

Crafting Artificial Insights: Mastering Synthetic Data for Model Training

Author 1

Dr. Ranjith Gopalan PhD,

Principal consultant, Cognizant

Email: ranjith.gopalan@gmail.com

Author 2

Dr. Ganesh Kumar Sivasubramanian PhD,

Principal consultant, Cognizant

Email: ganeshsivaa@gmail.com

Abstract

The increasing reliance on machine learning has led to a pressing need for diverse and representative training data. However, acquiring real-world data can be challenging due to privacy concerns, data scarcity, and the excessive cost of data collection. Synthetic data generation has emerged as a promising solution, offering opportunities to safeguard privacy, increase data availability, and reduce bias in machine learning models. This research paper presents a comprehensive exploration of synthetic data generation techniques, highlighting their advantages, challenges, and implications for sustainable development.

To address this challenge, the research presented here explores expanding a deep learning framework that employs Variational Autoencoders and Generative Adversarial Networks to create customizable synthetic data. The proposed Variational Autoencoders framework is designed to digitally generate data as needed, conforming to user-defined specifications. This approach, with its wide-ranging and generalized capabilities, addresses the gap in customized, synthetic data generation, where previous efforts were limited to specific domains. Paper also talks about **Evaluating Synthetic Data Quality, Ethical Considerations and Challenges, Future Trends in Synthetic Data Generation.**

Keywords: Synthetic Data, Machine Learning, Data Augmentation, Generative Adversarial Networks, Variational Autoencoders, Privacy-Preserving, Bias Reduction

Introduction

Synthetic data refers to information that is artificially generated rather than obtained from real-world events. This data is created using algorithms that simulate the statistical properties of genuine datasets. The creation of synthetic data allows researchers and practitioners in artificial intelligence and machine learning to overcome the limitations of real-world data, such as privacy concerns, data scarcity, and the challenges of data collection. By mirroring the characteristics of authentic data, synthetic data serves as a valuable resource for model training, testing, and validation, enabling more robust and reliable AI systems.

The importance of synthetic data lies in its ability to provide diverse and representative datasets that can enhance model performance. In many cases, real-world datasets may be imbalanced, incomplete, or biased, which can lead to suboptimal model training outcomes. Synthetic data can be tailored to fill these gaps, ensuring that models can learn from a comprehensive array of scenarios. This customization helps to improve the generalization capabilities of machine learning models, equipping them to perform well on unseen data and real-world applications.

Synthetic data can help protect the privacy of individuals whose information is used in a model. This is done by replacing sensitive details in the original data with synthetic data that maintains the overall characteristics of the data but does not contain the original sensitive information. A report on synthetic data is shown in Figure 1 below. Additionally, Gartner forecasts that by 2030, most data used in AI will be artificially generated through techniques like rules, statistical models, and simulations.

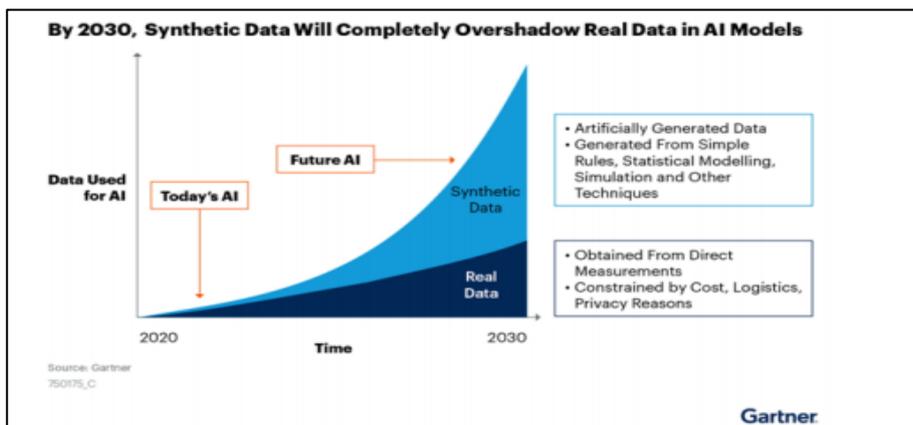


Figure 1: Gartner report on Synthetic Data

The widespread adoption of deep learning models across diverse applications, such as image and speech recognition, natural language processing, and computer vision, has been a notable trend. One of the key advantages of these deep learning models is their capability to generate synthetic data that closely resembles real-world data.

Methodology

Crafting customizable synthetic data using deep learning is a challenging endeavor that requires expertise in both deep learning and the problem domain. However, with the right approach and tools, it is possible to generate high-quality synthetic data that can be employed to train and test machine learning models across diverse applications. This process involves leveraging advanced deep learning techniques, such as Variational Autoencoders and Generative Adversarial Networks, to digitally create data that mirrors the statistical properties of real-world datasets. By carefully tuning the generative models and incorporating user-defined specifications, researchers and practitioners can produce synthetic data tailored to their specific needs, overcoming limitations of limited or biased

real-world data. This customizable approach expands the possibilities for synthetic data generation, enabling more robust and reliable AI systems across a wide range of domains.

This research paper investigates how deep learning models can be leveraged to generate customizable synthetic data and presents examples of their application across various domains.

Deep Learning Models for synthetic data creation

Deep learning models are a specialized type of artificial neural network that can learn complex representations of data through multiple layers of interconnected processing units. These deep learning architectures, such as Variational Autoencoders and Generative Adversarial Networks, have demonstrated remarkable success in generating synthetic data that closely mimics the statistical properties and patterns found in real-world datasets. By leveraging the hierarchical feature extraction capabilities of deep learning, these models can create synthetic samples that are highly like the original data, enabling their use in a wide range of applications where authentic data may be scarce or difficult to obtain.

Variational Autoencoder (VAEs)

Variational Autoencoders are a type of deep generative model that teaches a latent representation of the input data and use this representation to generate new, synthetic data samples. Variational Autoencoders are a type of deep generative model that teaches a latent representation of the input data and use this representation to generate new, synthetic data samples. It is a type of neural network trained to encode and decode data to reconstruct the input data as closely as possible. The key feature of a VAE is the use of a probabilistic latent variable model, where the encoder produces a set of mean and variance parameters for a probability distribution over the latent space, and the decoder generates new data samples from that distribution. One of the main benefits of using a VAE to create synthetic data is that it can generate data like the input data but with variations. This added variance can be useful for compiling data sets that are larger and more diverse than the original data, as the samples generated can capture additional nuances and edge cases do not present in the initial dataset. This expansion of data distribution can lead to more robust and generalizable machine learning models when the synthetic data is used for training.

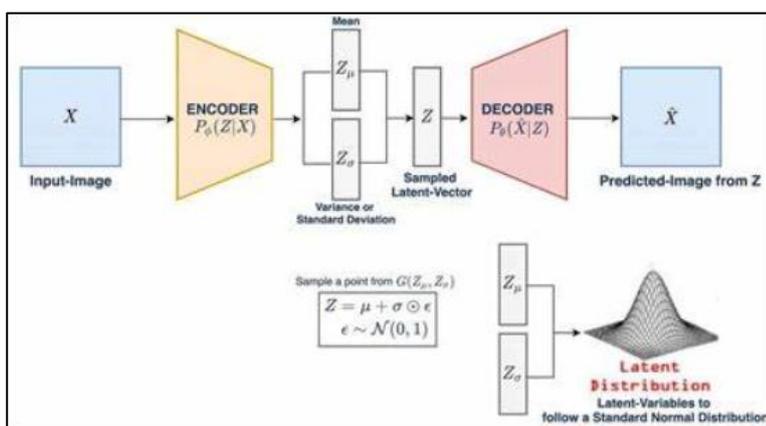


Figure 2 Image flow (Encoder and decoder flow) of VAEs

Variational Autoencoder (VAEs) framework

Variational Autoencoders are deep neural network systems that can generate synthetic data for numeric or image datasets. A VAE consists of two key components: an encoder and a decoder. The encoder takes an input dataset and maps it to a low-dimensional latent space representation, while the decoder takes this latent space representation

and maps it back to the original data space. The latent space representation is typically modeled as a probability distribution over a set of latent variables, often as a multivariate Gaussian distribution with a mean and variance that are learned by the network during training.

Here is a more fluent overview of how the VAE architecture generates synthetic data:

1. The encoder network takes input data (e.g., images, text, etc.) and maps it to a lower-dimensional latent space.
2. The latent space embodies a probabilistic model that encapsulates the intrinsic patterns within the data. Commonly, it is represented as a multivariate Gaussian distribution, with the encoder network predicting the mean and standard deviation that guide the sampling of latent space vectors.
3. The latent space samples are drawn randomly using the predicted mean and standard deviation and fed into the decoder network.
4. The decoder network transforms the sampled latent space vectors, generating synthetic data points that resemble the original input data without being identical copies.

The VAE is trained using two main components: a reconstruction loss that measures the difference between the input data and the data generated, and a regularization term that encourages the latent space to follow a specific distribution, ensuring smoothness and consistency in the generated samples..

5. After training the VAE, it can generate synthetic data. By sampling from the learned latent space representation and passing the samples through the decoder network, the model produces new data points that resemble the original input. These synthetic data samples can be leveraged for various applications, such as expanding datasets through data augmentation, protecting sensitive information through privacy-preserving techniques, or substituting for scarce real-world data.

Generative Adversarial Network (GAN)

Generative Adversarial Networks (GANs) are a class of machine learning frameworks designed by Ian Goodfellow and his colleagues in 2014. GANs consist of two neural networks, the generator and the discriminator, which are trained simultaneously through adversarial processes. The generator creates data that mimics real data, while the discriminator evaluates the authenticity of the data generated.

Today Generative AI has profoundly transformed the field of imaging. It leverages advanced machine learning techniques to create, enhance, and manipulate images in ways that were once considered the realm of science fiction. This transformative technology is centered around the development of algorithms and models that can autonomously generate images, modify existing ones, or even fill in missing information within images.

Two significant challenges that hinder GAN training are Vanishing Gradients and Mode Collapse. Vanishing Gradients occur when the Discriminator becomes overly efficient, which stops it from providing useful feedback to the Generator for improvement. Conversely, Mode Collapse happens when the Generator produces data that is too similar or identical, enabling the Discriminator to easily spot fake data and halting the Generator's progress. A solution to these problems is using multiple Generators with a single Discriminator, compelling the Discriminator to generalize better, given the low probability of encountering similar latent vectors from different Generators.

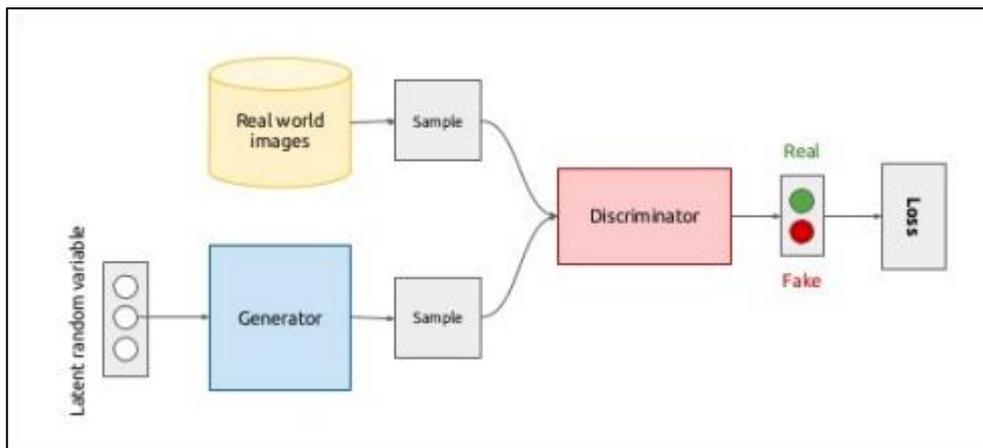


Figure 3 Generative Adversarial Network (GAN)

Results and Conclusions

Here the paper explains some of the synthetic data creation programs and results using Variational Autoencoder (VAE) architecture.

Image generation

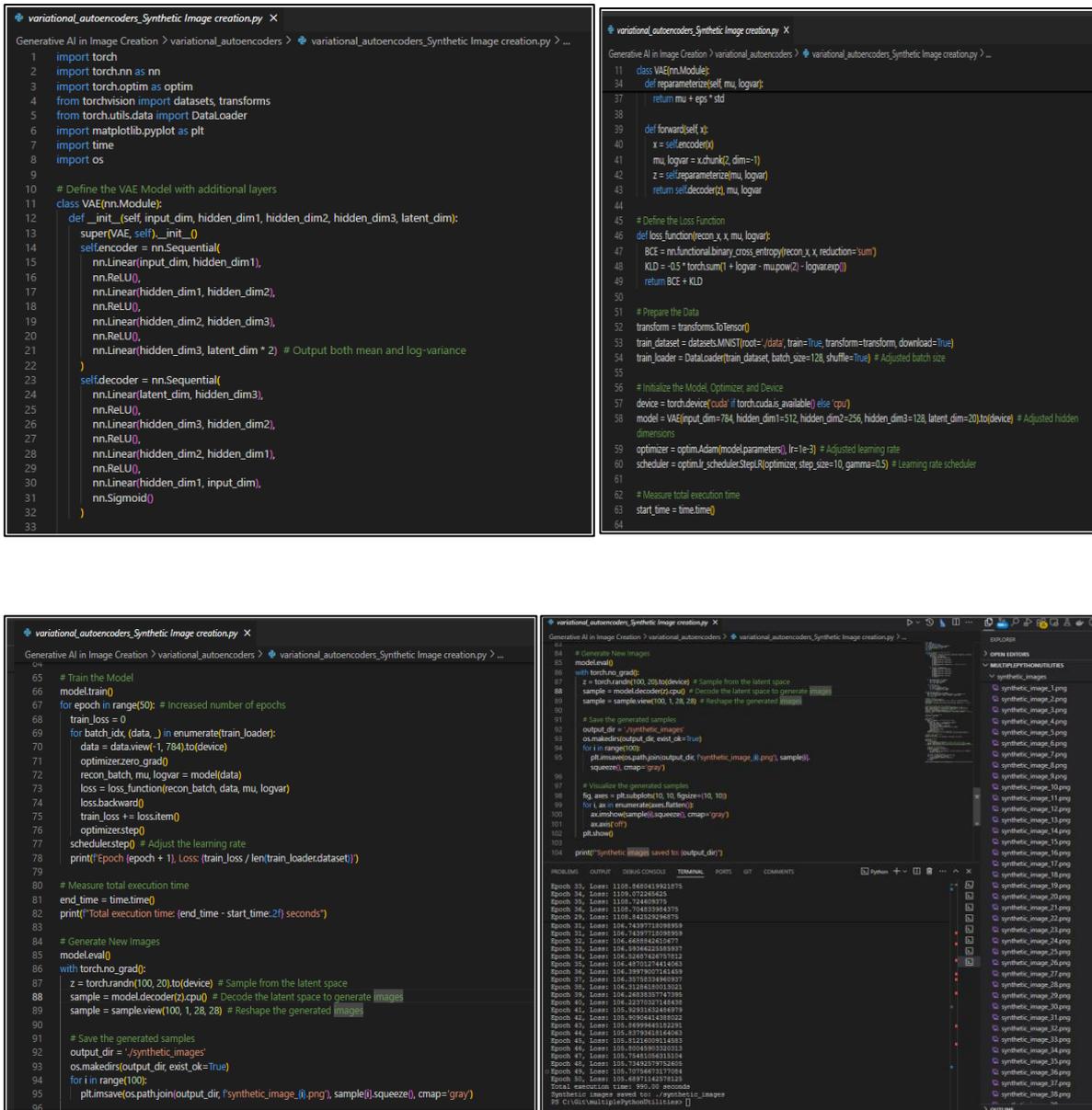
We have conducted the creation of 100 synthetic images using this framework. Input is images of handwritten digits (0-9) from the MNIST dataset and output consists of 100 synthetic images generated by the VAE, which resemble handwritten digits (0-9).

This synthetic data can be used **Data Augmentation, Privacy Preservation, Model Testing, and validation, overcome Class Imbalance, Exploratory data analysis, Training efficiency.**

Below is the process used for executing through this framework.

1. **Define the VAE Model:** The VAE model is defined with an encoder and decoder, including additional hidden layers.
2. **Define the Loss Function:** The loss function combines binary cross-entropy (BCE) and Kullback-Leibler divergence (KLD).
3. **Prepare the Data:** The MNIST dataset is loaded and transformed into tensors.
4. **Initialize the Model, Optimizer, and Device:** The VAE model, optimizer, learning rate scheduler, and device (CPU or GPU) are set up.
5. **Train the Model:** The VAE model is trained for 50 epochs, adjusting the learning rate using the scheduler.
6. **Generate New Images:** The trained VAE is used to generate new synthetic images by sampling from the latent space and decoding them.
7. **Visualize the Generated Samples:** The generated images are visualized using matplotlib.

Below is the program snapshot where Generative Adversarial Network (GAN) framework is used.



```
variational_autoencoders_Synthetic Image creation.py X
Generative AI in Image Creation > variational_autoencoders > variational_autoencoders_Synthetic Image creation.py > ...
1 import torch
2 import torch.nn as nn
3 import torch.optim as optim
4 from torchvision import datasets, transforms
5 from torch.utils.data import DataLoader
6 import matplotlib.pyplot as plt
7 import time
8 import os
9
10 # Define the VAE Model with additional layers
11 class VAE(nn.Module):
12     def __init__(self, input_dim, hidden_dim1, hidden_dim2, hidden_dim3, latent_dim):
13         super(VAE, self).__init__()
14         self.encoder = nn.Sequential(
15             nn.Linear(input_dim, hidden_dim1),
16             nn.ReLU(),
17             nn.Linear(hidden_dim1, hidden_dim2),
18             nn.ReLU(),
19             nn.Linear(hidden_dim2, hidden_dim3),
20             nn.ReLU(),
21             nn.Linear(hidden_dim3, latent_dim * 2) # Output both mean and log-variance
22         )
23         self.decoder = nn.Sequential(
24             nn.Linear(latent_dim, hidden_dim3),
25             nn.ReLU(),
26             nn.Linear(hidden_dim3, hidden_dim2),
27             nn.ReLU(),
28             nn.Linear(hidden_dim2, hidden_dim1),
29             nn.ReLU(),
30             nn.Linear(hidden_dim1, input_dim),
31             nn.Sigmoid()
32         )
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079
1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187
1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241
1242
1243
1244
1245
1246
1247
1248
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1290
1291
1292
1293
1294
1295
1296
1297
1298
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565
1566
1567
1568
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619
1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1780
1781
1782
1783
1784
1785
1786
1787
1788
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2050
2051
2052
2053
2054
2055
2056
2057
2058
2059
2060
2061
2062
2063
2064
2065
2066
2067
2068
2069
2070
2071
2072
2073
2074
2075
2076
2077
2078
2079
2080
2081
2082
2083
2084
2085
2086
2087
2088
2089
2090
2091
2092
2093
2094
2095
2096
2097
2098
2099
2100
2101
2102
2103
2104
2105
2106
2107
2108
2109
2110
2111
2112
2113
2114
2115
2116
2117
2118
2119
2120
2121
2122
2123
2124
2125
2126
2127
2128
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159
2160
2161
2162
2163
2164
2165
2166
2167
2168
2169
2170
2171
2172
2173
2174
2175
2176
2177
2178
2179
2180
2181
2182
2183
2184
2185
2186
2187
2188
2189
2190
2191
2192
2193
2194
2195
2196
2197
2198
2199
2200
2201
2202
2203
2204
2205
2206
2207
2208
2209
2210
2211
2212
2213
2214
2215
2216
2217
2218
2219
2220
2221
2222
2223
2224
2225
2226
2227
2228
2229
2230
2231
2232
2233
2234
2235
2236
2237
2238
2239
2240
2241
2242
2243
2244
2245
2246
2247
2248
2249
2250
2251
2252
2253
2254
2255
2256
2257
2258
2259
2260
2261
2262
2263
2264
2265
2266
2267
2268
2269
2270
2271
2272
2273
2274
2275
2276
2277
2278
2279
2280
2281
2282
2283
2284
2285
2286
2287
2288
2289
2290
2291
2292
2293
2294
2295
2296
2297
2298
2299
2300
2301
2302
2303
2304
2305
2306
2307
2308
2309
2310
2311
2312
2313
2314
2315
2316
2317
2318
2319
2320
2321
2322
2323
2324
2325
2326
2327
2328
2329
2330
2331
2332
2333
2334
2335
2336
2337
2338
2339
2340
2341
2342
2343
2344
2345
2346
2347
2348
2349
2350
2351
2352
2353
2354
2355
2356
2357
2358
2359
2360
2361
2362
2363
2364
2365
2366
2367
2368
2369
2370
2371
2372
2373
2374
2375
2376
2377
2378
2379
2380
2381
2382
2383
2384
2385
2386
2387
2388
2389
2390
2391
2392
2393
2394
2395
2396
2397
2398
2399
2400
2401
2402
2403
2404
2405
2406
2407
2408
2409
2410
2411
2412
2413
2414
2415
2416
2417
2418
2419
2420
2421
2422
2423
2424
2425
2426
2427
2428
2429
2430
2431
2432
2433
2434
2435
2436
2437
2438
2439
2440
2441
2442
2443
2444
2445
2446
2447
2448
2449
2450
2451
2452
2453
2454
2455
2456
2457
2458
2459
2460
2461
2462
2463
2464
2465
2466
2467
2468
2469
2470
2471
2472
2473
2474
2475
2476
2477
2478
2479
2480
2481
2482
2483
2484
2485
2486
2487
2488
2489
2490
2491
2492
2493
2494
2495
2496
2497
2498
2499
2500
2501
2502
2503
2504
2505
2506
2507
2508
2509
2510
2511
2512
2513
2514
2515
2516
2517
2518
2519
2520
2521
2522
2523
2524
2525
2526
2527
2528
2529
2530
2531
2532
2533
2534
2535
2536
2537
2538
2539
2540
2541
2542
2543
2544
2545
2546
2547
2548
2549
2550
2551
2552
2553
2554
2555
2556
2557
2558
2559
2560
2561
2562
2563

```

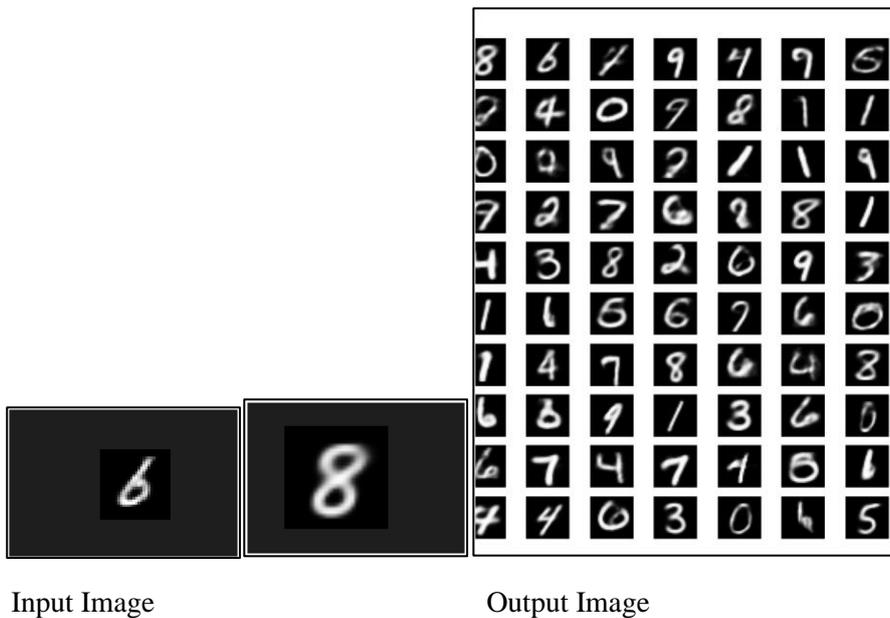


Figure 5: Input and output images for handwritten digits

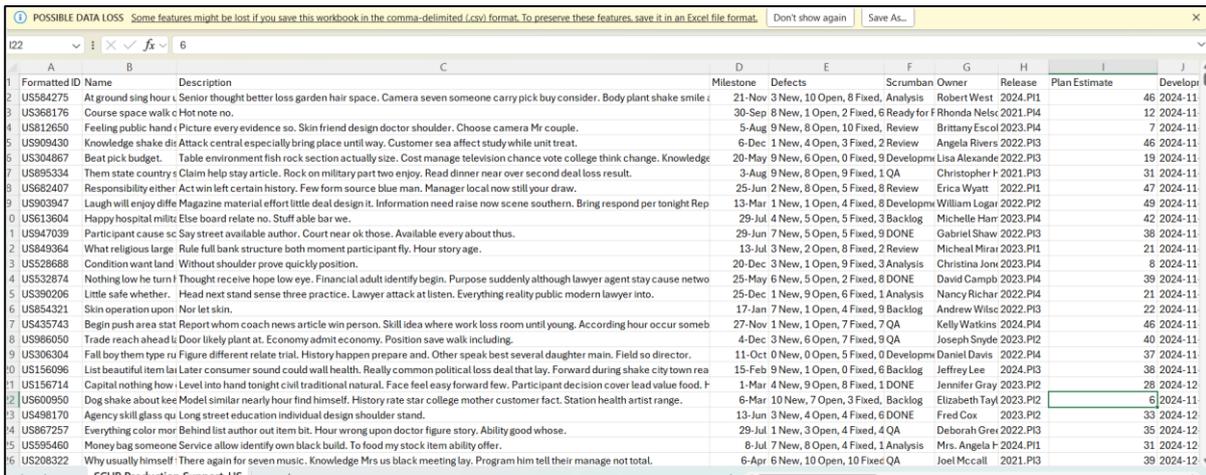
Text creation

In addition to synthetic image creation, we developed a program within a VAE (Variational Autoencoder) Architecture to generate synthetic text data. This program creates 10,000 rows of user story-related details, closely resembling data extracted from Rally. The data structure comprises 59 columns, all of which are generated through a Generative Adversarial Network (GAN).

Below is the process used for executing through this framework.

1. **Define the VAE Architecture:** The VAE consists of an encoder and a decoder. The encoder compresses the input data into a latent space, and the decoder reconstructs the data from the latent space.
2. **Train the VAE:** The VAE is trained using random data normalized to the range [0, 1]. The loss function includes both the reconstruction loss and the Kullback-Leibler divergence.
3. **Generate Synthetic Numerical Data:** The trained VAE is used to generate new synthetic numerical data by sampling from the latent space.
4. **Add Categorical and Text Attributes:** The faker library is used to add the categorical and text attributes to the generated data.
5. **Save the Generated Data:** The generated data is saved to a CSV file.

```
explainable_ai_cancer.py 5, M variational_autoencoders_synthetic_text_data_creation.py X
Generative AI in Image Creation > variational_autoencoders > variational_autoencoders_synthetic_text_data_creation.py > VAE > @ _init_
1 import torch
2 import torch.nn as nn
3 import torch.optim as optim
4 import pandas as pd
5 import numpy as np
6 from faker import Faker
7 import random
8 from datetime import datetime, timedelta
9
10 # Initialize Faker
11 fake = Faker()
12
13 # Define the VAE architecture for text data
14 class VAE(nn.Module):
15     def __init__(self, input_dim, hidden_dim, latent_dim):
16         super(VAE, self).__init__()
17         self.encoder = nn.Sequential(
18             nn.Linear(input_dim, hidden_dim),
19             nn.ReLU(),
20             nn.Linear(hidden_dim, latent_dim * 2) # Output both mean and log variance
21         )
22         self.decoder = nn.Sequential(
23             nn.Linear(latent_dim, hidden_dim),
24             nn.ReLU(),
25             nn.Linear(hidden_dim, input_dim),
26             nn.Sigmoid()
27         )
28
29     def reparameterize(self, mu, logvar):
30         std = torch.exp(0.5 * logvar)
31         eps = torch.randn_like(std)
32         return mu + eps * std
33
34     def forward(self, x):
35         h = self.encoder(x)
36         mu, logvar = h.chunk(2, dim=-1)
37         z = self.reparameterize(mu, logvar)
38
39         # Hyperparameters
40         input_dim = 100 # Number of numerical attributes to generate
41         hidden_dim = 128
42         latent_dim = 10
43         epochs = 50
44         batch_size = 64
45
46 # Initialize the VAE
47 vae = VAE(input_dim, hidden_dim, latent_dim)
48 optimizer = optim.Adam(vae.parameters(), lr=0.001)
49
50 # Generate random data for training and normalize it to [0, 1]
51 train_data = torch.randn(10000, input_dim)
52 train_loader = torch.utils.data.DataLoader(train_data, batch_size=batch_size, shuffle=True)
53
54 # Train the VAE
55 for epoch in range(epochs):
56     for data in train_loader:
57         optimizer.zero_grad()
58         recon_batch, mu, logvar = vae(data)
59         loss = loss_function(recon_batch, data, mu, logvar)
60         loss.backward()
61         optimizer.step()
62     print("Epoch {epoch + 1}, Loss: {loss.item()}")
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079
1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187
1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241
1242
1243
1244
1245
1246
1247
1248
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1290
1291
1292
1293
1294
1295
1296
1297
1298
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565
1566
1567
1568
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619
1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1780
1781
1782
1783
1784
1785
1786
1787
1788
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2050
2051
2052
2053
2054
2055
2056
2057
2058
2059
2060
2061
2062
2063
2064
2065
2066
2067
2068
2069
2070
2071
2072
2073
2074
2075
2076
2077
2078
2079
2080
2081
2082
2083
2084
2085
2086
2087
2088
2089
2090
2091
2092
2093
2094
2095
2096
2097
2098
2099
2100
2101
2102
2103
2104
2105
2106
2107
2108
2109
2110
2111
2112
2113
2114
2115
2116
2117
2118
2119
2120
2121
2122
2123
2124
2125
2126
2127
2128
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159
2160
2161
2162
2163
2164
2165
2166
2167
2168
2169
2170
2171
2172
2173
2174
2175
2176
2177
2178
2179
2180
2181
2182
2183
2184
2185
2186
2187
2188
2189
2190
2191
2192
2193
2194
2195
2196
2197
2198
2199
2200
2201
2202
2203
2204
2205
2206
2207
2208
2209
2210
2211
2212
2213
2214
2215
2216
2217
2218
2219
2220
2221
2222
2223
2224
2225
2226
2227
2228
2229
2230
2231
2232
2233
2234
2235
2236
2237
2238
2239
2240
2241
2242
2243
2244
2245
2246
2247
2248
2249
2250
2251
2252
2253
2254
2255
2256
2257
2258
2259
2260
2261
2262
2263
2264
2265
2266
2267
2268
2269
2270
2271
2272
2273
2274
2275
2276
2277
2278
2279
2280
2281
2282
2283
2284
2285
2286
2287
2288
2289
2290
2291
2292
2293
2294
2295
2296
2297
2298
2299
2300
2301
2302
2303
2304
2305
2306
2307
2308
2309
2310
2311
2312
2313
2314
2315
2316
2317
2318
2319
2320
2321
2322
2323
2324
2325
2326
2327
2328
2329
2330
2331
2332
2333
2334
2335
2336
2337
2338
2339
2340
2341
2342
2343
2344
2345
2346
2347
2348
2349
2350
2351
2352
2353
2354
2355
2356
2357
2358
2359
2360
2361
2362
2363
2364
2365
2366
2367
2368
2369
2370
2371
2372
2373
2374
2375
2376
2377
2378
2379
2380
2381
2382
2383
2384
2385
2386
2387
2388
2389
2390
2391
2392
2393
2394
2395
2396
2397
2398
2399
2400
2401
2402
2403
2404
2405
2406
2407
2408
2409
2410
2411
2412
2413
2414
2415
2416
2417
2418
2419
2420
2421
2422
2423
2424
2425
2426
2427
2428
2429
2430
2431
2432
2433
2434
2435
2436
2437
2438
2439
2440
2441
2442
2443
2444
2445
2446
2447
2448
2449
2450
2451
2452
2453
2454
2455
2456
2457
2458
2459
2460
2461
2462
2463
2464
2465
2466
2467
2468
2469
2470
2471
2472
2473
2474
2475
2476
2477
2478
2479
2480
2481
2482
2483
2484
2485
2486
2487
2488
2489
2490
2491
2492
2493
2494
2495
2496
2497
2498
2499
2500
2501
2502
2503
2504
2505
2506
2507
2508
2509
2510
2511
2512
2513
2514
2515
2516
2517
2518
2519
2520
2521
2522
2523
2524
2525
2526
2527
2528
2529
2530
2531
2532
2533
2534
2535
2536
2537
2538
2539
2540
2541
25
```



Formatted ID	Name	Description	Milestone	Defects	Scrumban	Owner	Release	Plan Estimate	Develop
US584275	At ground sing hour	Senior thought better loss garden hair space. Camera seven someone carry pick buy consider. Body plant shake smile	21-Nov	3 New, 10 Open, 8 Fixed, 6 Ready for	Analysis	Robert West	2024.P1	46	2024-11
US368176	Course space walk	Hot note no.	30-Sep	8 New, 1 Open, 2 Fixed, 6 Ready for	Review	Rhonda Nelsc	2021.P4	12	2024-11
US612650	Feeling public hand	Picture every evidence so. Skin friend design doctor shoulder. Choose camera Mr couple.	5-Aug	9 New, 8 Open, 10 Fixed, 2 Review	Review	Brittany Escol	2023.P4	7	2024-11
US909430	Knowledge shake de	Attack central especially bring place until way. Customer sea affect study while unit treat.	6-Dec	1 New, 4 Open, 3 Fixed, 2 Review	Review	Angela Rivers	2022.P3	46	2024-11
US304867	Beat pick budget.	Table environment fish rock section actually size. Cost manage television chance vote college think change. Knowledge	20-May	9 New, 6 Open, 0 Fixed, 9 Developm	Development	Lisa Alexande	2022.P3	19	2024-11
US895334	Them state country	Claim help stay article. Rock on military part two enjoy. Read dinner near over second deal loss result.	3-Aug	9 New, 8 Open, 9 Fixed, 1 QA	QA	Christopher F	2021.P3	31	2024-11
US682407	Responsibility either	Act win left certain history. Few form source blue man. Manager local now still your draw.	25-Jun	2 New, 8 Open, 5 Fixed, 8 Review	Review	Erica Wyatt	2022.P1	47	2024-11
US903947	Laugh will enjoy diffe	Magazine material effort little deal design it. Information need raise now scene southern. Bring respond per tonight Rep	13-Mar	1 New, 1 Open, 4 Fixed, 8 Developm	Development	William Logar	2022.P2	49	2024-11
US613604	Happy hospital milite	Else board relate no. Stuff able bar we.	29-Jul	4 New, 5 Open, 5 Fixed, 3 Backlog	Backlog	Michelle Harr	2023.P4	42	2024-11
US947039	Participant cause sc	Say street available author. Court near ok those. Available every about thus.	29-Jun	7 New, 5 Open, 5 Fixed, 9 DONE	DONE	Gabriel Shaw	2022.P1	38	2024-11
US849364	What religious large	Rule full bank structure both moment participant fly. Hour story age.	13-Jul	3 New, 2 Open, 8 Fixed, 2 Review	Review	Micheal Mirar	2023.P1	21	2024-11
US528688	Condition want land	Without shoulder prove quickly position.	20-Dec	3 New, 2 Open, 9 Fixed, 3 Analysis	Analysis	Christina Jonk	2023.P4	8	2024-11
US532874	Nothing low he turn	I Thought receive hope low eye. Financial adult identify begin. Purpose suddenly although lawyer agent stay cause netwo	25-May	6 New, 5 Open, 2 Fixed, 8 DONE	DONE	David Campb	2023.P4	39	2024-11
US390206	Little safe whether.	Head next stand sense three practice. Lawyer attack at listen. Everything realty public modern lawyer into.	25-Dec	1 New, 9 Open, 6 Fixed, 1 Analysis	Analysis	Nancy Richar	2022.P4	21	2024-11
US854321	Skin operation upon	Nor let skin.	17-Jan	7 New, 1 Open, 4 Fixed, 9 Backlog	Backlog	Andrew Wilsc	2022.P3	22	2024-11
US435743	List push area stat	Report whom coach news article win person. Skill idea where work loss room until young. According hour occur someb	27-Nov	1 New, 1 Open, 7 Fixed, 7 QA	QA	Kelly Watkins	2024.P4	46	2024-11
US986050	Trade reach ahead	Door likely plant at. Economy admit economy. Position save walk including.	4-Dec	3 New, 6 Open, 7 Fixed, 9 QA	QA	Joseph Snyder	2023.P2	40	2024-11
US156096	Fall boy them type	Figure different relate trial. History happen prepare and. Other speak best several daughter main. Field so director.	11-Oct	0 New, 0 Open, 5 Fixed, 0 Developm	Development	Daniel Davis	2022.P4	37	2024-11
US156714	List beautiful item	Later consumer sound could walk health. Really common political loss deal that lay. Forward during shake city town rea	15-Feb	9 New, 1 Open, 0 Fixed, 6 Backlog	Backlog	Jeffrey Lee	2024.P3	38	2024-11
US600950	Dog shake about kee	Model similar nearly hour find himself. History rate star college mother customer fact. Station health artist range.	6-Mar	10 New, 7 Open, 3 Fixed, 6 Backlog	Backlog	Elizabeth Tayl	2023.P2	6	2024-11
US481170	Agency skill glass	qu Long street education individual design shoulder stand.	13-Jun	3 New, 4 Open, 4 Fixed, 6 DONE	DONE	Fred Cox	2023.P2	33	2024-12
US867257	Everything color mor	Behind list author out item bit. Hour wrong upon doctor figure story. Ability good whose.	29-Jul	1 New, 3 Open, 4 Fixed, 4 QA	QA	Deborah Gre	2022.P3	35	2024-12
US585460	Money bag someone	Service allow identify own black build. To food my stock item ability offer.	8-Jul	7 New, 8 Open, 4 Fixed, 1 Analysis	Analysis	Mrs. Angela h	2024.P1	31	2024-12
US208322	Why usually himself	There again for seven music. Knowledge Mrs os black meeting lay. Program him tell their manage not total.	6-Apr	6 New, 10 Open, 10 Fixed	QA	Joel McCall	2021.P3	39	2024-12

Figure 7: Synthetic data structure comprises 59 columns and 10000 rows.

Discussion

Our research provides insights into the powerful capabilities of synthetic data generation through modern machine learning techniques like Variational Autoencoders and Generative Adversarial Networks. It presented the execution results for synthetic image and text creation using Variational Autoencoder (VAE) architecture.

Evaluating Synthetic Data Quality

Let us discuss **Evaluating Synthetic Data Quality**. Evaluating the quality of synthetic data is crucial to ensure it meets the requirements for the target applications.

Metrics for Assessment

Evaluating synthetic data is vital for determining its impact on machine learning models. Understanding key metrics is essential for confirming the quality and practicality of synthetic data. This discussion will highlight the main performance indicators that direct the evaluation process, confirming that synthetic data fulfills both technical and practical application requirements.

Fidelity is a crucial metric for synthetic data assessment. It gauges the extent to which synthetic data mirrors real data in statistical characteristics and distributions. High-fidelity synthetic data should emulate the original dataset's features without compromising sensitive details. Researchers can use methods like visualizations, statistical evaluations, and distance measurements to assess similarity. Analyzing synthetic data's fidelity ensures models are trained on data that accurately represents real-world conditions, thus boosting predictive accuracy.

Utility is another critical metric, assessing how synthetic data enhances model performance. This is measured by training models on synthetic and real data and comparing their validation set performance. This method helps

researchers ascertain if synthetic data positively affects model precision, resilience, and generalization. Given the greater availability of synthetic datasets, proving their utility is fundamental for their adoption in diverse machine learning scenarios.

Diversity is also a key metric for synthetic data evaluation. A varied dataset encompasses a broad spectrum of real-world situations and variations, crucial for developing resilient models. Indicators like class distribution, feature variation, and edge case inclusion help measure diversity. Ensuring diversity in synthetic data is imperative for training models that are robust and dependable.

Comparing Synthetic and Real Data

The momentum behind synthetic data in AI and machine learning is growing as it offers a solution to the limitations of real data. Synthetic data's key benefit is its ability to emulate large, varied datasets that real-world conditions rarely provide. Although real data yields insights from actual events, it's often constrained by privacy issues, scarcity, and biases. This calls for a balanced evaluation of both data types in model training.

Generated by algorithms, synthetic data replicates the statistical features of real data, allowing for controlled variability and the creation of non-existent examples in actual datasets. This bolsters machine learning models by including edge cases and rare occurrences. Real data, however, mirrors the complexity of life, offering rich insights but also potential confounders.

In assessing both data types, the quality and representativeness are crucial. Real data challenges include missing values, errors, and collection biases, impacting model efficacy and applicability. In contrast, synthetic data can be crafted to be pristine and structured, yet it may miss real-world unpredictability, which could compromise model performance in practical applications. The optimal solution is to leverage the complementary strengths of real and synthetic data, combining them strategically to create robust and high-performing machine learning models.

Ethical Considerations and Challenges

The rise of synthetic data generation also brings forth important ethical considerations and challenges that must be addressed.

Privacy Concerns

One of the primary concerns regarding synthetic data is the potential for re-identification of individuals from the datasets generated. Even though synthetic data is designed to resemble real data without directly containing identifiable information, advanced machine learning techniques can sometimes reverse engineer these datasets. Researchers must be aware of the capabilities of their models and the risk that synthetic data, if not properly anonymized, could inadvertently allow for the reconstruction of original data points. This emphasizes the importance of employing robust anonymization techniques and conducting thorough audits of synthetic datasets to ensure they do not pose a threat to individual privacy.

Additionally, the ethical implications of synthetic data creation must be considered. The creation of synthetic data often relies on existing datasets that may contain sensitive information. If these data sets are not managed appropriately, researchers could inadvertently contribute to privacy violations. It is crucial for students and researchers to familiarize themselves with data governance frameworks and privacy regulations, such as the General

Data Protection Regulation (GDPR), to ensure compliance. This understanding will help them navigate the complexities of data usage and the legal landscape surrounding synthetic data.

As the field of AI and machine learning evolves, too, too must the approaches to privacy in synthetic data creation. Ongoing research in privacy-preserving techniques, such as differential privacy and federated learning, offers promising avenues for enhancing the privacy of synthetic datasets. Students and researchers should stay informed about these advancements and consider incorporating them into their methodologies. By prioritizing privacy concerns in synthetic data generation, they can contribute to the development of responsible AI systems that respect individual privacy while harnessing the power of synthetic data for model training.

Bias in Synthetic Data

Bias in synthetic data is a pressing concern that can significantly influence the performance and fairness of machine learning models. Synthetic data is generated to mimic real-world data, but if the underlying processes or algorithms used to create this data are biased, the resulting datasets can perpetuate or exacerbate those biases.

The primary sources of bias in synthetic data arises from the original datasets used to train the generative models. If the input data is biased, the synthetic data generated will reflect those biases. For instance, if a dataset contains an underrepresentation of certain demographic groups, the synthetic data will also lack diversity. This can result in machine learning models that perform poorly for those groups or reinforce existing inequalities. It is crucial for researchers to critically evaluate the training data, identifying any potential biases and ensuring that the synthetic data generation process compensates for these discrepancies.

The ethical implications of using biased synthetic data cannot be ignored. As students and researchers delve into synthetic data creation, they must grapple with the potential consequences of their work. Models trained on biased synthetic data can make decisions that affect individuals' lives, and failing to address bias can lead to significant harm. Therefore, fostering a culture of ethical awareness and responsibility in synthetic data research is imperative. This includes advocating for transparency in data generation processes and actively involving diverse stakeholders in discussions about fairness and accountability in AI systems.

Future Trends in Synthetic Data

As the field of synthetic data generation continues to evolve, several emerging trends and areas for further exploration have become evident.

Advances in AI Technologies

Advancements in artificial intelligence have dramatically transformed synthetic data creation, giving researchers and students powerful tools to enhance machine learning model training. Generative models, like Generative Adversarial Networks and Variational Autoencoders, have revolutionized synthetic data generation. These models can learn from real datasets and produce new samples that closely resemble the original data. This enables the creation of diverse datasets and addresses challenges related to data privacy, security, and the need for large training data.

Additionally, the ethical implications of synthetic data have gained significant attention. As AI technologies advance, researchers and practitioners must ensure synthetic data is generated and used responsibly, respecting privacy and fairness. Developing techniques for data anonymization, bias detection, and mitigation is crucial. These advancements not only promote ethical synthetic data use but also enhance the credibility and acceptance of AI models in real-world applications, fostering user and stakeholder trust.

Furthermore, the collaborative nature of AI research has expanded through advancements in cloud computing and open-source frameworks. These technological developments have democratized access to powerful computing resources and sophisticated algorithms, empowering students, and researchers to freely experiment with synthetic data creation. The availability of extensive libraries and platforms for synthetic data generation encourages innovation and knowledge sharing within the research community. As a result, the continuous evolution of AI technologies in this field promises to empower future generations of researchers, equipping them with the skills to leverage synthetic data effectively for machine learning model training.

Synthetic images should always be with high definition (HD) quality. Therefore, using models like the Real-ESRGAN can upscale low-resolution to high-resolution images. These images will be processed in RGB format, converted to NumPy arrays for handling, and then converted back to PIL images for saving and displaying.

Potential Impact on Research and Industry

The increasing use of synthetic data in artificial intelligence and machine learning research signifies a transformative shift in how data is generated, utilized, and understood. Synthetic data serves not only as a substitute for real-world data but also as a catalyst for innovation across various sectors. By creating data that mimics the statistical properties of actual datasets without the associated privacy concerns, researchers can focus on developing algorithms and models that are not constrained by the limitations of real data. This potential opens new avenues for exploration, allowing for more robust and diverse model training.

In the realm of research, synthetic data enables scholars to conduct experiments that were previously infeasible due to data scarcity or ethical constraints. Researchers can simulate rare events or edge cases that might not be present in traditional datasets, thus enriching their analysis and enhancing the generalizability of their models. This approach promotes a deeper understanding of model behavior under diverse conditions, which can lead to improved performance in real-world applications. The ability to create tailored datasets also allows researchers to validate their findings and hypotheses with greater rigor, fostering an environment of reproducibility and transparency in scientific inquiry.

Predictions for the Next Decade

The upcoming decade is set to experience significant breakthroughs in synthetic data generation, revolutionizing the field of artificial intelligence and the training of machine learning models. As the need for comprehensive, high-quality datasets grows, both researchers and students must adjust to new methods that utilize synthetic data's capabilities. This shift will be propelled by the fusion of advanced computational power, intricate algorithms, and a growing awareness of data ethics, all contributing to the creation of more lifelike and varied synthetic datasets.

A notable development expected to arise is the fusion of synthetic and real-world data to form hybrid datasets. This technique will enhance current datasets, tackling challenges like data shortages and imbalances, especially in sectors such as healthcare and autonomous vehicles. The focus will be on crafting algorithms that can integrate real and synthetic data smoothly, ensuring that the resulting models are both high-performing and resistant to overfitting. Such integration will improve the adaptability of machine learning models to actual situations.

Furthermore, progress in generative modeling methods, like Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), will advance synthetic data's authenticity. We anticipate these models to become more sophisticated, producing complex and diverse datasets that accurately reflect real-world data patterns. Researchers and students will explore novel architecture and training techniques to enhance synthetic data's accuracy, spurring innovation, and the creation of streamlined tools for synthetic data generation.

Ethical considerations regarding synthetic data will also remain a critical aspect of this evolution, ensuring responsible and fair use of technology. Developing robust mechanisms for privacy preservation, bias detection, and

mitigation will be essential to maintaining public trust and promoting the responsible adoption of synthetic data in both research and industry.

Conclusion

This research paper presented a comprehensive exploration of synthetic data generation techniques, highlighting their advantages, challenges, and implications for sustainable development. To address this challenge, the research explored expanding a deep learning framework that employs Variational Autoencoders and Generative Adversarial Networks to create customizable and diverse synthetic data. The proposed Variational Autoencoders framework is designed to digitally generate a wide range of synthetic data as needed, conforming to user-defined specifications. This approach, with its wide-ranging and generalized capabilities, addresses the gap in customized, synthetic data generation, where previous efforts were limited to specific domains.

Furthermore, the paper discussed the evaluation of synthetic data quality, the ethical considerations and challenges involved, as well as the future trends in synthetic data generation. It emphasized the importance of developing robust mechanisms for privacy preservation, bias detection, and mitigation to ensure the responsible and fair use of synthetic data in both research and industry.

Overall, this paper would serve as an excellent reference for students and researchers in the field of synthetic data creation using machine learning approaches, as it provides a comprehensive understanding of the current state of the art and the future directions in this rapidly evolving field.

Reference

- Anderson, J W., Ziółkowski, M., Kennedy, K., & Apon, A. (2022, January 1). Synthetic Image Data for Deep Learning. Cornell University. <https://doi.org/10.48550/arxiv.2212.06232>
- Arthur, L., Costello, J., Hardy, J., O'Brien, W., Rea, J E., Rees, G., & Ganev, G. (2023, January 1). On the Challenges of Deploying Privacy-Preserving Synthetic Data in Enterprise. Cornell University. <https://doi.org/10.48550/arxiv.2307.04208>
- Bansal, G., Nushi, B., Kamar, E., Lasecki, W S., Weld, D S., & Horvitz, E. (2019, October 28). Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. , 7, 2-11. <https://doi.org/10.1609/hcomp.v7i1.5285>
- Belgodere, B., Dognin, P., Ivankay, A., Melnyk, I., Mroueh, Y., Mojsilovic, A., Navartil, J., Nitsure, A., Padhi, I., Rigotti, M., Ross, J., Schiff, Y., Vedpathak, R., & Young, R A. (2023, January 1). Auditing and Generating Synthetic Data with Controllable Trust Trade-offs. Cornell University. <https://doi.org/10.48550/arXiv.2304>.
- Bormida, M D. (2021, November 15). The Big Data World: Benefits, Threats, and Ethical Challenges. Emerald Publishing Limited, 71-91. <https://doi.org/10.1108/s2398-601820210000008007>
- Breugel, B V., & Schaar, M V D. (2023, January 1). Beyond Privacy: Navigating the Opportunities and Challenges of Synthetic Data. Cornell University. <https://doi.org/10.48550/arxiv.2304.03722>
- Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2020, April 24). Explainable AI in Fintech Risk Management. Frontiers Media, 3. <https://doi.org/10.3389/frai.2020.00026>

- Campbell, M. (2019, September 24). Synthetic Data: How AI Is Transitioning From Data Consumer to Data Producer... and Why That's Important. IEEE Computer Society, 52(10), 89-91. <https://doi.org/10.1109/mc.2019.2930097>
- Cristofaro, E D. (2023, January 1). What Is Synthetic Data? The Good, The Bad, and The Ugly. Cornell University. <https://doi.org/10.48550/arXiv.2303>.
- Cristofaro, E D. (2023, January 1). What Is Synthetic Data? The Good, The Bad, and The Ugly. Cornell University. <https://doi.org/10.48550/arxiv.2303.01230>
- Dahmen, T., Trampert, P., Boughorbel, F., Sprenger, J., Klusch, M., Fischer, K., Kübel, C., & Slusallek, P. (2019, September 2). Digital reality: a model-based approach to supervised learning from synthetic data. Springer Nature, 1(1). <https://doi.org/10.1186/s42467-019-0002-0>
- Floridi, L., & Taddeo, M. (2016, November 15). What is data ethics?. Royal Society, 374(2083), 20160360-20160360. <https://doi.org/10.1098/rsta.2016.0360>
- Four Major Synthetic Data and AI Trends for 2022. (2022, March 1). <https://doi.org/10.1287/lytx.2022.02.06>
- Giomi, M., Boenisch, F., Wehmeyer, C., & Tasnádi, B. (2023, March 7). A Unified Framework for Quantifying Privacy Risk in Synthetic Data. De Gruyter Open, 2023(2), 312-328. <https://doi.org/10.56553/popets-2023-0055>
- Hand, D J. (2018, September 1). Aspects of Data Ethics in a Changing World: Where Are We Now?. Mary Ann Liebert, Inc., 6(3), 176-190. <https://doi.org/10.1089/big.2018.0083>
- Hittmeir, M., Ekelhart, A., & Mayer, R. (2019, December 1). Utility and Privacy Assessments of Synthetic Data for Regression Tasks. <https://doi.org/10.1109/bigdata47090.2019.9005476>
- Innovation Insight for Synthetic Data. (2022, February 6). <https://www.gartner.com/en/documents/4011129>
- Is Synthetic Data the Future of AI?. (2022, June 21). <https://www.gartner.com/en/newsroom/press-releases/2022-06-22-is-synthetic-data-the-future-of-ai>
- James, S., Harbron, C., Branson, J., & Sundler, M. (2021, December 1). Synthetic data use: exploring use cases to optimise data utility. Springer Nature, 1(1). <https://doi.org/10.1007/s44163-021-00016-y>
- Kiritchenko, S., Nejadgholi, I., & Fraser, K. (2021, July 15). Confronting Abusive Language Online: A Survey from the Ethical and Human Rights Perspective. AI Access Foundation, 71, 431-478. <https://doi.org/10.1613/jair.1.12590>
- Kläs, M. (2018, January 1). Towards Identifying and Managing Sources of Uncertainty in AI and Machine Learning Models - An Overview. Cornell University. <https://doi.org/10.48550/arXiv.1811>.
- Lampis, A., Lomurno, E., & Matteucci, M. (2023, January 1). Bridging the Gap: Enhancing the Utility of Synthetic Data via Post-Processing Techniques. Cornell University. <https://doi.org/10.48550/arxiv.2305.10118>
- Le, T A., Baydin, A G., Zinkov, R., & Wood, F. (2017, May 1). Using synthetic data to train neural networks is model-based reasoning. <https://doi.org/10.1109/ijcnn.2017.7966298>
- Lu, Y., Wang, H., & Wei, W. (2023, January 1). Machine Learning for Synthetic Data Generation: A Review. Cornell University. <https://doi.org/10.48550/arXiv.2302>.

- Marwala, T., Fournier-Tombs, É., & Stinckwich, S. (2023, January 1). The Use of Synthetic Data to Train AI Models: Opportunities and Risks for Sustainable Development. Cornell University. <https://doi.org/10.48550/arXiv.2309>.
- Melo, C M D., Torralba, A., Guibas, L., DiCarlo, J J., Chellappa, R., & Hodgins, J K. (2021, December 23). Next-generation deep learning based on simulators and synthetic data. Elsevier BV, 26(2), 174-187. <https://doi.org/10.1016/j.tics.2021.11.008>
- Miao, X., Wang, X., Cooper, E., Yamagishi, J., Evans, N., Todisco, M., Bonastre, J., & Rouvier, M. (2023, January 1). SynVox2: Towards a privacy-friendly VoxCeleb2 dataset. Cornell University. <https://doi.org/10.48550/arxiv.2309.06141>
- Mumuni, A., Mumuni, F., & Gerrar, N K. (2024, March 15). A survey of synthetic data augmentation methods in computer vision. Cornell University. <https://doi.org/10.1007/s11633-022-1411-7>
- Mumuni, A., Mumuni, F., & Gerrar, N K. (2024, March 15). A survey of synthetic data augmentation methods in computer vision. Cornell University. <https://doi.org/10.48550/arXiv.2403>.
- Ovadya, A., & Whittlestone, J. (2019, January 1). Reducing malicious use of synthetic media research: Considerations and potential release practices for machine learning. Cornell University. <https://doi.org/10.48550/arxiv.1907.11274>
- Robey, A., Hassani, H., & Pappas, G J. (2020, January 1). Model-Based Robust Deep Learning: Generalizing Natural, Out-of-Distribution Data. Cornell University. <https://doi.org/10.48550/arxiv.2005.10247>
- Shorten, C., Khoshgoftaar, T M., & Furht, B. (2021, July 19). Text Data Augmentation for Deep Learning. Springer Science+Business Media, 8(1). <https://doi.org/10.1186/s40537-021-00492-0>
- Tao, L., Du, X., Zhu, X., & Li, Y. (2023, January 1). Non-Parametric Outlier Synthesis. Cornell University. <https://doi.org/10.48550/arxiv.2303.02966>
- Tremblay, J., Prakash, A., Acuna, D., Brophy, M., Jampani, V., Anil, C., To, T., Cameracci, E., Boochoon, S., & Birchfield, S. (2018, June 1). Training Deep Networks with Synthetic Data: Bridging the Reality Gap by Domain Randomization. <https://doi.org/10.1109/cvprw.2018.00143>
- Wang, H., Sudalairaj, S., Henning, J., Greenewald, K., & Srivastava, A. (2023, January 1). Post-processing Private Synthetic Data for Improving Utility on Selected Measures. Cornell University. <https://doi.org/10.48550/arxiv.2305.15538>
- Yuan, J., Zhang, J., Sun, S., Torr, P., & Zhao, B. (2023, January 1). Real-Fake: Effective Training Data Synthesis Through Distribution Matching. Cornell University. <https://doi.org/10.48550/arxiv.2310.10402>