Identifying Outliers in Data Using Variational Autoencoders

April 6, 2025

Anurag Sarva (CS24MTECH14003)
Gulshan Hatzade (CS24MTECH14006)
Mainak Adhikari (CS24MTECH14018)

Abstract

This report presents an approach to identify outliers in high-dimensional data using Variational Autoencoders (VAEs) combined with K-means clustering. VAEs compress the data into a lower-dimensional latent space where patterns are easier to identify. K-means clustering is then applied to group similar data points, allowing for the identification of outliers as boundary points that deviate significantly from their cluster centers.

1 Introduction

Outlier detection is essential for identifying anomalies, detecting fraud, and finding unusual patterns in data. Traditional methods often struggle with high-dimensional data. Our approach uses VAEs to reduce dimensionality while preserving the underlying data structure, followed by clustering to identify anomalous points.

2 Methodology

2.1 Data Preprocessing

- The dataset contained 10 features and 1199 data points
- Data was normalized using MinMaxScaler to ensure all features fell within the 0-1 range
- No missing values were found in the dataset

2.2 Variational Autoencoder (VAE) Architecture

- Encoder: Input layer (10 features) $\rightarrow$ 32 neurons $\rightarrow$ 16 neurons $\rightarrow$ latent space (size determined experimentally)
- Two parallel outputs for mean and log variance (for the variational aspect)

- Decoder: Latent space → 16 neurons → 32 neurons → 10 features (output)
- ReLU activations used throughout, with Sigmoid for the final layer
- Trained using Adam optimizer with learning rate of 1e-3

2.3 Optimal Compression Size Determination

- Tested compression dimensions from 2 to 10
- Evaluated using reconstruction loss (MSE + KL divergence)
- Dimension 3 found to be optimal with lowest reconstruction error (0.032306)

2.4 Clustering in Latent Space

- Applied K-means clustering to the compressed data points
- Used the elbow method to determine the optimal number of clusters
- Found 3 as the optimal cluster count
- Cluster sizes: 636, 326, and 237 points

2.5 Outlier Detection Strategy

- Boundary outliers: Points beyond the 90th percentile distance from their cluster center
- Small cluster outliers: Points in clusters smaller than 10% of the average cluster size
- 121 total outliers were identified, all categorized as boundary outliers

3 Results and Analysis

3.1 Compression Performance

- Optimal latent space dimension: 3
- Reconstruction loss by dimension:
  - Dim 2: 0.037541
  - Dim 3: 0.032306 (best)
  - Dim 4: 0.033129
  - Dim 5: 0.035284

3.2 Cluster Analysis

- 3 distinct clusters identified
- Cluster sizes: 636, 326, and 237
- No small clusters were found (less than 10% of average size)

3.3 Outlier Analysis

- 121 boundary outliers detected (points far from their cluster centers)
- 0 small cluster outliers detected
- PCA was used to visualize the 3D compressed space in 2D for better interpretation

3.4 Visualization

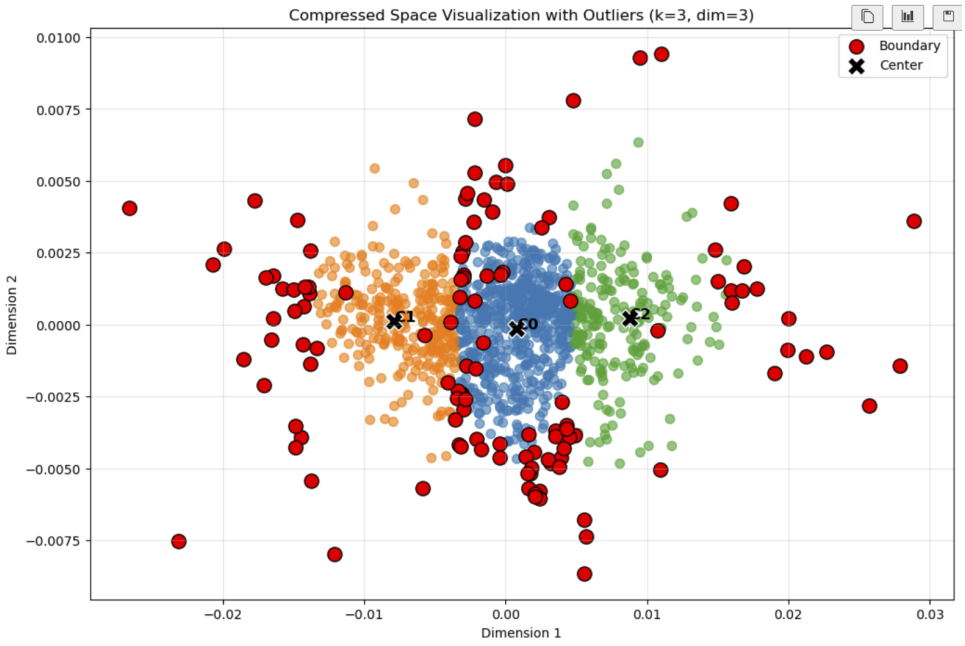

Figure 1: Visualization of outliers

- 2D visualization of compressed space shows clear cluster boundaries
- Outliers marked in red, appear at the periphery of their respective clusters
- Cluster centers shown as black X markers

Index	Group	Outlier_Type	Feature_1	Feature_2	Feature_3	Feature_4	Feature_5	Feature_6	Feature_7	Feature_8	Feature_9	Feature_10
256	0	Boundary	1.000000	0.970786	−0.081369	−0.072967	0.000000	0.573944	−0.478665	−0.032600	−1.023559	−0.043989
258	0	Boundary	0.997425	0.756470	0.911067	0.606161	0.495914	0.451248	0.889136	−0.031071	1.203017	−0.054227
261	0	Boundary	0.985373	0.981808	−0.549108	−0.458748	0.522045	0.719086	−0.132245	−0.032735	−0.497513	−0.053758
1031	0	Boundary	0.999947	0.999540	−0.204506	−0.216741	0.000000	0.837546	−0.269845	−0.032748	−0.828124	−0.052684
8	0	Boundary	0.999928	0.922748	−0.444438	−0.371287	0.000000	0.528038	−0.221645	−0.032692	−1.065439	0.381914
521	0	Boundary	0.999644	0.952079	−0.487173	−0.647487	0.000000	0.884155	0.313261	−0.031744	−0.378983	−0.053696
527	0	Boundary	0.275310	0.955129	0.881384	0.064600	0.875513	0.194844	0.851715	−0.010118	−0.008582	−0.054020
1050	0	Boundary	1.000000	0.937671	−0.212494	0.035146	0.000000	0.474058	0.186214	−0.035103	−0.996974	−0.052308
794	0	Boundary	0.999659	0.999898	0.597644	0.596845	0.159101	0.149742	0.987453	−0.033858	0.382540	−0.054299
283	0	Boundary	0.999674	0.229259	0.663739	−0.109842	0.000000	0.528487	0.509179	−0.031046	−0.751295	−0.052478
545	0	Boundary	0.996066	0.971404	−0.415240	−0.446575	0.000000	0.714367	0.256862	−0.032756	−0.548834	−0.053656
36	0	Boundary	0.876536	0.974149	0.895181	0.750179	0.441621	−0.277212	0.801766	−0.031696	0.257575	−0.053998
808	0	Boundary	0.987922	0.999883	−0.575969	−0.539010	0.000000	0.862905	0.340585	−0.034464	−1.012951	−0.048892
42	0	Boundary	0.999733	0.942155	−0.659336	−0.705233	0.000000	0.428490	0.991244	−0.033874	0.081275	−0.054245
558	0	Boundary	0.998720	0.937802	0.852196	0.723002	−0.362788	0.554720	0.817692	−0.032296	−0.529022	−0.053589
1077	0	Boundary	0.999977	0.621584	−0.593346	−0.419085	0.000000	0.975407	0.704715	−0.031386	−0.887923	−0.051167
56	0	Boundary	0.987637	0.284656	0.776732	−0.163130	0.000000	0.510213	0.546358	−0.029488	−0.493566	−0.053054
572	0	Boundary	1.000000	0.999380	−0.234672	−0.234872	0.000000	0.902117	0.010365	−0.032664	−1.054146	−0.018469
67	0	Boundary	0.995120	0.495787	0.427310	−0.133896	0.511221	0.274859	0.987462	−0.031382	1.563899	−0.054286
1091	0	Boundary	1.000000	0.952880	−0.176931	−0.092106	0.000000	0.170895	−0.378465	−0.034260	−1.008950	−0.050102
836	0	Boundary	0.984832	0.991054	−0.251314	−0.255289	−0.324182	0.920497	−0.038077	−0.034103	−0.947990	−0.051329
332	0	Boundary	0.596161	0.503281	0.093297	−0.250968	0.864256	−0.032444	0.231188	−0.031252	−0.508906	−0.053624
333	0	Boundary	0.986406	0.526701	0.679449	0.093505	−0.124262	0.789929	0.737497	−0.031485	−0.633219	−0.053128
847	0	Boundary	0.999878	0.991276	−0.530155	−0.544024	0.000000	0.558001	0.018859	−0.032756	−1.045182	−0.034123
337	0	Boundary	0.999998	0.920736	−0.338875	−0.373950	0.000000	0.024728	0.839860	−0.033334	−0.357383	−0.053993
340	0	Boundary	0.988807	0.981617	−0.185342	−0.242158	−0.286573	0.975199	−0.223720	−0.033759	−1.066023	0.958523
604	0	Boundary	0.972432	0.989007	−0.443711	−0.580728	0.000000	0.349282	0.866461	−0.033526	−0.621650	−0.053690
93	0	Boundary	0.999971	0.993816	−0.229454	−0.155943	0.000000	0.615072	0.648532	−0.032777	−1.018552	−0.045523
1117	0	Boundary	0.999941	0.834916	−0.378852	−0.330238	0.000000	0.253284	−0.015179	−0.033676	−0.980283	−0.050737
351	0	Boundary	0.594583	0.794384	−0.146733	−0.459772	0.718797	0.397309	0.672754	−0.032361	−0.160616	−0.054206
619	0	Boundary	0.995413	0.998809	−0.360961	−0.355929	0.000000	0.435781	−0.389250	−0.032927	−1.029619	−0.042901
621	0	Boundary	1.000000	0.490193	0.775470	0.009381	0.000000	0.946571	0.387488	−0.031458	−0.672893	−0.053014
623	0	Boundary	0.984413	0.607323	−0.044790	−0.552202	0.000000	0.807503	0.720666	−0.032840	−0.258638	−0.053969

Figure 2: Part 1 of Outlier Table

623	0	Boundary	0.984413	0.607323	−0.044790	−0.552202	0.000000	0.807503	0.720666	−0.032840	−0.258638	−0.053969
875	0	Boundary	1.000000	0.688198	−0.099761	0.424825	0.000000	0.740244	0.260112	−0.032763	−0.470520	−0.053782
368	0	Boundary	0.708776	0.999759	−0.314027	0.193558	0.000000	0.739550	0.489766	−0.033117	−0.773390	−0.053008
114	0	Boundary	1.000000	0.968133	−0.462474	−0.472678	0.000000	−0.080617	0.986356	−0.032841	1.990638	−0.054368
372	0	Boundary	0.999961	0.892770	−0.332195	−0.295481	0.000000	0.735786	−0.475183	−0.032859	−1.065031	0.241633
1141	0	Boundary	0.796886	0.989159	0.735492	0.679159	0.897119	−0.282697	0.891608	−0.028493	1.411144	−0.054326
377	0	Boundary	0.999955	0.543011	0.006852	0.808604	0.000000	0.745495	0.030912	−0.032783	−0.501838	−0.053739
634	0	Boundary	1.000000	0.973775	0.000000	0.000000	0.000000	0.139212	−0.370801	−0.031962	−1.065460	0.476889
383	0	Boundary	0.983083	1.000000	−0.417261	−0.348643	0.000000	0.007628	0.013886	−0.034129	−0.977304	−0.050602
389	0	Boundary	0.996778	0.911724	0.695309	0.639592	0.253531	0.766399	0.969019	−0.032135	0.963989	−0.054257
902	0	Boundary	0.983730	0.127498	−0.158601	−0.370553	−0.101687	0.953207	0.887201	−0.032757	−0.723482	−0.053156
657	0	Boundary	0.936496	0.966787	0.476147	0.081710	0.000000	0.267635	0.720551	−0.032333	−1.061796	0.060108
663	0	Boundary	0.999983	0.952087	−0.163766	−0.247688	0.000000	0.382840	−0.132714	−0.032976	−0.698205	−0.053401
1176	0	Boundary	1.000000	0.000000	1.000000	0.000000	0.000000	0.287130	0.505315	−0.032872	−0.121875	−0.054063
677	0	Boundary	0.969866	0.975547	−0.600598	−0.649724	0.000000	0.526340	0.063062	−0.033838	−1.063128	0.068908
421	0	Boundary	1.000000	0.833811	−0.091348	0.128108	0.000000	0.196290	−0.066422	−0.032763	−0.964579	−0.050269
935	0	Boundary	0.973113	0.904080	−0.281079	−0.306501	0.000000	−0.201455	0.972672	−0.034253	0.161957	−0.054291
940	0	Boundary	1.000000	0.999355	−0.161472	−0.165730	0.000000	−0.123388	−0.077092	−0.035115	−0.988122	−0.052575
947	0	Boundary	0.999951	0.999489	−0.231996	−0.230795	0.000000	0.957332	−0.059712	−0.032950	−0.917020	−0.051770
708	0	Boundary	0.999993	0.374278	−0.252093	0.761044	0.000000	0.741160	−0.162361	−0.032660	−1.062836	0.068797
453	0	Boundary	0.812004	0.197821	−0.324095	−0.327846	0.723411	0.326996	−0.348182	−0.031032	−0.612744	−0.053296
200	0	Boundary	1.000000	0.998155	−0.113500	−0.124083	0.000000	0.899461	−0.083396	−0.032444	−1.045900	−0.031652
202	0	Boundary	1.000000	0.903017	−0.665221	−0.674896	0.000000	0.947314	0.591421	−0.032109	−1.066436	33.188277
720	0	Boundary	0.383745	0.996927	0.696408	0.131907	0.000000	0.872291	0.767737	−0.029627	−0.661350	−0.053486
481	0	Boundary	1.000000	0.962770	−0.377539	−0.419321	0.000000	0.862902	0.466010	−0.032867	−1.011086	−0.046935
996	0	Boundary	0.998851	0.959415	0.551929	0.546367	0.605376	0.495742	0.805618	−0.032820	1.435093	−0.054339
491	0	Boundary	0.999196	0.441214	0.378727	−0.170753	0.000000	0.979000	0.774615	−0.032215	−0.634485	−0.053343
493	0	Boundary	1.000000	0.998877	−0.232468	−0.247011	0.000000	0.966363	−0.001469	−0.032736	−1.063390	0.087502
1009	0	Boundary	0.999989	0.275968	0.103305	−0.709024	0.000000	0.776001	−0.208982	−0.032425	−1.059755	0.016045
248	0	Boundary	0.999991	0.900182	−0.793428	−0.777158	0.000000	0.656264	0.093283	−0.033191	−0.852278	−0.052728
763	0	Boundary	0.738368	0.077807	0.704938	−0.035316	0.285711	0.268176	0.833152	−0.030235	−0.157963	−0.053690
510	0	Boundary	0.996074	−0.291742	0.568118	−0.313115	0.000000	0.894797	0.132003	−0.032515	−0.794575	−0.052754
37	1	Boundary	0.459678	0.993079	0.644322	0.221562	0.944413	0.750197	0.257200	−0.029378	0.137436	−0.054355
297	1	Boundary	1.000000	0.999899	0.365856	0.366458	0.000000	0.988900	−0.371052	−0.032690	−1.050195	−0.027422
557	1	Boundary	1.000000	0.954833	−0.316802	−0.202977	0.000000	0.935213	−0.286255	−0.034554	−0.645684	−0.053700

Figure 3: Part 2 of Outlier Table

557	1	Boundary	1.000000	0.954833	−0.316802	−0.202977	0.000000	0.935213	−0.286255	−0.034554	−0.645684	−0.053700
816	1	Boundary	0.999953	0.999987	−0.254486	−0.257876	0.000000	0.992984	−0.714136	−0.032657	−0.529335	−0.053678
50	1	Boundary	0.990172	0.995557	0.843943	0.774902	0.000000	0.999384	0.720489	−0.031222	−0.589625	−0.053552
66	1	Boundary	0.931183	0.999999	0.595929	0.590790	0.839027	0.399217	0.866698	−0.030785	−0.253276	−0.053778
591	1	Boundary	0.245619	0.999049	0.300700	−0.059870	0.000000	0.793884	0.786821	34.367195	0.391459	−0.054193
608	1	Boundary	−0.122685	0.951900	0.780717	0.080046	0.000000	0.883477	0.619918	−0.031172	−0.918893	−0.051589
861	1	Boundary	0.972903	1.000000	0.324285	0.404371	0.000000	0.991967	−0.101965	−0.029902	−0.819236	−0.052683
870	1	Boundary	0.995630	0.990237	0.598251	0.528546	0.505975	0.844636	0.690681	−0.030807	−0.936718	−0.049418
1129	1	Boundary	0.997847	0.988473	0.187795	0.135449	0.000000	0.985564	−0.433271	−0.032461	−1.063750	0.124056
874	1	Boundary	0.996589	0.999892	0.902333	0.880036	0.000000	0.871109	−0.520048	−0.031821	−1.064662	0.270302
884	1	Boundary	0.999992	0.999795	0.250071	0.247964	0.000000	0.986353	−0.490465	−0.032624	−0.870684	−0.052214
1144	1	Boundary	0.972239	0.962191	0.755388	0.771590	0.000000	0.981845	−0.164407	−0.032542	−0.928345	−0.050949
899	1	Boundary	0.062760	1.000000	0.937388	0.156098	0.000000	0.879822	0.841506	−0.026928	−0.907726	−0.048969
901	1	Boundary	0.999997	0.998106	0.286426	0.275216	0.000000	0.987939	−0.310903	−0.032724	−1.045003	−0.034150
145	1	Boundary	0.996320	0.993229	0.235133	0.191745	0.000000	0.974802	−0.449440	−0.033206	−0.041099	−0.054069
919	1	Boundary	1.000000	1.000000	0.910277	0.910277	0.000000	0.998373	0.114479	−0.032665	−0.203561	−0.053992
678	1	Boundary	0.998342	0.939386	−0.162496	−0.197115	0.572340	0.925730	−0.289822	−0.032818	−1.043784	−0.035243
1195	1	Boundary	0.999503	0.994784	0.847388	0.807294	0.000000	0.992071	0.779629	−0.032152	−0.867659	−0.051979
186	1	Boundary	0.997258	0.591218	0.536522	0.232970	0.511241	0.359473	0.810828	−0.030441	−0.691102	−0.052602
446	1	Boundary	0.948005	0.999525	0.529232	0.500563	0.000000	0.966060	−0.189248	−0.032866	−1.050162	−0.027035
969	1	Boundary	0.999397	0.858855	0.768574	0.573231	0.493004	0.471939	0.402177	−0.029136	−0.654173	−0.052383
462	1	Boundary	−0.312219	0.993686	0.546496	−0.083316	0.000000	0.317602	0.633745	−0.009160	−0.790955	−0.052700
471	1	Boundary	0.901081	0.431261	0.480171	0.046497	0.914528	0.820986	0.839478	−0.032150	0.180663	−0.054101
986	1	Boundary	0.999285	0.998506	0.827404	0.809625	0.956735	0.429413	0.370118	−0.031753	−0.844550	−0.052205
742	1	Boundary	1.000000	0.758236	0.520414	0.052706	−0.473722	0.947093	0.709609	−0.032642	−0.647499	−0.053441
488	1	Boundary	0.923444	0.607046	0.257076	0.509499	0.892259	0.715119	0.773417	−0.032364	−0.461781	−0.053756
747	1	Boundary	0.993749	0.975589	0.742672	0.588345	0.580290	0.925117	−0.087125	−0.031915	−0.965487	−0.049823
496	1	Boundary	0.979983	1.000000	−0.089506	−0.033871	0.727559	0.837977	0.506225	−0.033836	−0.720336	−0.053324
243	1	Boundary	0.970961	−0.062647	0.488377	−0.422728	0.409454	0.929040	0.799291	−0.032085	0.281000	−0.054137
1018	1	Boundary	0.999952	0.840449	0.624517	0.281141	0.509874	0.998318	0.234584	−0.032395	−0.735635	−0.053074
509	1	Boundary	−0.009317	0.998840	0.389204	0.047300	0.000000	0.721618	0.994127	−0.030924	1.934928	−0.054342
5	2	Boundary	0.595378	−0.531958	0.679654	−0.126799	0.455487	0.432046	0.988092	−0.029813	0.768742	−0.054167
263	2	Boundary	0.713135	0.311196	0.526621	−0.182596	−0.283029	−0.131136	0.053187	−0.032945	−0.698098	−0.053131
1035	2	Boundary	0.999970	−0.314415	0.811221	0.064862	0.000000	0.150068	0.974928	−0.030721	1.744308	−0.054266

Figure 4: Part 3 of Outlier Table

1035	2	Boundary	0.999970	−0.314415	0.811221	0.064862	0.000000	0.150068	0.974928	−0.030721	1.744308	−0.054266
533	2	Boundary	0.540865	1.000000	0.083330	−0.128718	0.000000	0.407346	0.925293	−0.028315	1.150191	−0.054314
28	2	Boundary	0.842654	0.391817	0.283967	0.541146	0.110819	0.584928	0.993057	−0.031174	1.348210	−0.054323
800	2	Boundary	0.625106	0.834977	0.636258	−0.027956	0.000000	0.209238	0.525758	−0.028113	−0.313087	−0.052980
45	2	Boundary	0.188614	0.031166	0.832167	0.342503	−0.153480	0.824574	0.152444	−0.013936	−0.296467	−0.052397
1076	2	Boundary	0.714275	0.025578	0.732727	−0.157460	0.000000	0.307395	0.422614	−0.029698	−0.603351	−0.052161
57	2	Boundary	0.809388	0.618737	−0.185972	−0.179937	−0.335617	0.581522	0.349141	−0.032771	−0.633104	−0.053479
314	2	Boundary	0.247143	0.031195	0.260139	0.261260	0.882335	0.581720	0.979954	−0.032836	1.370903	−0.054448
68	2	Boundary	0.443443	0.740395	0.279947	0.370327	0.000000	0.625446	0.931012	−0.032430	1.260689	−0.054279
580	2	Boundary	0.999788	0.893116	−0.585581	−0.384694	0.000000	0.873736	0.039651	−0.032249	−0.216801	−0.053925
73	2	Boundary	0.945316	0.235552	0.770774	0.380522	−0.369490	0.362169	0.971654	−0.031125	0.973996	−0.054222
852	2	Boundary	0.126200	0.258904	0.146875	−0.178838	0.000000	−0.213063	0.985947	−0.035313	0.914433	−0.054336
606	2	Boundary	0.752158	0.443502	0.000000	0.000000	0.000000	0.710299	−0.330908	−0.033128	−1.066153	1.340107
106	2	Boundary	−0.133945	0.283608	0.970914	−0.171278	0.000000	0.976835	0.905625	−0.030864	−0.375026	−0.053857
618	2	Boundary	0.999011	0.999923	−0.172646	−0.158803	0.000000	0.521042	−0.250651	−0.030617	0.331231	−0.054010
371	2	Boundary	0.784137	0.605739	0.617510	0.759004	0.000000	0.167165	0.989077	−0.032890	1.969029	−0.054354
141	2	Boundary	0.812237	0.067582	0.000000	0.000000	0.131656	0.007968	0.969886	−0.032212	1.172919	−0.054243
689	2	Boundary	0.552319	0.939522	−0.232991	−0.261600	0.000000	0.092729	0.457548	−0.034397	−0.909822	−0.052138
180	2	Boundary	0.961096	0.280749	0.177005	−0.013683	−0.177218	0.956326	−0.026494	−0.032963	0.251647	−0.054172
201	2	Boundary	0.033564	−0.146953	−0.142238	−0.215256	−0.133407	0.653039	0.547611	−0.029229	−0.849317	−0.052171
225	2	Boundary	0.729880	0.987338	0.502494	0.427280	0.000000	−0.412246	0.998549	−0.032672	2.109923	−0.054356
252	2	Boundary	0.999995	0.984277	−0.345149	−0.318476	−0.317350	0.736551	−0.222035	−0.032901	−0.314227	−0.053953

Figure 5: Part 4 of Outlier Table

All 121 detected outliers are classified as boundary outliers, suggesting they represent extreme cases within their respective clusters rather than forming isolated small clusters.

4 Interpretation

The combination of VAEs and clustering effectively identified 121 boundary outliers in the dataset. These outliers represent approximately 10.1% of the total data points. The absence of small cluster outliers suggests that the data forms well-defined, balanced clusters in the latent space.

The optimal latent dimension of 3 indicates that the 10 original features contained significant redundancy, and the underlying structure could be effectively captured in just 3 dimensions. This confirms the VAE's ability to learn meaningful compressed representations.

5 Conclusion

This work demonstrates the effectiveness of combining VAEs with clustering techniques for outlier detection. The approach successfully identified boundary outliers while reducing dimensionality from 10 to 3 features. The VAE's ability to capture non-linear relationships in the data made it particularly effective for this dataset.

Future work could explore different VAE architectures, alternative clustering algorithms, or apply this technique to other domains where outlier detection is critical, such as fraud detection or anomaly identification in medical data.