to have a confidence quality to be used. The output is synthetic 5000 data records.

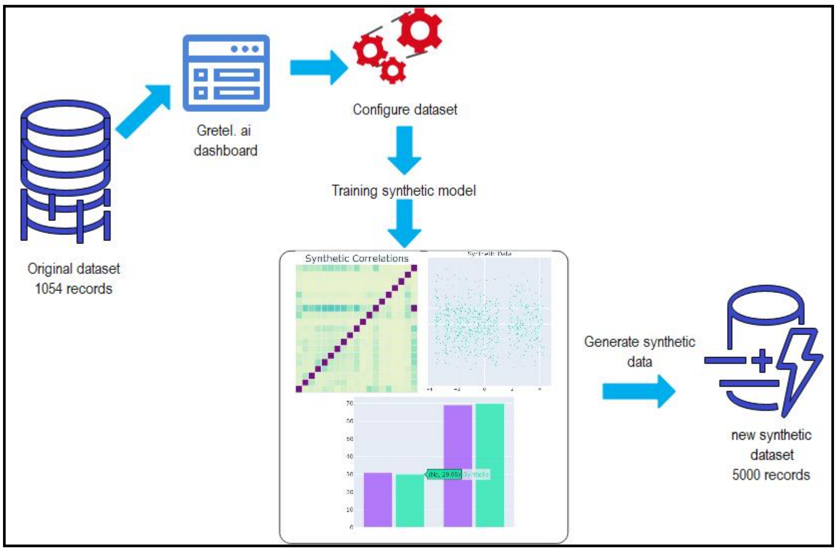


Figure 3. The framework of synthetic application

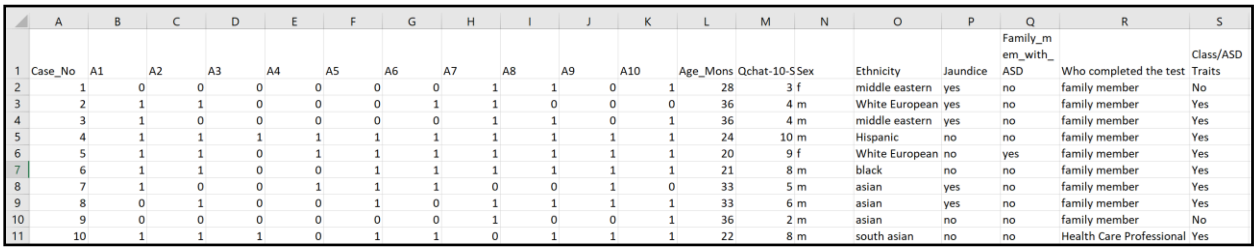
2.1 Dataset

The dataset used in this study was generated and made publicly available on the website https://www.kaggle.com/fabdelja/autism-screening-for-toddlers [13]. In addition, the website is verified and permitted to use the dataset for research context only. The detailed data description is summarized in Table 1.

Table 1. Attribute List Summary.

Attribute No	Attribute	Value
1	Case No	Number (1-1054)
2-11	A1 -A10	Binary (0,1)
12	Age_mons	Number between 12-36
13	Qchat-10-Score	Number between (1-10)
14	Sex	String
15	Ethnicity	String
16	Jaundice	Boolean (True or False)
17	Family_mem_with_ASD	Boolean (True or False)
18	Who completed the test	String
19	Class/ASD Traits	ASD / Non ASD class (String)

Table 1 summarizes the detailed description of the dataset. The name of the dataset is "Toddler Autism dataset July 2018", which contains 1054 records and 19 variables indicating various features, ten of which are questions used to establish whether a toddler has ASD, denoted by items A1 through A10 in the table above. These ten questions are taken from the clinically validated Autism Quotient 10 (AQ10) for toddlers. The toddlers are considered positive for ASD if the score is six or above [14]. Figure 4 shows a sample of 10 records in comma-separated values (CSV) format of the Toddler dataset.

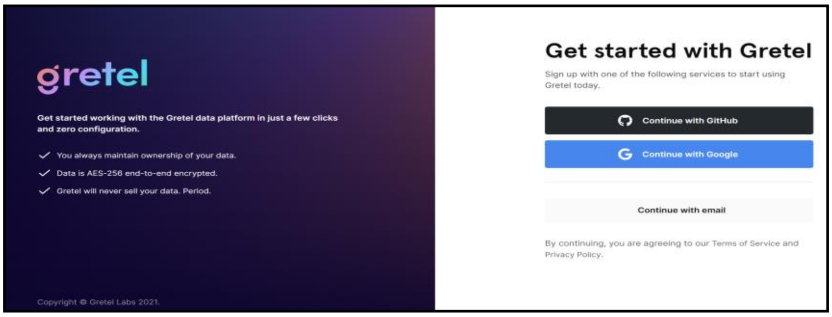


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Case_No	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	Age_Mons	Qchat-10-S	Sex	Ethnicity	Jaundice	Family_mem_with_ASD	Who completed the test	Class/ASD Traits
2	1	0	0	0	0	0	0	1	1	0	1	28	3	f	middle eastern	yes	no	family member	No
3	2	1	1	0	0	0	1	1	0	0	0	36	4	m	White European	yes	no	family member	Yes
4	3	1	0	0	0	0	0	1	1	0	1	36	4	m	middle eastern	yes	no	family member	Yes
5	4	1	1	1	1	1	1	1	1	1	1	24	10	m	Hispanic	no	no	family member	Yes
6	5	1	1	0	1	1	1	1	1	1	1	20	9	f	White European	no	yes	family member	Yes
7	6	1	1	0	0	1	1	1	1	1	1	21	8	m	black	no	no	family member	Yes
8	7	1	0	0	1	1	1	0	0	1	0	33	5	m	asian	yes	no	family member	Yes
9	8	0	1	0	0	1	0	1	1	1	1	33	6	m	asian	yes	no	family member	Yes
10	9	0	0	0	0	0	0	1	0	0	1	36	2	m	asian	no	no	family member	No
11	10	1	1	1	0	1	1	0	1	1	1	22	8	m	south asian	no	no	Health Care Professional	Yes

Figure 4. Sample of 10 ASD records in CSV format

2.2 The Gretel Dashboard

In this step, the Gretel tool is used to generate synthetic data. The application is at https://gretel.ai/. To get started, sign in or sign-up email is compulsory. Figure 5 illustrates the gretel.ai dashboard.



Get started working with the Gretel data platform in just a few clicks and zero configuration.

- ✓ You always maintain ownership of your data.
- ✓ Data is AES-256 end-to-end encrypted.
- ✓ Gretel will never sell your data. Period.

Copyright © Gretel Labs 2021.

Get started with Gretel

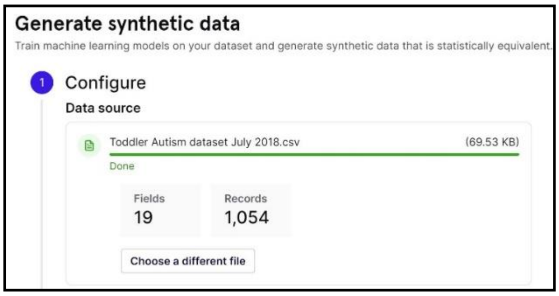
Sign up with one of the following services to start using Gretel today.

By continuing, you are agreeing to our Terms of Service and Privacy Policy.

Figure 5. The gretel.ai dashboard

2.3 Generate Synthetic Data

The Gretel tool trains the data and generates synthetic data. The original data source is configured since there are no missing values in the dataset, as shown in Figure 6. Each activity states the status configured from the model created until results are generated via report form. Figures 7 and 8 depict the log activity in the gretel.ai tool application.



Generate synthetic data

Train machine learning models on your dataset and generate synthetic data that is statistically equivalent.

Configure

Data source

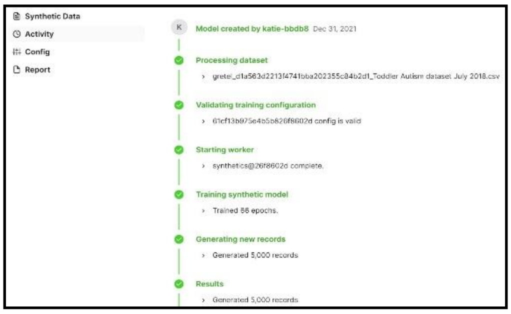
Toddler Autism dataset July 2018.csv (69.53 KB)

Done

Fields 19

Records 1,054

Figure 6. Configuration of synthetic data



Model created by katie-bbdb8 Dec 31, 2021

Processing dataset
› gretel_d1a563d2213f4741bba202355c84b2d1_Toddler Autism dataset July 2018.csv

Validating training configuration
› 61cf13b975e4b5b826f8602d config is valid

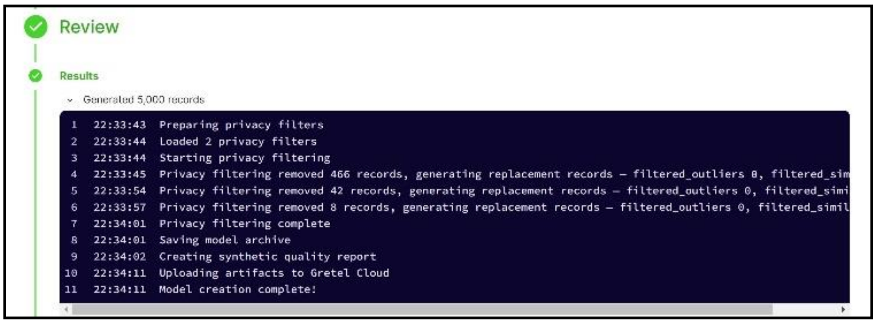
Starting worker
› synthetics@26f8602d complete.

Training synthetic model
› Trained 56 epochs.

Generating new records
› Generated 5,000 records

Results
› Generated 5,000 records

Figure 7. The activity in the gretel.ai



Review

Results

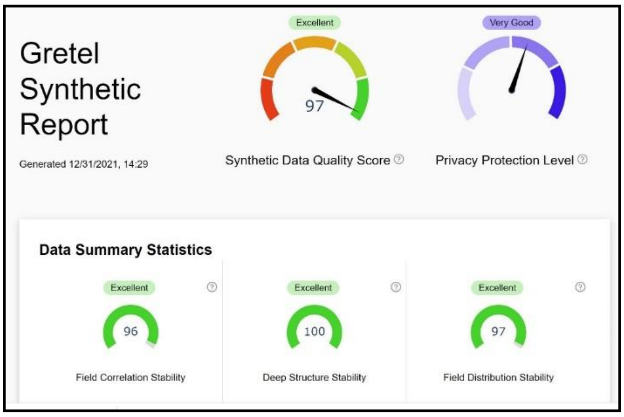
⌄ Generated 5,000 records

```
22:33:43  Preparing privacy filters
22:33:44  Loaded 2 privacy filters
22:33:44  Starting privacy filtering
22:33:45  Privacy filtering removed 466 records, generating replacement records — filtered_outliers 0, filtered_sim
22:33:54  Privacy filtering removed 42 records, generating replacement records — filtered_outliers 0, filtered_simi
22:33:57  Privacy filtering removed 8 records, generating replacement records — filtered_outliers 0, filtered_simil
22:34:01  Privacy filtering complete
22:34:01  Saving model archive
22:34:02  Creating synthetic quality report
22:34:11  Uploading artifacts to Gretel Cloud
22:34:11  Model creation complete!
```

Figure 8. Configuration of synthetic data

3.0 RESULT

The Gretel application produced a report, namely Gretel Synthetic Report, consisting of sections of Synthetic Data Quality Score, Privacy Protection Level, Data Summary Statistics, Privacy Protection Summary, Training Field Overview, Training, and Synthetic Data Correlation, Principal Component Analysis and Field Distribution Comparisons. Each of the sections compares the synthesized and the original training data. The report is attached with the question mark symbol for a detailed explanation of the report. Figure 9 shows the score of the Synthetic Data Quality Score that measures how well the generated synthetic data maintains the original dataset's statistical features. The statistics show that the confidence score is 97 percent, and the synthetic data is a nail for quality confidence. On the other hand, the Privacy Protection Level measures protect synthetic data against adversarial attacks, which shows a suitable protection level.



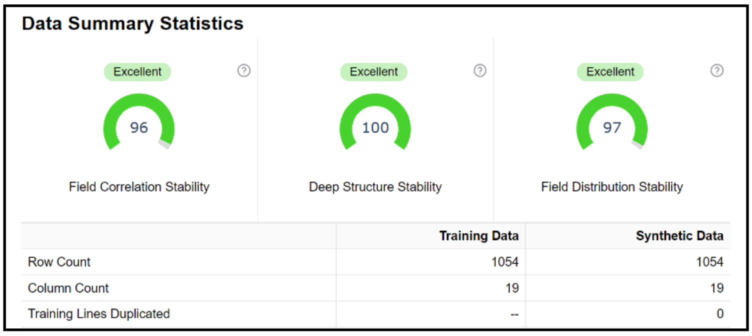
Gretel Synthetic Report

Generated 12/31/2021, 14:29

Data Summary Statistics

Figure 9. Gretel Synthetic Report

As shown in Figure 10, the Data Summary Statistics section includes scores for Field Correlation, Deep Structure, and Field Distribution Stability. The summary also includes the row and column counts, which are the same for training and synthetic data and have no duplicate training lines.



Data Summary Statistics

	Training Data	Synthetic Data
Row Count	1054	1054
Column Count	19	19
Training Lines Duplicated	--	0

Figure 10. Data Summary Statistic

Gretel's effect on the original data yielded the following data files: training and synthetic data. The synthetic data files calculated average, and the results are closer to the original dataset. The report includes a heatmap that depicts the relationships contributing to a 96 per cent score. The X-axis of the heat map is Class/ASD Traits, and the Y-axis is A9, which shows a correlation difference of 0.03, the train correlation is 0.31, and the synthetic correlation is 0.28. The higher the score shows the stability of the data to aid in the comparisons. Figure 11 shows the heat map of the Toddlers dataset.

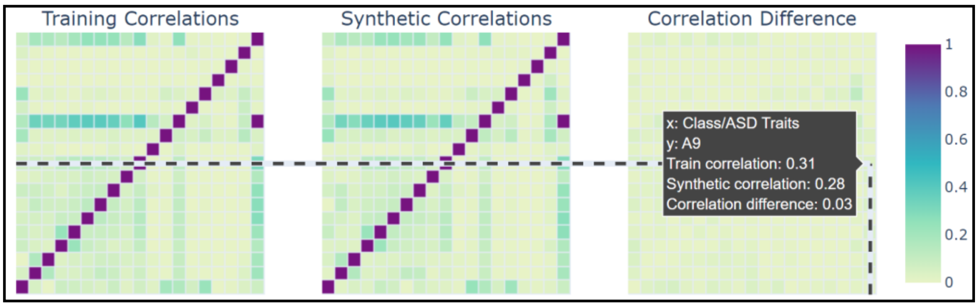


Figure 11. Heat map for the Toddlers dataset

The Deep Structure Stability checks the relationships between attributes through Principal Component Analysis (PCA). Gretel compares PCA by computing first on the original data, then comparing the distribution dependence again on the synthetic data. As a result, the 100% statistical integrity in the synthetic score given the higher score of the quality, confidence, and data distributions. Figure 12 illustrates the distribution dependence relationship between attributes.

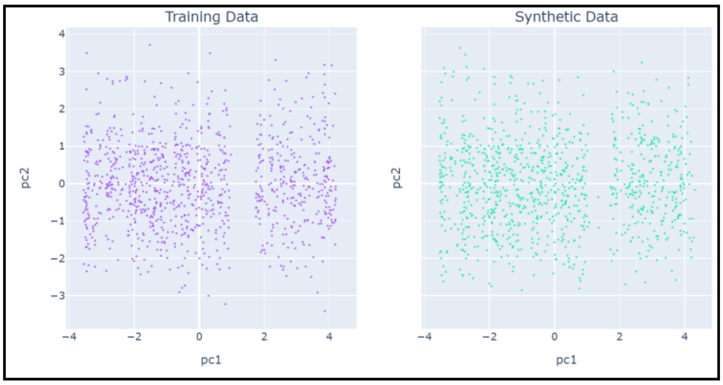


Figure 12. Distribution dependence relationship between attributes

The Field Distribution Stability statistic quantifies how the synthetic data replicates the original data. Gretel constructed a marginal histogram with a 97 percent distribution on synthetic data. As illustrated in Figure 13, the purple histogram represents the differences of original data, whereas the green histogram represents synthetic data.

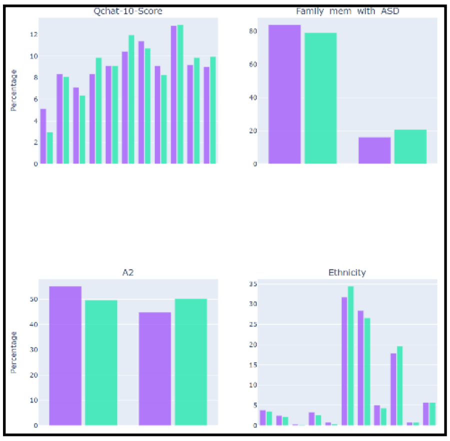
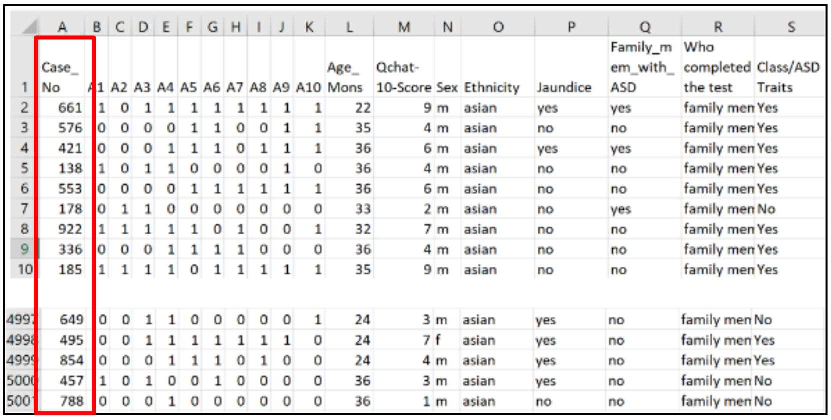


Figure 13. Marginal Histogram

The Gretel tool transformed the original data of 1054 records into synthetic data of 5000 records with 19 attributes, as shown in Table 2. Figure 14 shows the synthetic data with randomized Case No attribute displayed after training in the gretel.ai application.

Table 2. Comparison of Original and Synthetic Data Attribute.

Attribute No	Attribute	Original data	Synthetic data
1	Case No	Number (1-1054) Ascending	Number (1-5000) Randomized
2-11	A1 -A10	Binary (0,1)	Binary (0,1)
12	Age_mons	Number between 12-36	Number between 12-36
13	Qchat-10-Score	Number between (1-10)	Number between (1-10)
14	Sex	String	String
15	Ethnicity	String	String
16	Jaundice	Boolean (True or False)	Boolean (True or False)
17	Family_mem_with_ASD	Boolean (True or False)	Boolean (True or False)
18	Who completed the test	String	String
19	Class/ASD Traits	ASD / Non ASD class (String)	ASD / Non ASD class (String)



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Case_ No	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	Age_ Mons	Qchat- 10-Score	Sex	Ethnicity	Jaundice	Family_m em_with_ ASD	Who completed the test	Class/ASD Traits
2	661	1	0	1	1	1	1	1	1	1	1	22	9	m	asian	yes	yes	family men	Yes
3	576	0	0	0	0	1	1	0	0	1	1	35	4	m	asian	no	no	family men	Yes
4	421	0	0	0	1	1	1	0	1	1	1	36	6	m	asian	yes	yes	family men	Yes
5	138	1	0	1	1	0	0	0	0	1	0	36	4	m	asian	no	no	family men	Yes
6	553	0	0	0	0	1	1	1	1	1	1	36	6	m	asian	no	no	family men	Yes
7	178	0	1	1	0	0	0	0	0	0	0	33	2	m	asian	no	yes	family men	No
8	922	1	1	1	1	1	0	1	0	0	1	32	7	m	asian	no	no	family men	Yes
9	336	0	0	0	1	1	1	1	0	0	0	36	4	m	asian	no	no	family men	Yes
10	185	1	1	1	1	0	1	1	1	1	1	35	9	m	asian	no	no	family men	Yes
4997	649	0	0	1	1	0	0	0	0	0	1	24	3	m	asian	yes	no	family men	No
4998	495	0	0	1	1	1	1	1	1	1	0	24	7	f	asian	yes	no	family men	Yes
4999	854	0	0	0	1	1	1	0	1	0	0	24	4	m	asian	yes	no	family men	Yes
5000	457	1	0	1	0	0	1	0	0	0	0	36	3	m	asian	yes	no	family men	No
5001	788	0	0	0	1	0	0	0	0	0	0	36	1	m	asian	no	no	family men	No

Figure 14. Synthetic Data with Randomized Case No Attribute

4.0 CONCLUSION

This paper investigated the existing artificial intelligence tool, the Gretel.ai, for generating synthetic data and intelligent data analysis and applications. Using the ASD Toddler dataset, which is publicly available, the proposed framework can generate synthetic data ranging from 1054 to 5000 records without changing the original features. By viewing the graphics using the ASD dataset, the tool also provides a quick report, namely the Gretel Synthetic Report, which can help quantify their utility on exploratory data analysis. With these benefits and the availability of synthetic data, it will likely become the future of Artificial Intelligence. In due course, the synthetic data will replace actual data to become primary data generation for future references.

REFERENCES

[1] T. Davenport and R. Kalakota, “The potential for artificial intelligence in Healthcare,” *Future Healthcare Journal*, vol. 6, no. 2, pp. 94–98, 2019.

[2] White, A.(2022). *By 2024, 60% of the data used for the development of A.I. and analytics projects will be synthetically generated - Andrew White*. [online] https://www.gartner.com/. Available at: <https://blogs.gartner.com/andrew_white/2021/07/24/by-2024-60-of-the-data-used-for-the-development-of-ai-and-analytics-projects-will-be-synthetically-generated/>

[3] J. Snoke, G. M. Raab, B. Nowok, C. Dibben, and A. Slavkovic, “General and specific utility measures for synthetic data,” *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 181, no. 3, pp. 663–688, 2018.

[4] M. Hittmeir, A. Ekelhart, and R. Mayer, “On the utility of Synthetic Data,” *Proceedings of the 14th International Conference on Availability, Reliability and Security*, 2019.

[5] Devaux, E., 2022. List of synthetic data startups and companies — 2021. [Blog] *elise-deux.medium.com*, Available at: <https://elise-deux.medium.com/the-list-of-synthetic-data-companies-2021-5aa246265b42>

[6] “Generatedata.Com,” *Generatedata.com*. [Online]. Available: https://generatedata.com/

[7] “Synth,” *Synth*. [Online]. Available: https://www.getsynth.com/

[8] “The Synthetic Data Vault. Put synthetic data to work!,” *Sdv. dev*. [Online]. Available: https://sdv.dev/

[9] “MOSTLY AI, the synthetic data company,” *MOSTLY AI*, 03-Sep-2021. [Online]. Available: https://mostly.ai/

[10] T. Rinker, *Wakefield: Generate random data sets*.

[11] M. J. Willemink *et al.*, “Preparing medical imaging data for machine learning,” *Radiology*, vol. 295, no. 1, pp. 4–15, 2020.

[12] “Gretel.Ai - privacy engineering as a service,” *Gretel.ai*. [Online]. Available: https://gretel.ai/

[13] F. Thabtah, “Autism screening data for Toddlers,” *Kaggle*, 23-Jul-2018. [Online]. Available: https://www.kaggle.com/fabdelja/autism-screening-for-toddlers.

[14] Allison C, Auyeung B, Baron-Cohen S. Toward brief “Red Flags” for autism screening: The Short Autism Spectrum Quotient and the Short Quantitative Checklist for Autism in toddlers in 1,000 cases and 3,000 controls [corrected]. J Am Acad Child Adolesc Psychiatry. 2012 Feb;51(2):202-212.e7. DOI: 10.1016/j.jaac.2011.11.003. Epub 2011 Dec 30. Erratum in: J Am Acad Child Adolesc Psychiatry. 2012 Mar;51(3):338. PMID: 22265366.